可変スニペットとキーワード相関グラフを利用した検索補助インタフェースの提案

上條　朋彦†　　小山　聡†　　栗原　正仁†

北海道大学大学院　情報科学研究科†

1. はじめに

1.1 背景

近年, WEBページ総数の爆発的増加が進んでいる. すでにその数は80億を超え, 現在もなお増加し続けている. もはや人手による分類・利用は不可能であり, WWWからユーザの目的に応じた情報抽出・情報検索を行う研究が数多くなされてきた. 中でも現在, 最も普及し利用されているサービスが, GoogleやYahoo!に代表される検索エンジンである. これらはPageRankやHITSなどのランキングアルゴリズムを応用したもので, キーワードに関連するページを抽出する事に優れる.

しかし, 現在の検索エンジンにも問題点が残されている. ランキングアルゴリズムの特性もあり, ユーザごとの微妙な目的の違いはさほど加味されず, 一般的に著名なサイトを上位にランキングする傾向が強い. また, 目的の情報に辿り着かない場合, ユーザは検索ワードを修正して試行錯誤を行うが, 使用経験の浅いユーザにとっては試行錯誤する事自体が難しいといった問題がある.

1.2 検索エンジンのパーソナライズ

現状の検索エンジンの問題点を解決するため, 検索エンジンのパーソナライズを目的とした研究が行われ始めた. 検索結果の再ランキング手法やスニペット(検索結果に表示される要約文)のクラスタリング[2]や最適化[3], 検索ワード操作支援[4]など様々なアプローチの研究が進められている. すでに稼動しているサービスもあり, Googleパーソナライズ検索ツール各種やGoogleウェブ履歴, Yahoo! Rerankなどが存在する.

本研究では現状の検索エンジンの問題点に対して, 特に以下の2点に着眼し, これを実現する検索補助インタフェースの提案を目指す.

- リンク先を判断するための重要な要素であるスニペットは, 検索毎・目的毎の興味の度合いに応じて, 固定長ではなくユーザが操作可能な可変長である必要性
- 検索エンジンとユーザの目的との間で齟齬が発生した場合, 即座にそれを理解し検索ワードを修正できる視覚的情報掲示の必要性

2. 提案インタフェース

インタフェース構築の共通処理を説明する.

(1) ユーザが入力するキーワードの検索結果上位N件を検索エンジンから取得する.

(2) N件のリンク先の文書の中身を取得し, HTMLタグやスクリプトを除去した本文dの集合Dを作成する.

(3) 各本文を形態素解析し, 重要語となりやすい名詞・形容詞・形容動詞を抽出して単語集合Wを作成する.

(4) Wの各単語のTF・IDFを算出する. 文書$d(\in D)$中の単語$w(\in W)$のTF・IDFは以下の式で表される. ただし$f_{w,d}$は文書d中の単語wの出現回数, $n_{w,d}$は文書d中の総単語数, DF_wは単語wの出現する文書数$(\leq N)$である

$$TF \cdot IDF_{w,d} = \frac{f_{w,d}}{n_{w,d}} \cdot \log_2 \frac{N}{DF_w} \qquad (1)$$

2.1 可変スニペットインタフェースの構築

(1) 文書$d(\in D)$内のテキストを1文ごとに分割し, 文集合S_dを作成する.

(2) 文書$d(\in D)$中の各文$s(\in S_d)$に対して, 以下のスコアを与える

$$SCORE_{d,s} = \sum_{w \in s} f_{w,s} \cdot TF \cdot IDF_{w,d} \qquad (2)$$

$f_{w,s}$は, 文s中の単語wの出現回数である.

スコアの高い文sから順にスニペットに追加していく事で可変スニペットを生成する.

