知 能 と 複 雑 系 128−19
(2002.5.24)

ILPによるWebページ作成者の意図抽出について

櫻井 成一朗*
*東京工業大学大学院情報理工学研究科

Webコミュニティ手法は，同種の内容を持つWebページ同士を関連付ける方法であり，Webのグラフ構造に基づいてWebコミュニティが構成される．有用なWebコミュニティが得られる場合も多いが，得られるWebコミュニティ自身はWeb作成者の意図を必ずしも考慮したものではない．本論文では，Webページ作成者の作成意図を自らの分類階層の提示にあるととらえて，帰納論理プログラミングによって分類階層知識を抽出する方法を提案する．分類階層知識を抽出する事によって，特定のカテゴリがなぜ導入されたのかと言う事を説明できるようになり，その結果，抽出された分類階層知識を用いることで，バックリンクページが何を表しているのかというページ作成者の意図を検索者が推定できるようになる．

Extracting Web Page Writer's intention by using ILP

SAKURAI Seiichiro*
* Department of Computer Science, Tokyo Institute of Technology

This paper focuses on extracting Web page writer's intention by using ILP. It is assumed that the writer's intention is to present the writer's implicit classification knowledge. At first, backward-link pages are collected by a search engine. URLs of the pages are categorized by syntactic constraints. Then the classificatin knowledge of URLs are extracted by ILP. If the classificatin knowledge is obtained successfully, a backward-link page will be identified by the classification knowledge.

1 はじめに

WWWではその大規模さゆえに重要な情報が埋没してしまう可能性が高く，検索効率を高めるために様々な試みが行われている．しかしながら，検索エンジンやディレクトリサービスによって情報を引き出すしか方法が提供されていないのが現状である．検索エンジンを高速化するためにはアルタビスタ [AltaVista] のようにハードウェアの高速化とソフトウェアの効率化を組み合わせて実現したものもあれば，Google [Google] のように有用なページの特徴付けを行い，有用なページを検索するアルゴリズムを工夫した検索エンジンもあり，非常に多くの検索エンジンが実現されている．このような検索エンジンを利用した知識獲得の試みも活発に行われている．特定のトピックに関して有用なページ集合を求める方法としては，HITS [Kleinberg 98]，Webコミュニティ発見手法 [村田 01] がある．HITSは入力キーワードのトピックに関する有用なページを出力するアルゴリズムである．HITSの特異値分解による手法は必ずしも有効では無く [廣川 01]，検索語と関連ないページが出力される場合 (topic drift) がある．Chakrabarti[Chakrabarti 01] らはHTML構文情報を利用してWebページを分割し，テキスト情報の利用と組み合わせてtopic driftを軽減している．Webコミュニティ発見手法 [村田 01] は，特定のURLに対するリンクを逆に辿ることで完全2部グラフを抽出し，URLのグルーピングを行っている．Kumar [Kumar 99] はWebのスナップショットからWebコミュニティを数え上げるアルゴリズム

を提案している．HITS や Web コミュニティ発見手法では，Web のグラフ構造に基づいて知識を抽出しており，これらのアプローチでは多量の知識ゆえに質の良い知識を得ることができると考えられる．

これらの方法は Web ページ間のグラフ構造から知識を抽出しようという試みであるが，必ずしも Web 製作者の知識を十分に抽出できているとは言えない．Web 作成者が Web ページを作成する際には，利用者の利便性を考えて様々な労力を払っていることが少なくないはずである．それらの労力の一端でも抽出できれば，利用者の利便性の向上が期待できるであろう．任意の Web ページから作成者の意図抽出を行う事はもちろん困難であるので，本論文では特定の構文制約の下で Web ページを作成する際の作成者の意図を抽出する方法について考察する．特定の構文制約としては，単一の HTML ファイルとしてリンク集を記載する場合に限定する．単一の HTML ファイルで数百から数千のリンクを有するリンク集を作成する際には，利用者の利便性を考慮すればリンクの分類が必要であり，実際に多くのリンク集ではリンクの分類が行われている．したがって，リンク集を作成する事は，利用者の利便性向上のためにリンクの分類を行う事であると言うことができるであろう．Web ページを分類知識獲得の知識源として考えることによって，分類知識の動的性や多様性を反映させられるのではないかと考えられる．

