

6U-06 Web ページ間のリンク構造に着目した 製品解説記事の自動収集

島津光伸 藤本和則

日本電信電話(株)コミュニケーション科学基礎研究所

1 はじめに

我々は、インターネット利用者に様々なアドバイスを与える「DSIU システム」の研究を進めている[1,2]。ここに、DSIU は「デシユウ」と読み、*Decision Support for Internet Users* の略である。DSIU はアドバイスのための推論知識を WWW 上のテキストから自動獲得する。このため、DSIU は、まず、知識獲得の源となるテキスト、特に「対象とする分野についての解説文が含まれるテキスト(以下、解説テキスト)」を WWW から収集する。DSIU においては、こうした解説テキストを高速に収集する枠組みの実現が重要となる。

WWW 上の情報のほとんどは、リンクで参照付けされたハイパーテキスト(以下、ページと呼ぶ)として提供される。したがって、手がかりとなる URL をもとに一連の情報を集めるには「URL の指すページ」に合わせて「URL の指すページの周辺のページ」もリンクを辿りながら収集する必要がある。こうした収集での「周辺のページ」には、集める必要のあるページとそうでないページとが存在する。例えば、メーカーの Web サイトから製品記事を集める場合には、製品の特徴や、スペックの詳細などのページのみを収集すればよい(メーカーのトップページや、問い合わせ先などのページは必要ない)。こうした状況で効率的な収集を実現するには、リンク先が必要なページかどうか推定し、必要な見込みの高いページを優先して収集するような仕組みが必要となる。

本稿では、ページのもつ意味的な階層構造に着目して、一連の情報を効率的に集める情報の収集法を提案する。まず、ある時点で収集されたページ間のリンク構造に着目して、各ページの意味的な親子関係を推定する方法を示す。そして、この推定の結果に基づいて収集先を絞り込む方法を示す。最後に、デジタルカメラの製品記事の収集を例に、リンク先を全て収集する方法との比較により、収集先を絞り込む本手法の有効性を評価する。

2 意味的な階層構造の推定に基づく収集法

2.1 ページの意味的な階層構造

WWW 上のページには意味的な階層構造があることが多い。例えば、メーカーの提供する製品紹介ページでは、製品の層(デジタルカメラ、MDなどの製品を並べた層)、機種層(ある製品の各機種を並べた層)、仕様特徴の層(ある機種の詳細情報を並べた層)といった構造がある。情報の収集にあたっては、あるページの参照先を全て集めるというよりもむしろ、こうした意味的な木構造の一部を集めたい場合が多い。例えば、デジタルカメラについての情報を集めるには、デジタルカメラのトップページの下に位置するページを全て集めれば良い。我々は、ある時点で収集されたページについて、こうした意味的な木構造が推定できれば、木構造の必要な部分のみを集めるという効率的な収集が可能となるのではないかと考えた。

2.2 収集アルゴリズム

意味的な木構造を推定する方法としては、様々な方法が考えられるが、本稿では、ページ間のリンク構造に着目して推定する方法を示す。これは、上階層へ向けたリンクの数は、下階層へ向けたリンクの数より多いという傾向を利用したものである。こうした傾向は、ユーザがインデックスページに戻り易いように配慮して、上階層へのリンクを多く張ることにより自然と現われるものである。以下に、あるアドレス r^0 の参照先ページとして集められた $R^0 = \{r_1^0, \dots, r_{n_1}^0\}$ について、木構造(親子関係)を推定するアルゴリズムを示す。

1. R^0 中の各ページについて、 R^0 中のページから参照される被参照回数 $\text{Ref}(r_i^0)$ をカウントする。
2. R^0 中の各ページについて、 $\text{Ref}(r_i^0) \leq T$ ならば r_i^0 は r^0 の子供、 $\text{Ref}(r_i^0) > T$ ならば r_i^0 は r^0 の親候補とする。ここに T は、閾値である。
3. 手順2.にて、親と推定された r_i^0 のうち、 r^0 を参照するページを r^0 の親と推定する。

