社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (13)
2008－BIO－15 (13)
2008／12／17

構造同値に基づいたWebページの分類の高速化技術の提案

山下 長義 †, 森山 甲一 ††, 沼尾 正行 ††, 栗原 聡 ‡‡

† 大阪大学情報科学研究科情報数理学専攻 †† 大阪大学産業科学研究所 ‡ 科学技術振興機構 戦略的創造研究推進事業

著者らは，これまでWebのリンク構造の類似性に着目してWebページを分類する手法を提案し，その有効性を検証してきた．しかし，実用化を考えた場合，Webページの収集に時間がかかるという問題があった．この問題に対し，現在では検索サイトのWeb APIの利便性が大幅に向上し，実際のWebページにアクセスせずに情報を収集できるようになったことから，検索の結果得られるWebページを構造同値に基づいて分類する処理を高速化する新しい手法を提案する．実験の結果，Webページの収集時間は大幅に短縮され，精度はほぼ変わらず分類できることがわかった．

A Proposal of Speed up Technology for Classification of Web Pages Based on Structural Equivalence

Nagayoshi YAMASHITA † Koichi MORIYAMA †† Masayuki NUMAO †† Satoshi KURIHARA ‡‡

† Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University

†† The Institute of Scientific and Industrial Research, Osaka University ‡ JST CREST

The authors proposed a methodology to classify web pages focusing on the similarity of link structure. However, for practical use, there is a problem that collecting Web pages require considerable time. Therefore, we propose a new methodology to classify web pages from a search engine because improved convenience of Web API enables to obtain a number of links immediately without accessing to the actual Web. A experiment shows that time for collecting web pages has been shorten drastically and the classification precision is almost the same.

1 はじめに

Web上の情報量の増加にともない，ユーザーが欲しい情報を見つけるコストが増大している．さらにユーザーが情報を探すために検索式を用いるが，一度の検索で用いられる平均キーワード数は2個足らずであり情報を絞り込むには十分ではない．このような現状を解消するためには，情報システムによる支援が必要である．そこで，検索結果の全体を把握し欲しい情報を見つけやすくするために，検索の結果得られたWebページを分類する研究が行われている[1, 2]．著者らは，Webのリンク構造の類似性に着目してWebページを分類する方法を提案してきた[3]．あるWebページ同士に類似性が認められる場合，そのWebページ同士はそれぞれ類似するリンク構造を持っていると考えられる．そこでWebページを分類するために，社会科学などで古くより開発されてきた「構造同値」という概念を導入した．ネットワーク内の2つの点が他のすべての点と完全に同じ関係を持つ場合，2つの点は構造同値であるという．ただし，構造同値の定義は制約が強いため，我々は相関関係を計算する方法により構造同値の度合いを求め，それを利用してWebページを分類する手法を提案し，その有効性を検証してきた．

実用化を考えた場合，このWebページの収集に時間がかかるという課題があった．しかし近年，検

索サイトのWeb APIの利便性が大幅に向上し，まとまった数のインバウンドリンクを瞬時に得られるようになったことから，本論文では，検索サイトからインバウンドリンクを収集し，検索の結果得られるWebページを分類する新しい手法を提案する．このWebページ集合に対しても提案手法が有効であることが検証されれば，収集時間が大幅に短縮されるため，提案手法の実用化に近づくと考えられる．

以下，2節では関連手法について簡単に述べ，3節で構造同値に基づいたWebページの分類方法について説明し，4節で提案手法を説明する．そして，5節で実験を説明し，6節で評価し，7節にてまとめを述べる．

2 関連研究

本論文において導入する構造同値という概念はさまざまなネットワークに適用されている．たとえば，企業間関係の分析[4]や論文の参照関係から研究トピックを抽出する研究[5]に用いられている．

Web構造マイニングにおいては，リンク構造の共参照関係から任意のWebページに対する類似サイトを発見する研究[6]が行われている．また，巨大で複雑なネットワークの中の似た結合パターンをもつサブネットワークをEMアルゴリズムを用いて発見する方法[7]が提案されている．

3 構造同値に基づくWebページの分類

検索の結果得られるWebページを構造同値に基づいて分類[3]する．まず，構造同値の度合いを求めるためWebページの隣接関係の相関を求め，それを基にデンドログラムを形成する．そして，そのデンドログラムにおいてどのレベルに形成されるクラスタを出力するかを決定する．