2 検索エンジンから収集可能な情報

2.1 バックリンクページ

Google や AltaVista では，キーワードの代わりに link:URL という入力によって URL をアンカーとして含むようなページを検索できる．特定の URL をアンカーとして含むようなページを当該 URL のバックリンクページと呼ぶ．有名企業の URL に対するバックリンクページを検索してみると，その多くがリンク集となることが観察できる．複数の URL のバックリンクページの共通集合を取る事で，リンク集の割合が増加する．AltaVista を使って同業種の３つの有名企業についてバックリンクページを調べた結果を表 1 に示す．

表 1 でリンク数が 0 となっているページについては，別ページにリダイレクトしているページや HTML のフレーム機能を使って記述されたものであるので，除外しておく．一つの URL のバックリンクを検索すると，50 以上のリンクを持つものが少なくない事がわかる．Altavista ではバックリンクページの共通集合を直接検索できるので，共通集合として検索した結果ではより顕著になっている．表 1 では，共通集合を検索することでほぼ半数近くが 50 以上のリンクを持つページであることがわかる．

このように一つの話題に限定して，その話題に関する著名な URL に対するバックリンクページはリンクを大量に持つものが少なくない事がわかる．このようなリンクを大量に持つページにおいては，何らかの工夫をしなければ使い勝手の悪いページとなる事から，以下ではこれらのリンク集を作成した際の作者の意図の抽出方法について考察する．

2.2 リンク集の特徴

本論文で対象としている Web ページは単一の HTML ファイルでリンク集を構成している場合である．リンクの数が数十であれば何の工夫もなしに HTML ファイルを記述してもブラウザの検索機能を利用することでも困らないが，数百からなる Web ページへのリンク集を作成する場合には何らかの工夫が必要となる．リンク集を観察してみると，単一の話題に関するリンク集以外はカテゴリを導入する事でリンクの分類を行っている．カテゴリを導入するための構文は Web ページによって異なるが，hr タグや div タグあるいは table 等の区切りとなる HTML タグが導入されているものが多い．

図 1 の (a) では，h3 タグがカテゴリの区切りであり，li タグによって一つのカテゴリが表されている．同じく (b) では，列が一つのカテゴリを表現している．

表 2: 区切りで用いられる HTML タグの例

h1	h2	h3	...
font	em	b	...
div	hr	...	...

カテゴリの区切りをなす HTML タグの例を表 2 に示す．表 2 に示したように，カテゴリの区切り

表 1: バックリンクページのリンク数

リンク数	0	1〜50	51〜100	101〜150	151〜200	201〜250	251〜300	301〜350	351〜
トヨタ	114	520	133	53	28	10	11	14	39
日産	150	444	148	64	36	14	12	14	38
本田	110	594	107	44	22	5	13	9	21
3社	83	410	185	91	40	25	16	20	45

```
<h3> カテゴリ名A</h3>
<ul>
  <li><a href="...">...</a>...</li>
  ...
</ul>
<h3> カテゴリ名B</h3>
<ul>
  <li><a href="...">...</a>...</li>
  ...
</ul>
```

(a) リストによるリンク集

```
<table>
  <tr>
    <td>カテゴリ名A</td>
    <td> カテゴリ名B</td>
  </tr>
  <tr>
    <td><a href="...">...</a>...</td>...
  </tr>
  <tr>
    <td><a href="...">...</a>...</td>...
  </tr>
</table>
```

(b) テーブルによるリンク集

図 1: リンク集で用いられる HTML 構文の例

をなす HTML タグは必ずしも一意ではないものの，HTML の木構造におけるアンカータグの出現位置情報を利用する事ができる．すなわち，同一のカテゴリであれば木構造の中で同一の深さにアンカーが出現する事が多いという事を利用して，カテゴリ間の区切りをなすタグを同定すれば，各 HTML ファイルにおけるカテゴリを分別する事ができる．言い換えれば，各 HTML ファイルのカテゴリを収集する事で，特定の話題の Web コミュニティを得るだけでなく，特定の話題に関連する別の話題の Web コミュニティも得る事ができるのである．例えば，自動車会社の URL を元にして，新聞社の Web コミュニティや検索エンジンの Web コミュニティを得る事ができる [若月 02]．

