Ｗｅｂ情報検索のための簡単な検索補助システム*

名古屋孝樹**　藤本直樹***　武富　敬****

A Simplified Support System for Web Information Retrieval

Kouki NAGOYA, Naoki FUJIMOTO and Hiroshi TAKETOMI

１．はじめに

Web ページ(以下、ページと略する)を検索していくうえで、より詳しい情報を知るためには、絞り込み検索は必要不可欠なものである。しかし、絞り込み検索に必要なキーワードは、どう選んだらよいか迷うことが多い。

そこで、ここでは、ある一つのキーワード入力で、それと関係するであろうキーワード候補を提示し、利用者がその中から取捨選択することで、絞り込み検索を手助けするシステムを提案する。本稿では、その実現手法と、開発したプロトタイプシステムについて報告する。さらに、関連するキーワードを選択する際の単語の重みづけの考察等、このシステムの高度化についても議論する。

２．システムの開発

2-1 システム全体

本システムの操作の流れは、図1のようになっている。

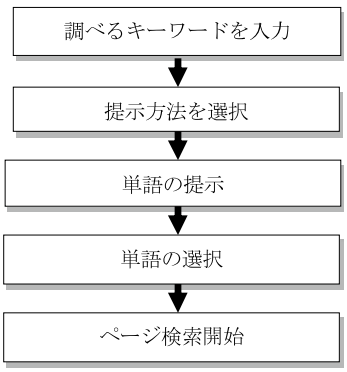


図１　本システムの操作の流れ

本システムでは、単一キーワードで直接ページ検索を行う前に、関係する単語を提示することで、効率のよいページ検索を目指す。

本システムのメインウィンドウを、図２に示す。

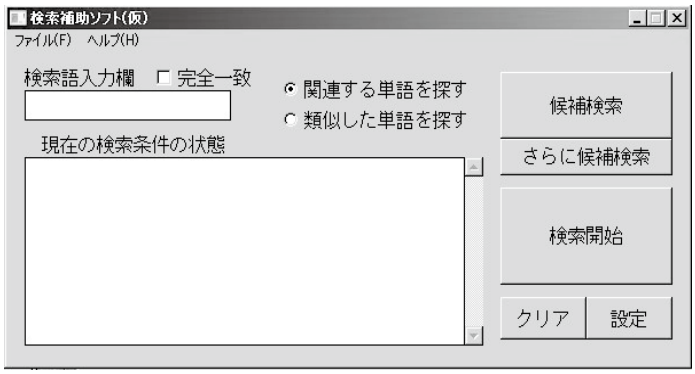


図２　メインウィンドウ

本システムを使用する場合、検索語入力欄に、ページ検索に用いるキーワードを入力して、候補検索ボタンを押下することで、単語の提示を開始する。

提示する単語には

・関連性のある単語

・類似した単語

の２種類があり、検索の目的に応じて単語の提示方法を使い分けることができる。

それぞれの提示方法の実現手法を、以下に示す。

2-2 関連する単語の提示

関連する単語の提示を行う場合、最初に入力されたキーワードで、ページ検索を行う。そして、検索結果から、該当したページの内、適当なページを３ページ、文書として取り込み、単語を切り出す。切り出した単語のうち、入力したキーワードと関連性のあると考えられるものをページごとに 10 個ずつ取り出し、さらにそれら 30 個の中で関連性の強いと考えられる 10 個の単語を、関連性のある単語として提示する。

＊　原稿受付　平成 21 年９月２５日

＊＊　信州大学理学部　数理・自然情報科学科

＊＊＊　佐世保工業高等専門学校　専攻科
　　電気電子工学専攻

＊＊＊＊　佐世保工業高等専門学校　電子制御工学科

図 3 に、単語の入力から、関連する単語が提示されるまでの流れを示す。

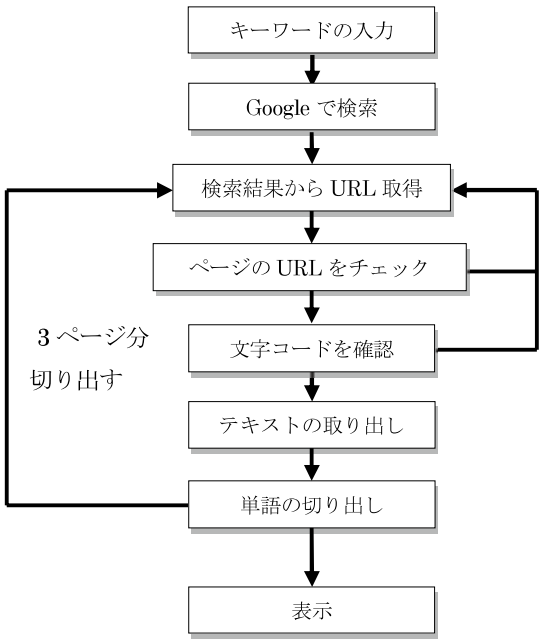


図 3　関連する単語を提示する過程

単語として切り出されるものは、文書のうち

・英文字、数字

・漢字、カタカナ

のどちらかで 2 文字以上連続しているものである。ひらがなは、単語と送り仮名が混じるので本システムでは無視される。

単語の関連性の強さは、単語のページ内での出現頻度を数値化することで判断される。出現頻度の数値は、表 1 に従って加算されていく。増加量は、試行した結果から定めた。