外部から同時に参照されていることが多いクラスタ内のページや，外部に対して同時に参照していることが多いクラスタ内のページは互いに関連している可能性が高い．そこで，このような外部のWebページとの間にリンクが多数存在する極大のクラスタをデンドログラムにおけるクラスタの境界とし，このようなクラスタを複数抽出することでWebページの分類を行う．

まず，デンドログラムにおいて相関係数が0から1の間で融合されるクラスタごとに以下の定義1から定義3に基づいて極大関連クラスタかどうかを判定する．そして，求めた複数の極大関連クラスタを分類結果として出力する．

定義 1 任意の Web ページ i とクラスタ C_k との間にリンクが存在する割合 $\Delta_{i\to C_k}$ を以下のように定義する．

$$\Delta_{i\to C_k} = \frac{\sum_{j\in C_k} X_{ij}}{|C_k|} \tag{1}$$

ただし，

$$X_{ij} = \begin{cases} 1 & Web\text{ページ}i\text{から}Web\text{ページ}j\text{へリンクがあるとき} \\ 0 & Web\text{ページ}i\text{から}Web\text{ページ}j\text{へリンクがないとき} \end{cases} \tag{2}$$

とする．

定義 2 任意の Web ページ i とクラスタ C_l があるとする．

$\Delta_{i\to C_l} \geq \alpha$ のとき，Web ページ i をクラスタ C_l のハブと呼ぶ．そして，ハブとなる Web ページが存在するクラスタ C_l を関連クラスタと呼ぶ．

定義 3 C が関連クラスタ かつ $C \subsetneq D$ となる任意のクラスタ D が関連クラスタでないとき，C を極大関連クラスタと呼び，デンドログラムにおけるクラスタの境界とする．

4 提案手法

本論文では，検索の結果得られるWebページに対するインバウンドリンクを収集し，構造同値に基づいた分類をこれらのWebページに適用する手法を提案する．このWebページ集合に対しても分類手法が有効であることが検証されれば収集時間が短縮されることが期待でき，分類手法の実用化に近づくと考えられる．従来のWebページの収集方法と本論文で提案するWebページの収集方法の詳細は以下の通りである．

4.1 従来のWebページの収集方法

これまでは，検索の結果得られるWebページを収集し，これらのWebページに隣接しているWebページを収集して，さらに，これらすべてのWebページを分類するために，隣接ページのアウトバウンドリンクも収集していた[3]．この中で，点線のリンクはアウトバウンドリンクであるため，収集時間の増加の原因となる．

4.2 提案するWebページの収集手法

本論文で提案するWebページの収集方法では検索の結果得られるWebページを収集し，これらのWebページに対するインバウンドリンクを収集する．

5 実験

リンク解析を行うWebページをYahoo!デベロッパーネットワークで以下のようにして収集する．検索サイトにあるキーワードを入力し，結果上位m(最大1000[1])までのWebページのURLを収集する．そして，これらのWebページに対するインバウンドリンクを1ページにつき最大nページを収集することとした．ただし，異なるドメイン間のリンクのみを用いる．

このようにして得られたWebページ集合によって，検索の結果得られたmページを分類する．本論文では，nを100とし，mは次節において決定し，さらに検索語としてS社(電機メーカー)を用いて，Webページを収集する．

このようにして収集したWebページ間のリンク構造に対して相関係数を計算しデンドログラムを作成し，極大関連クラスタを出力する．ただし，クラスタに対してハブとなるために必要なリンクが存在する割合αを0.5とし，評価においては極大関連クラスタ内のページ数が1のものは除外する．

6 評価

Webページ集合に対して構造同値に基づいた分類手法を適用し，その結果の評価を行った．

6.1 分類対象のWebページ数と構造同値性の関係

適切なmを決定するために，検索の結果得られるWebページ数とこのWebページ集合内における構造同値性の関係の検証を行った．検索の結果得られるWebページ数であるmが50，200，500，1000の場合ごとにデンドログラムを作成し，それぞれのデンドログラムにおける相関の値とその相関値までに融合したクラスタの割合の関係を調査した．相関が1に近いときに融合するクラスタが多数存在すれば，リンク構造が似ているWebページが多数存在し，リンク構造の類似度による分類に適したWebページ集合であるということができる．

[1] Yahoo!デベロッパーネットワークでは，最大1000位までのWebページを収集できる．

図1は，相関と融合したクラスタの割合との関係を表している．たとえば図1のS社の例では，mが1000の場合，横軸の値が0.3のとき縦軸の値は0.6の値を示した．これは，相関が0.3以上のWebページを1つのクラスタに分類したとき，集合内のWebページの60%は構造が似ている他のWebページと融合して1つのクラスタを形成していることを示している．一方，mが50の場合では，縦軸の値はおよそ0.1を示し，10数%程度しか他のクラスタと融合しなかった．