3 意図抽出の基本的考え方

3.1 著者の意図モデル

リンク集は構文的な制約が強いので，カテゴリを分別する事は比較的容易に行う事ができる．しかしながら，導入されるカテゴリは検索の出発点となった URL とは一見関連の薄いものが少なくない．例えば，自動車会社から検索したのであれば，自動車関連の話題に関するカテゴリが出現する事は容易に予想できる．実際，自動車会社を起点とするのであれば，タイヤやＦ１に関するカテゴリ等自動車に関連が深いものが出現している．一方，現実の Web ページでは主話題に関連の薄いと思われるようなカテゴリも同時に出現しているのである．[若月 02] では，自動車会社を起点として検索したにもかかわらず，放送局や旅行関連会社あるいはコンピュータ関連会社というカテゴリも抽出されたのである．

この事実から，バックリンクページを辿る事によって汎用のリンク集が見つけられたと考える事もできるが，HTML ファイルに含まれるリンクの数が高々千数百しか無い事を考慮すると，汎用のリンク集で

あるとは考えにくい．一例を挙げてみると,「経営学」関連のリンク集の中の「企業リンク」の中に自動車会社が含まれており，その中に自動車会社へのリンクが記述されているのである．このリンク集の作者は「経営学」という大項目,「企業リンク」という中項目,「自動車会社」という小項目を導入していた事になる．放送局のリンクが記述されていたのも実はこの「経営学」関連のリンク集であった．また,ある自動機関連リンクの中には「自動車会社」だけでなく,「電機・電気」,「製薬」,「ガラス加工」,「鋼材」がサブカテゴリとして導入されていた．これらの項目分類を表 3 に示す.

表 3: リンク集における項目分類の例

大項目	経営学	企業リンク
中項目	企業リンク	自動機関連リンク
小項目	自動車会社, 放送局 旅行会社	自動車会社, 電機・電気 製薬

以上から，Web ページ作成者のリンク集作成の意図と検索者の意図が異なっているために，検索者の意図しないようなカテゴリが出現していたと考えられる．このように Web ページ作成者の作成した分類階層と利用者の分類階層が一致しない事はままある事であるが，このような分類階層知識を作者の Web ページ作成の意図として抽出できれば，利用者は分類階層知識をうまく利用することで、所望の Web ページに速やかにたどり着く事が期待できる.

3.2 意図抽出と ILP

Web ページ作成者の意図を分類階層知識として考えるのであれば，意図抽出とは

作成者はなぜカテゴリを導入したのか

を推定する事に他ならない．表 3 のような分類階層知識を予め持っていれば，なぜ一見無関係に思える「旅行会社」や「放送局」といったカテゴリが出現するのかと言う事を説明できる．このような分類階層知識を抽出するために，帰納的論理プログラミング (ILP)[Muggleton 95] を利用する.

ILP とは観察事実 o と背景知識 Σ が与えられた時に，$\Sigma \cup \Pi \vdash o$ となる論理プログラム Π を推論する枠組である．ILP を利用して分類階層知識を抽出するために，構文情報を用いて分類された各カテゴリに対して述語記号 cat_i を割り当て，$cat_i(url_j)$ という観察事実として表現する．このように考えると，ILP で推論すべき関係は，カテゴリ間の包摂関係 $\Pi = \{cat_i(X) \leftarrow cat_j(X), \cdots\}$ となることがわかる.