A User Interface for Search Engines with Variable-Length Snippets and Keyword-Relation Graphs

†Tomohiko KAMIJO, Satoshi OYAMA, Masahito KURIHARA, Graduate School of Information Science and Technology, Hokkaido University

2.2 キーワード相関グラフ部の構築

共通処理で算出済の文書d中の単語wにおけるTF・IDF値を降順にソートしたものを特徴語集合とし，これを利用してキーワード相関グラフを生成する．

キーワード相関グラフの生成手順

(1) グラフGにN個のノードを追加する．(これらを親ノードとする)
(2) 各親ノードに，特徴語集合からm個の単語ノードを追加する．(mはパラメータ)
(3) 追加したm個の単語ノードと親ノード間のエッジを作成する．単語ノードが他の親ノードに追加済であった場合，エッジのみを繋ぐ．
(4) グラフGを適切なグラフレイアウト手法で表示する

実装段階では$m=20$，グラフレイアウトにはKK法[1]を使用している．

3. 実装実験

計算機上に提案インタフェースの実装を行った．実装にはJava言語を使用し，形態素解析器sen，グラフ表示ライブラリJungを用いている．実際の動作画面は図1，2のようになる．(検索ワード「北海道大学」，$N=10$の場合)

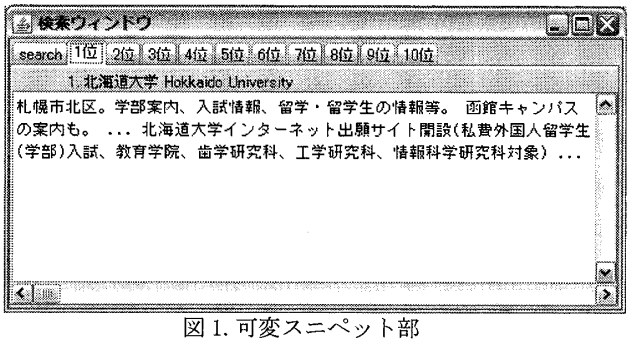

図1. 可変スニペット部

可変スニペット部(図1)では，ユーザがスクロールバーを調整することによりスニペットの長さを調節可能となっている．ユーザの興味の度合いに応じて情報を効率的に得る事が出来る．可変スニペットの表示内容も従来型とは異なる性質を持ち，リンク先全体を要約した内容を含んでいるため，ユーザはリンク先を開く前に目的情報と一致しているかを把握しやすい．図1の実装ではスニペット操作部にスクロールバーを採用したが，マウスホイールやマウスオーバーなどの様々な実装法が考えられ，場面に応じた表示が可能である．

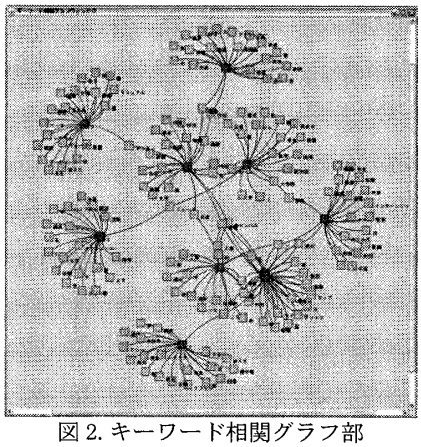

図2. キーワード相関グラフ部

キーワード相関グラフ部(図2)は，検索結果1件1件を表す親ノード(赤)の周辺に，内容を端的に表す単語ノード(灰)が配置される形となる．KK法によるグラフ配置により，関連語は距離的に近傍に配置され，多くの文書の共通重要単語(ピンク)ほど多くのエッジに繋がれ強調される．また，単語ノードをクリックする事で対応する単語を即座にAND検索に追加でき，検索ワードの修正支援にも繋がる．その他，マウスホイールによる拡大縮小機能も備えており，検索結果の直観的理解と検索ワード修正を支援する．

4. まとめ

検索エンジンの検索結果に対するパーソナライズ手法として，可変スニペットとキーワード相関グラフを用いた検索補助インタフェースモデルを提案し，計算機上へ実装を行った．今後の課題としては，インタフェース全体の適切な性能評価がある．本研究の応用として，Web検索エンジンだけでなくオフラインのドキュメント検索にも適用可能である．

参考文献

[1] T. Kamada and S. Kawai. "An Algorithm for Drawing General Undirected Graphs", Information Processing Letters, 31, 7-15, 1989.
[2] Ferragina, P. and Gulli, A. 2005. A personalized search engine based on web-snippet hierarchical clustering. In Special interest Tracks and Posters of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM, New York, NY, 801-810.
[3] 高見真也・田中克己．検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化．信学技報，vol. 107, no. 131, DE2007-69, pp. 283-288, 2007.7.
[4] 吉田 大我，小山 聡，中村 聡史，田中 克己，Web検索結果におけるキーワード出現相関の可視化と対話的な質問変換，DEWS2007，c7-2，2007.3.