# WWWにおける関連リンク集の自動生成

田村 雅樹 (410080)

北陸先端科学技術大学院大学 情報科学研究科

2006年2月9日

**キーワード**: WWW，ポータルサイト，関連リンク集，キーワードの曖昧性，クラスタリング.

近年のWWWの普及により，ウェブ上で多種多様な情報を容易に入手できるようになった．また，誰もが簡単にウェブサイトを開設できるようになり，ウェブ上には膨大な情報が蓄積されている．しかし，WWWには様々な情報が無秩序に存在しており，有用な情報を探し出すのは困難である．このような背景の下，ウェブへのアクセスを支援する手法の1つにポータルサイトの利用がある．しかし，多種多様なユーザの要求に合ったポータルサイトがウェブ上に存在するとは限らない．したがって，ユーザの興味に応じてポータルサイトを自動的に構築することが望ましい．

本研究では，自動的に構築するポータルサイトのコンテンツの1つとして関連リンク集の自動生成を目指す．関連リンク集とは，あるテーマがいくつかのキーワードとして与えられたとき，そのテーマに関するページを自動的に収集し，リンク集として出力したものである．

リンク集生成の際にはキーワードの曖昧性に留意する．テーマとして与えられたキーワードが複数の意味をもつとき，ユーザがどの意味でそのキーワードを入力したのか判断できない．例えば，キーワードとして「松井」が与えられたとき，それが「松井秀喜」なのか「松井稼頭央」なのかそれ以外の誰か・何かなのか判断できない．そこで，本研究ではキーワードの意味の曖昧性を自動的に判断し，その意味ごとに関連するページを集めてリンク集を作成する．

本研究で提案するシステムの処理の流れは，(1) テーマの入力，(2) 候補ページの取得，(3) 候補ページの追加，(4) 不要なページの除去，(5) クラスタリング，(6) 出力の6ステップからなる．ステップ (1) ではユーザにテーマをキーワードとして入力してもらう．キーワードは1個以上の名詞である．ステップ (2) では検索エンジンGooを用いて，キーワードをクエリとしたウェブ検索を行い，上位500件をリンク集に掲載するページの候補として取得する．ステップ (3) では既存のリンク集のリンクを辿ることで，ステップ (2) では得られなかったがテーマと関連があると考えられるページを取得し，候補ページに加える．リンク集の判定はパターンマッチングにより行う．ステップ (4) では候補ページの中から

リンク集のみからなると考えられるものを削除する．これは，リンク集から別のリンク集へ飛び，更に別のリンクに飛ぶのはユーザにとって二度手間になるためである．ステップ (5) では従来のトピックをまとめるクラスタリングとは異なり，キーワードの曖昧性に着目して，同じ意味で使われるキーワードをまとめるクラスタリングを行う．このクラスタリングでは，まずキーワードの前後の名詞がキーワードの意味を表していると考え，キーワード前後の名詞が同じであるページをまとめる．そして，まとめられたページのうちページ数の多いものを基本クラスタとする．次に，基本クラスタと基本クラスタに属さなかったクラスタの間で類似度を計算し，それが閾値を超えた場合，そのページを基本クラスタに追加する．類似度はコサイン類似度を，また各ページやクラスタの単語ベクトルにはキーワードの前後 50 単語を用いる．単語の重みは TF 値に，クラスタを特徴付ける値として定義した ICF(Inverse Cluster Frequency) 値の積である TF-ICF 値を用いる．最後に，ステップ (6) ではステップ (5) で作成されたクラスタを元にリンク集を構築し，出力する．

上記の手法の評価実験を行ったところ，ステップ (3) やステップ (4) で行うリンク集の検出については適合率が 67.7%，再現率が 55.1%であった．ステップ (5) のクラスタリングで作成された基本クラスタには「松井秀喜」と「松井稼頭央」,「プロ野球」と「高校野球」のようにキーワードの意味をうまく反映しているものもあればそうでないものもあった．また，基本クラスタの精度 (クラスタ中のページのうちリンク集に掲載するページとして適切なページの割合) は 42.6%であった．これについて，ステップ (2) で得た初期の候補ページとステップ (3) で追加されたページとで分けて評価を行ったところ，精度はそれぞれ 49.4%，34.5%であった．更に，クラスタリング処理で基本クラスタに追加されたページについて，ページ中のキーワードが基本クラスタの定義するキーワードと同じ意味をもつかどうかの精度は 49.6%であった．ステップ (2) で得た初期の候補ページとステップ (3) で追加されたページを分けた場合の精度はそれぞれ 48.7%，50.6%であった．