4 ILP の適用に関する考察

ILP 適用の実際は，まず適当な複数の URL に対してバックリンクページを検索し，検索したバックリンクページを構文情報に基づいてカテゴリの弁別を行う．このとき，カテゴリ間の語彙の統一を行っておく必要があるので，同義語に関する適当なシソーラスを予め用意しておき，用語を統一することにする．例えば,「車」,「自動車」などは同義語として予め与えておく．カテゴリの弁別結果を訓練例として ILP を適用することによって，Web 作成者の分類階層知識を学習する．表 3 に示した左側の分類階層からは,

$$\text{経営学}\ (X) \leftarrow \text{企業リンク}\ (X).$$

$$\text{企業リンク}\ (X) \leftarrow \text{自動車会社}\ (X).$$

という分類階層知識が候補として得られる事になる．また，訓練例の総数は高々1000 個程度であるので，計算時間よりも仮説候補を絞り込む事が重要となる．仮説候補を絞り込む一つの方法としては，背景知識を与えておき，背景知識によって仮説空間を制約する事が考えられる．例えば,「自動車会社」は「企業リンク」の下位に分類されるが,「自動車会社」の下位に「企業リンク」が分類されることはないという一貫性制約の形で背景知識を与えておけばよい.

リンク集作成者の分類階層知識を抽出する際に，注意しなければならない事は汎化のための十分な訓練例が集められない事である。極端な場合には，特定のカテゴリは単一の HTML ファイルにしか存在しないのである．実際，放送局のカテゴリは 1000 個のバックリンクを検索しても，一つの HTML ファイルにしか存在していなかった．単一例から汎化しなければならない場合にも，信頼性を向上させなければならない場合には，適切な背景知識を与えておく.

本論文で対象としている分類階層知識は，一つの HTML ファイルに限定しているため，いつでも分

類階層知識が抽出できるわけではない．表 3 の右側の項目分類については単一の HTML ファイルではなく，中項目や大項目は当該 URL へのリンクページを辿らないと，出現しない記号である．そのため，単一の HTML ファイルでない場合には，ILP を適用することはできない．このような場合に対応するためには，一つの Web サイト内の HTML ファイルについてはリンクを展開して置かなければならない．

5 おわりに

本論文では，リンク集を作成する際に，作者がどのような観点でグルーピングを行っているのかと言う作者の意図の抽出方法について提案した．一つの HTML ファイルによるリンク集であると制限する事で，当該 HTML ファイルがどのような意図で作成された HTML ファイルなのかと言う事を ILP によって探索する．ILP によって分類階層知識が抽出できれば，この分類階層知識を用いて，バックリンクページを分類できる．この分類階層知識の下では，Web ページ作成者の意図を部分的に推定した事になるので，利用者の検索を助けられる可能性がある．まだ本格的な実験を行っていないので，現実の Web ページについての意図抽出を行って，本方法の有効性を確認するのは今後の課題である．

参考文献

[AltaVista] AltaVista The Search Company, http://www.altavista.com/

[Chakrabarti 01] Soumen Chakrabarti, Mukul M. Joshi, Vivek B. Tawde, "Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks", SIGIR 2001, http://www.cs.berkeley.edu/~soumen/

[Google] Google, http://google.com/

[廣川 01] 廣川佐千男, 池田大輔, "Web グラフの構造解析", 人工知能学会誌,Vol16,No.4,pp.525-529,2001.

[Kleinberg 98] Jon M. Kleinberg, "Authoritative Source in Hyper linked Enviroment", *Proc.ACM-SIAM Symp.on Discrete Algorithms*, pp.668-677(1998)

[Kumar 99] Kumar,R. et al., "Trowling the Web for Emerging Cyber-Communities",*Proc. of the* 8th *WWW Conf.*

[村田 01] 村田剛志, "参照の共起性に基づく Web コミュニティの発見", 人工知能学会誌, Vol16, No.3, pp.316-323, 2001.

[page98] Page, L.Brin,S.Motwani R. and Winograd T., "The Page Rank Citation Ranking" Bringing Order to the Web,Online manuscript, http://www-db.stanford.edu/~backrub/pageranksub.ps, 1998.

[若月 02] 若月謙太郎："リンク解析による WWW からの知識獲得に関する研究"，東京工業大学大学院情報理工学研究科　修士論文 (2002)

[Muggleton 95] S. Muggleton: Inverse Entailment and Progol, New Gen. Comput., Vol. 13, pp. 245-286, 1995.