本収集法は、ユーザが指定したアドレス r^0 から順に1

段づつページの意味的な親子関係を推定する。そして、その推定結果をもとに、木構造のうち収集の必要のある部分のみに絞込みながら収集を進めるものである。

3 実験

3.1 実験方法

実験にあたっては、三つのメーカーサイト A,B,C を対象とした。デジタルカメラの機種インデックスページ数は、それぞれ 5,4,5 ページ(これは、各メーカーの機種は約5種であることを示す)、各機種に関連する仕様、特徴のページは、それぞれ2~5ページあった。これらを合わせて正解ページとして、提案手法による収集をおこなった(各メーカーサイトの正解ページの総数は、それぞれ、14,34,14 ページであった)。収集にあたっては、初期 URL として一つの機種のインデックスページの URL が(各メーカーサイトについて)与えられることを前提にした。比較手法としては、初期 URL のページの参照先を全て集める手法をとりあげた(参照先を辿る回数としては、1,2,3 とし、それぞれについて実験を行った)。なお、親子判定に用いるリンク数の閾値は、今回は、集められたページの 1/3 が子と判定されるように定めた。

3.2 実験結果

上記条件のもとに、各手法について、収集したページの総数、収集した正解ページ数を調べた。この結果を図1, 2, 3に示す。図において、横軸は収集したページの総数、縦軸は、収集できた正解ページ数をそれぞれ表す。また、条件として、それぞれの収集ページにおける同一 URL は、再び収集しないとした。

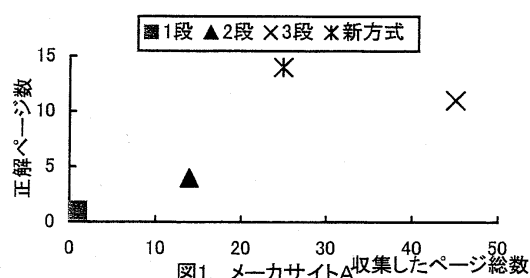


図1. メーカーサイトA

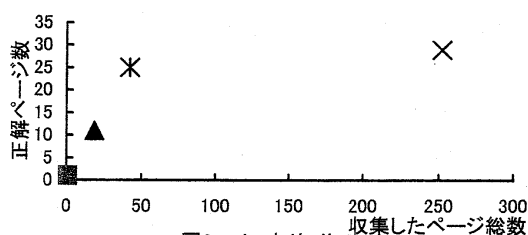


図2. メーカーサイトB

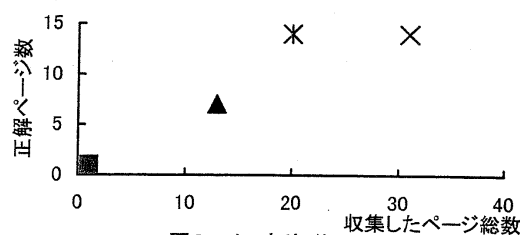


図3. メーカーサイトC

図において、参照先を全て集める手法(3 段)の収集効率(正解ページ数/収集ページ総数)は、平均27%である。これに対し、提案の手法の収集効率は、平均69%である。このように、提案の手法では、親子関係の推定に基づき収集先を絞り込むため、効率よく正解ページを収集できることがわかる。また、図 2 から本手法では、メーカーサイトBでは、正解テキストを全て収集できないことがわかる。これは、初期 URL より、上階層へ向けたリンクが無く親 URL の判定において、真の親を発見できなかったためである。このように、本手法は、全ての正解テキストを収集することを保証するものではない。しかし、収集時間に余裕がある場合には、リンク先を全て収集する方法と組み合わせることにより、全ての正解テキストを収集できる可能性がある。

4 おわりに

本稿では、ページのもつ意味的な階層構造に着目して、一連の情報を効率的に集める情報の収集法を提案した。そして、デジタルカメラの機種ページ及び、仕様、特徴のページを含む収集を例に、本手法の有効性を示した。今後は、様々な分野へ適用したときの有効性を調べる予定である。

参考文献

- [1] Kazunori Fujimoto and Kazumitsu Matsuzawa. Intelligent systems using web-pages as knowledge base for statistical decision making. New Generation Computing, Vol. 17, No. 4, pp. 349-358, 1999.
- [2] 藤本 他. Dsiu システム: Decision support for internet users 「ネット情報を使ってホットなものをあなたに!」. 人工知能学会論文誌(掲載予定), Vol. 15, No. 1, 2000.