JAIST Repository
https://dspace.jaist.ac.jp/

Title	第三者による解説・評価を含む関連リンク集の自動生成
Author(s)	平野, 健児
Citation	
Issue Date	2004-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1777
Rights	
Description	Supervisor:白井 清昭, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

第三者による解説・評価を含むWeb関連リンク集の自動生成

平野 健児 (210076)

北陸先端科学技術大学院大学 情報科学研究科

2004 年 2 月 13 日

キーワード: Web 探索支援，テキスト分類，リンク集.

本論文では，Web 探索を行うユーザが欲しい情報を簡単に Web から獲得するために，ユーザが要求するキーワードに関連した Web ページを収集し，関連リンク集を自動生成する事を目的とする．また，自動収集された Web ページを提示するだけでは，ユーザが一つ一つ Web ページを閲覧していかなければ，有用な情報がどこにあるのかわからない．そこで，各 Web ページに関する解説も提示する．本論文では，関連リンク集に掲載するにページ（以下，対象ページ）に関する解説は Web ページ自身から取り出すのではなく，その Web ページにリンクをはっているページ（以下，参照ページ）から取り出す．これにより，Web ページ自身からは得られない第三者による記述をユーザに提示することができる．また，Web ページに関する記述を説明・評価などに分類し，ユーザにわかりやすく提示する．また，関連リンク集をユーザに提示する際，掲載するページの収集・選別も重要であるが，本論文ではリンク先ページの説明・評価の記述に重点を置く．

本システムは以下のステップにより構成される．

1. ユーザーによるキーワード入力
2. キーワードに関するページの URL の収集
3. 参照ページからの対象ページに関する記述の抽出
4. 得られた記述の説明文・評価文などへの分類
5. 関連リンク集の出力

最初に，ステップ１～３を行う．具体的には，まずユーザーがキーワードを入力する．キーワードを Goo にかけ，キーワードに関する対象ページの URL を収集する．次に，各々の対象ページの参照ページを収集する．収集の方法は Goo のリンク先 URL 検索を用いて収集する．

次にステップ４について述べる，本論文では，HTML タグ及びサイト名を手掛かりとした二つの方法で Web ページに関する情報を抽出する．

Copyright © 2004 by Hirano Kenji

HTML タグを用いた方法は，リストタグ，br タグ，テーブルタグ等を手掛かりに，アンカー周辺にある Web ページの情報を抽出する．リストタグの場合，該当アンカーの前の li タグから，該当アンカーの後ろの li タグまでを Web ページに関する情報として抽出する．br タグは「アンカー＋文字列＋ br」が 3 回以上並んでいる場合，この文字列を Web ページに関する情報として抽出する．テーブルタグの場合，アンカーを含むセルの右側のセルにアンカー以外の文字列が記述されていれば，その文字列を参照箇所として取り出す．また，テーブルの同じ列にアンカーと文字列が交互にある場合，該当アンカーの下のセルにある文字列を Web ページに関する情報として抽出する．

サイト名を手掛かりとした方法は，まず，対象ページのサイト名を特定する．具体的には，サイト名として対象ページのアンカーの文字列を抽出する．ただし，長い文字列は，サイト名として抽出しない．また，サイト名を示す文字列は一つとは限らないので，複数のアンカーからサイト名を複数抽出する．次にサイト名を手掛かりとして Web ページに関する情報を抽出する．具体的には，サイト名の文字列の前にある HTML タグからサイト名の文字列の後ろにある HTML タグまでを Web ページに関する情報として抽出する．

次にステップ 5 について述べる．本論文では，参照ページから抽出した対象ページに関する記述を「評価：利便」「評価：情報量」「評価：その他」「説明：機能」「説明：記述」のカテゴリに自動的に分類する．カテゴリ「評価：利便」は Web ページに関する利便性，使い勝手といった記述を表す.「評価：情報量」は Web ページの規模，情報量を含む記述を表す.「評価：その他」は情報量と利便性以外の評価を含む記述を表わす.「説明：機能」は Web ページの機能について書かれた記述を表す.「説明：記述」はページの機能以外について書かれたページの説明を表わす．カテゴリの自動分類は,「説明：記述」を除く．それぞれのカテゴリに対してパターンとの人手で作成された．パターンマッチングにより行う．また，４つのカテゴリのいずれのパターンマッチにも失敗したときは，その記述のカテゴリを「説明：記述」とした．

本論文のシステムの評価実験を行った．参照ページから抽出された 467 個の記述に対し，カテゴリの自動分類を行った．カテゴリ [評価:利便] に対する精度は 0.8519，再現率: 0.766 であった．[評価:利便] の場合は，精度は 0.8519，再現率は 0.7692 であった．[評価:その他] の場合，精度は 0.3125，再現率は 0.4167 であった．[説明:機能] は精度 0.6442，再現率 0.7614 であった．[説明:記述] の場合は，精度 0.7797，再現率 0.7302 であった．また，全体の精度は 0.7410，再現率は 0.6617 であった．

Copyright © 2004 by Kenji Hirano