表 1　出現頻度の数値の増加表

条件	増加量
単語としての切り出し	単語の文字列の長さ
完全一致	(単語の文字列の長さ×2)^2
部分一致	(0〜一致する部分の長さ)×3

単語として切り出された時に、同じ単語が切り出されていない場合、新単語として、基本値として単語の文字列長さが与えられる。すでに同じ単語が切り出されていた場合、単語の文字列長さの 2 倍の二乗の値が、出現頻度として加えられる。また、既に切り出された単語と、部分的に一致した場合、まず、新単語として基本点を与えられ、一致した部分の長さだけ、システムの設定に従い、どちらかに加えられる。

ページ内の単語の切り出し、出現頻度の数値化が終了すると、その中で最も数値の高い 10 個の単語が切り出され、次のページの単語の切り出しが開始される。3 ページ分の切り出し・出現頻度の数値が完了すると、30 個の単語のうち、出現頻度の数値の高いほうから 10 個の単語を提示する。

図 4 に、関連する単語の提示画面を示す。

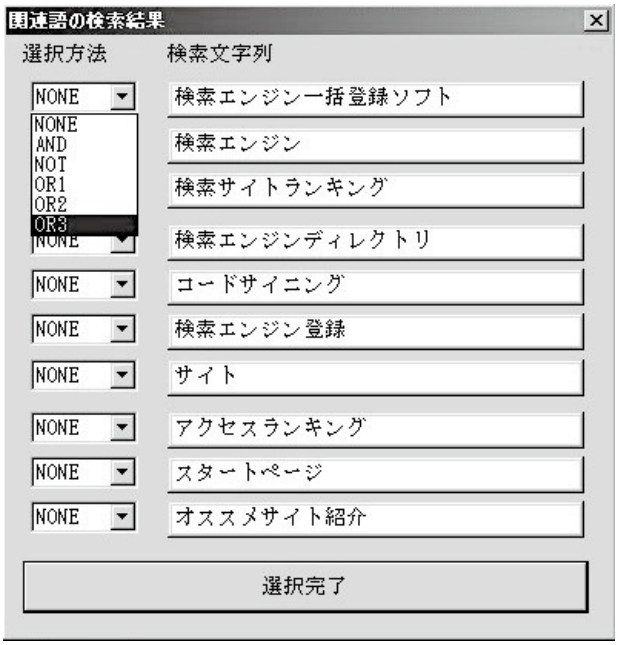


図 4　関連する単語の提示画面

図は、「検索サイト」に関連する単語の検索結果である。図のように提示された単語に対して、検索方法を決定して、関連する単語の選択を完了する。

検索方法には、NONE,AND,NOT,OR1,OR2,OR3 の 6 種類があり、それらは、次のような意味を持つ。

・NONE：検索条件に加えない。何もしない。

・AND　：通常の絞り込み検索と同じ。

・NOT　：その単語の含まれるページは検索結果から除かれる。

・OR　　：同じ番号の OR で選択された単語のいずれかが含まれるページを加える。

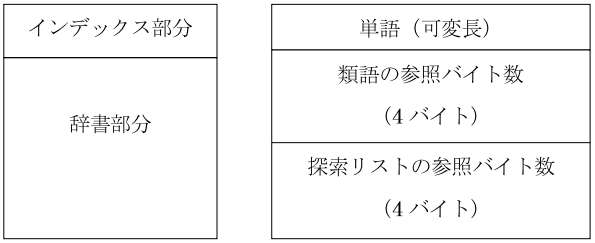
OR は、複数の単語のうち、いずれかの単語を含むページを検索結果に加えるものなので、一つだけ

OR で選択した場合、AND と同じ意味になる。

2-3 類語の提示

類語を求める場合、独自形式の類語辞書を用いることで、該当する単語を検索・提示する。

辞書全体の構造と単語単位での構造を、図 5 に示す。



(a)辞書全体の構造　　(b)単語単位での構造

図 5　類語辞書の構造

辞書内の単語はすべて、それぞれの類語リスト、探索リストの２つのリストに入っている。

類語を探索する際は、まず、入力した単語の１文字目を判別し、インデックス部分から、それぞれの文字に対応した探索リストで入力した単語を探す。複数の意味合いを持つ単語は、それぞれの意味をあらわす類語のグループに属している。そのため、複数のグループに属する単語を検索した場合、どの意味合いで検索したいのかを確認する必要がある。

