

# 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<sup>†</sup>, 森山 甲一<sup>††</sup>, 沼尾 正行<sup>††</sup>, 栗原 聡<sup>‡</sup>

<sup>†</sup> 大阪大学情報科学研究科情報数理学専攻 <sup>††</sup> 大阪大学産業科学研究所 <sup>‡</sup> 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<sup>†</sup> Koichi MORIYAMA<sup>††</sup> Masayuki NUMAO<sup>††</sup> Satoshi KURIHARA<sup>‡</sup>

<sup>†</sup> Department of Information and Physical Sciences, Graduate School of Information Science and  
Technology, Osaka University

<sup>††</sup> The Institute of Scientific and Industrial Research, Osaka University <sup>‡</sup> 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<sup>1, 2)</sup>。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<sup>3)</sup>。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

素サイトの 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<sup>4)</sup> や論文の参照関係から研究トピックを抽出する研究<sup>5)</sup> 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<sup>6)</sup> 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<sup>7)</sup> 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<sup>3)</sup> する。まず、構造同値の度合いを求め、その度合いに基づいて 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

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  $\Delta_{i \rightarrow C_k}$  を以下のように定義する。

$$\Delta_{i \rightarrow C_k} = \frac{\sum_{j \in C_k} X_{ij}}{|C_k|} \quad (1)$$

ただし、

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

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。

$\Delta_{i \rightarrow C_l} \geq \alpha$  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<sup>3)</sup>。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<sup>1</sup>) 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  $\alpha$  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出来る。

<sup>1</sup> 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。

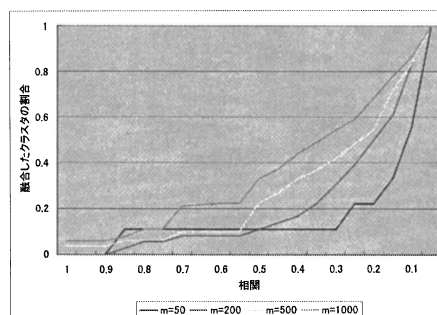


Fig. 1 相関と融合したクラスタの割合の関係 (S 社)

### 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)。

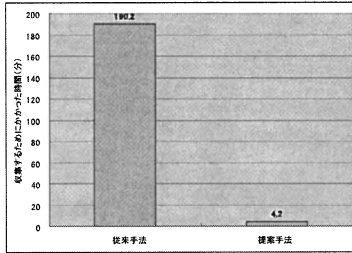


Fig. 2 Web ページ収集時間の比較

### 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)。

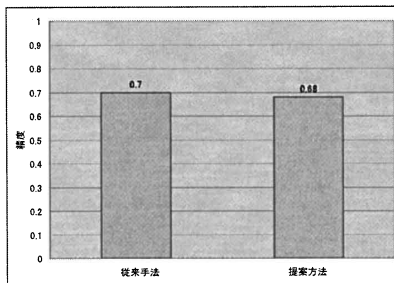


Fig. 3 これまでの収集方法と今回の収集方法との精度における比較

## 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. Newman and E. A. Leicht: "Mixture models and exploratory analysis in networks.", Proc Natl Acad Sci U S A (2007).