参照ページからの情報を利用したWeb探索支援

板橋 英夫 (910008)

北陸先端科学技術大学院大学 情報科学研究科

2002年2月15日

キーワード: WWW, Webブラウジング, 情報探索支援, 参照関係.

本論文では，Web探索支援を目的とし，ユーザがアンカーをクリックしてアンカーの参照先のページ (以下，対象ページと呼ぶ) を表示する前に，対象ページの内容や第三者による多面的な評価をユーザに提示する手法を提案する．まず，対象ページを指すアンカーを持つページ (以下，参照ページと呼ぶ) を収集する．次に，参照ページから，対象ページに関する内容，意見，感想などを記述した文章の部分 (以下，参照箇所と呼ぶ) を抽出し，これらをユーザに提示する．参照箇所を提示することによって，ユーザが対象ページの有用性を知る手がかりになると考えられる．本論文では，参照箇所を抽出，分類する手法とその結果について報告する．

まず，参照箇所を抽出する手法を検討するために，インターネットから参照・被参照関係にあるWebページを収集した．その方法は以下の通りである．まず，検索エンジンにクエリを入れ，上位200件（満たない場合は最大数）を対象ページとした．次に，それぞれの対象ページにリンクをはっているページを参照ページとして収集した．但し，参照ページが10件以下の対象ページは除いた．これにより21ページの対象ページと582ページの参照ページを得た．これらのWebページを分析し，参照箇所を自動的に抽出する方法と，参照箇所にどのような情報が記述されているかについての分析を行った．

参照箇所抽出は，HTMLタグを手がかりにして行った．参照箇所抽出を試みる前に該当アンカー（対象ページへの参照ページからのリンク）がナビゲーション目的での参照であるかを判定した．具体的には，該当アンカー文字列が、「戻る」「トップへ」などとなっている場合は，ナビゲーション目的の参照とみなした．この場合，該当アンカーの周辺には対象ページに関する記述は存在しないことが多かったため，参照箇所の抽出は行わない．

次に参照箇所抽出を試みる．その方法は以下の4つに分けられる．

1. リストタグを手がかりとする場合
2. brタグを手がかりとする場合
3. テーブルタグを手がかりとする場合
4. その他

1，2，3 の場合は，それぞれリストタグ，br タグ，テーブルタグを用いて，アンカーとそのリンク先ページの説明が列挙されている参照ページを想定している．1 の場合，該当アンカーの直前の li タグから次の li タグまでを参照箇所として抽出した．2 の場合，アンカー+文字列+br というパターンが 3 回以上並んだとき，該当アンカーの次の文字列を参照箇所として抽出する．また，3 の場合，アンカーが同じ列に並べられ，該当アンカーが存在するセルの右のセルに参照箇所がある場合と，該当アンカーが存在するセルの下のセルに参照箇所がある場合がある．そこで、テーブル全体のレイアウトを判別し，参照箇所を抽出した．一方，1，2，3 のいずれのパターンにも当てはまらない場合，4 の場合には，該当アンカーの近傍を参照箇所として抽出する．このとき参照箇所の境界は HTML タグによって決める．具体的には，該当アンカーの前に存在する HTML タグを探し参照箇所の先頭とする．同様に，該当アンカーの後に存在する HTML タグを探し，参照箇所の末尾とする．但し，文字修飾タグ，イメージタグ，a タグ，コメントは無視し，参照箇所の境界としない．また，br タグについては，最初に出現したときは参照箇所の境界とせず，2 回目に現れたときは参照箇所の境界とする．

次に抽出した参照箇所の内容を分析し，どのような情報が含まれているかについて調査した．その結果，参照箇所は大きく分けて以下の 3 つのタイプに分類できることがわかった．

1. 説明タイプ
2. 意見タイプ（ページ型）
3. 意見タイプ（コンテンツ型）

1 の説明タイプは対象ページの内容を説明しているタイプである．意見タイプ（ページ型）は対象ページに対する意見を述べているタイプである．これは対象ページのレイアウトや雰囲気など対象ページに対する様々な意見が得られる．このような他者の客観的な意見は対象ページからは得られない情報であり，参照ページからの情報を収集することの利点である．3 のコンテンツ型は，対象ページそのものでなく，対象ページが紹介しているコンテンツに関する意見を述べているページである．このような対象ページのコンテンツに関する意見も，対象ページそのものからは得られない情報であり，ユーザに対象ページの有用性を判断させる重要な材料となる．複数の参照ページから得られた参照箇所を 1，2，3 のタイプに分類し，整理して提示すれば，ユーザも対象ページの内容を理解しやすくなるだろう．参照箇所のタイプを自動的に分類することは今後の課題である．

次に，提案した参照箇所抽出アルゴルズムの評価実験を行った．ここではクローズドテストとオープンテストの 2 種類の実験を行う．クローズドテストは先に述べた方法で収集した Web ページの集合を用いた．一方，オープンテストは，Web ページ作成のための素材を提供する「まゆ工房」と他者による商品の評価を掲載するページ「リブラ」の 2 つを対象ページとし，それぞれの参照ページの中から参照箇所を抽出した．参照ページの数はそれぞれ 86,40 である．これらのページは参照箇所抽出アルゴリズムの検討に用いていな

い．オープンテスト，クローズドテストとともに，人手で抽出した参照箇所を正解として評価を行った．その結果，クローズドテストでは完全一致で再現率 0.57，精度 0.49，部分一致で再現率 0.85，精度 0.82 を達成した．またオープンテストでは，完全一致で再現率 0.32，精度 0.35 を得た．部分一致では，再現率 0.62，精度 0.69 を得た．