このように，mが1000のときに互いに構造が似たWebページがより多く存在していることから，上位1000までのWebページ集合が構造の類似度による分類に適したWebページ集合であることがわかった．

よって，以下ではmを1000とした．

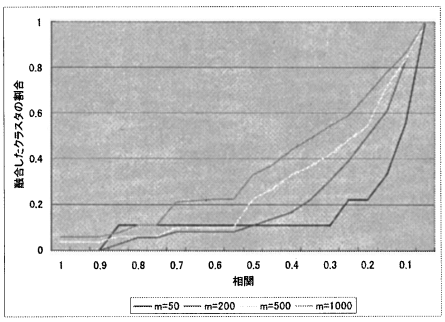

Fig. 1 相関と融合したクラスタの割合の関係（S社）

6.2 Webページの収集にかかる時間の比較

従来手法でWebページの収集にかかった時間と今回のWebページの収集方法でかかった時間の比較を行った．

Webページの収集に用いたサーバのスペックは，Dual AMD Opteron 2.2GHz(CPU)，6GB(RAM)である．

従来の収集方法では1969ページの隣接関係を収集するために190.2分かかっていたが，提案する収集方法では1000ページに対するインバウンドを収集するために4.2分かかった．このように，収集する時間を45分の1に大幅に短縮することができた（図2）．

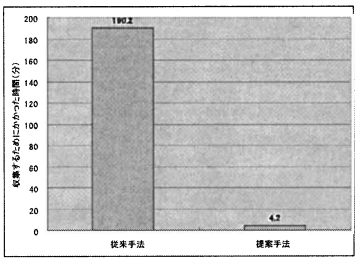

Fig. 2　Web ページ収集時間の比較

6.3　被験者による分類精度の評価

それぞれの極大関連クラスタ内の Web ページが互いに関連しているかを評価するために，被験者 10 人に極大関連クラスタ内のそれぞれのページを見てもらい，それぞれの Web ページを 3 つ以内の言葉で表現してもらった．その結果からそれぞれの極大関連クラスタ内の Web ページが互いに関連しているかどうかを判断した．クラスタ内の Web ページが共通の概念を持つ 1 つ以上の言葉で被験者によって表現されていれば，その極大関連クラスタは内容が関連しているとみなした．

精度は，全極大関連クラスタ数に対する内容が関連しているものの割合と定義する．

その結果，これまでの収集方法と今回の収集方法における精度はそれぞれ 0.70 と 0.68 となり，ほぼ変わらなかった（図 3）．

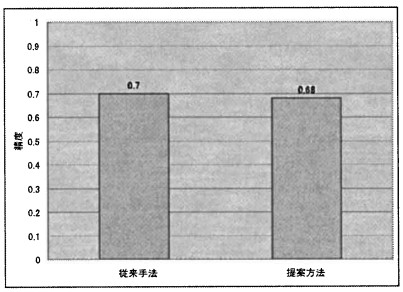

Fig. 3　これまでの収集方法と今回の収集方法との精度における比較

7　まとめ

本論文では，インバウンドリンクのみを収集して検索の結果得られる Web ページを分類する手法を提案した．本手法は Web ページ集合を収集するために要する時間を 45 分の 1 に短縮する一方で，分類精度はこれまでの方法と変わらず分類できたことを示した．

参考文献

1) H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma and J. Ma: “Learning to cluster web search results”, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2004).

2) O. Zamir and O. Etzioni: “Web document clustering: A feasibility demonstration”, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998).

3) 山下, 森山, 栗原, 沼尾：“リンク構造に基づいた www からのトピック抽出”, 情報処理学会論文誌 数理モデル化と応用 (2008).

4) 渡邊, 小坂：“日本における企業間関係の社会ネットワーク分析”, 経営情報学会春季全国研究発表大会, pp. 356–359 (2005).

5) 榊, 松尾, 市瀬, 武田, 石塚：“論文データベースからの研究トピック抽出”, 人工知能学会第 19 回全国大会 (2005).

6) J. Dean and M. R.Henzinger: “Finding related pages in the world wide web”, Computer Networks(Amsterdam,Netherlands) (1994).

7) M. E. J. E. Newman and E. A. A. Leicht: “Mixture models and exploratory analysis in networks.”, Proc Natl Acad Sci U S A (2007).