表示する類語のグループの確認は、グループ内で先頭の単語を表示することで行う。提示させるグループを選択すると、類語の検索結果が一覧となって表示される。

類語の検索結果は図６のように表示される。

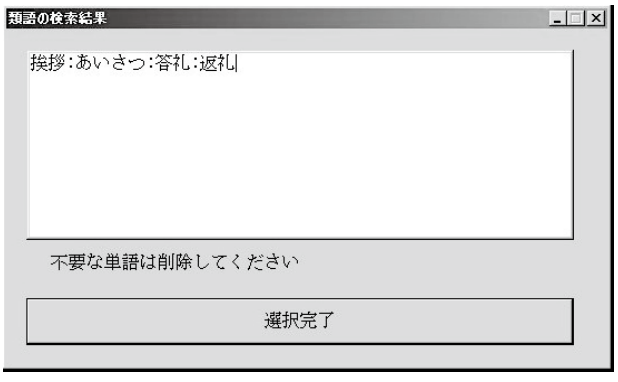


図６　類語の検索結果

図６は、「挨拶」の類語を検索したものである。この中から、ページ検索に不要な単語を除き、選択完了ボタンを押すことで、類語の提示を終了する。

残された単語はすべて、OR 検索でページ検索に用いられ、いずれかの単語が使われているページは検索結果に該当することになる。

2-4 絞り込み検索

提示された単語を取捨選択した後のメインウィンドウを、図７に示す。

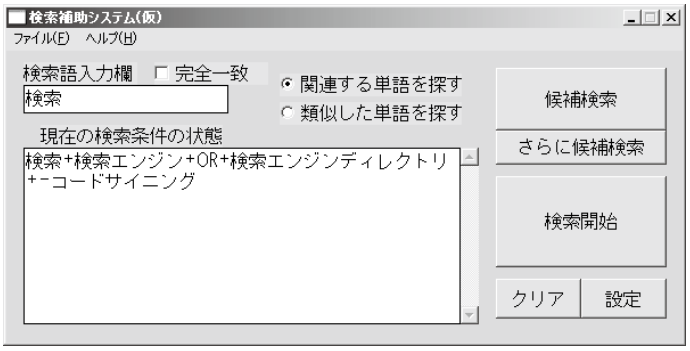


図７　単語を提示したメインウィンドウ

単語を取捨選択した後では、現在の検索条件の状態というスペースにその結果が反映される。ここで +は AND 検索を、+OR+は OR 検索を、-は NOT 検索を示す。

最後に検索開始ボタンを押下して Web ブラウザを呼び出すことで再検索(絞り込み検索)する。

３．システム評価

3-1 評価の必要性と方法

システム開発においては、システムの信頼性や有用性について考慮しなければならない。そこで、本システムの有用性や使いやすさの実証のために、テスト試行の実施によるシステム評価を行った。評価方法は、「シュナイダーマンのユーザインタフェース満足度評価シート　QUIS」[1]に基づく質問紙法による評価を採用した。

3-2 テスト試行とその結果

テスト試行は、以下の条件で実施した。

実施日時：2009 年 1 月 16 日　～1 月 23 日

被験者　：電子制御工学科 5 年生 12 名

16 日に簡単な使い方の説明を行って本システムを配布し、1 週間後に質問紙を配り、アンケートを実施した。

アンケートは、5 段階評価の質問 9 項目と、2 択の質問 1 項目、2 項目の自由回答からなる。

5 段階評価での集計結果を、表 2 に示す。

表 2　システム評価の集計結果

質問内容	平均値
提示される単語は適切か	3.58
提示される単語数は十分か	4.17
OR、NOT などの操作は役立つか	3.92
関連語の検索は役立つか	4.00
システム全体の使用感はよいか	3.25
システムの外観はよいか	3.75
システムの使用方法はわかりやすいか	3.25
各種設定は役立つか	3.67
Web ページの検索に役立つか	3.83

提示される単語は適切か、という質問や、システム全体の使用感はよいか、という質問の平均値は低いものとなった。このことから、まだ本システムには、改善の余地がある。Web ページの検索に役立つか、という質問の平均値が高いことから、本システムは、絞り込み検索に対し有効であることがいえる。

アンケートの自由回答欄では、ユーザインタフェース面に関する指摘が多かった。[2]

3-3 関連する単語の検索結果

関連する単語の検索は、切り出される単語の出現頻度から、その関連性の強さを調べるので、きちんと関連性のある単語を提示する場合と、全く関係のない単語を提示する場合とがあった。実際に関連する単語を、提示させた結果を示す。

(1) 成功例

検索ワード：ロボットタウン

提示された単語（関連度の強い順番）

・ロボットタウン
・ロボット
・カイトプレーン
・次世代ロボット共通プラットフォーム技術
・ロボットタウン実証実験公開
・スマートパル
・アイランドシティ
・ユニバーサルデザイン
・環境プラットフォーム
・次世代ロボット連携群

(2) 失敗例

検索ワード：近似曲線

提示された単語（関連度の強い順番）

・ブックマーク
・マイクロソフトセキュリティ
・ヘルプ
・エクセル
・近似曲線
・グラフ
・サイトマップ
・ブログ
・ドキュメント
・ログイン

失敗の原因としては、切り出したページの文書量が少なく、ページの構成に関わる単語の方が多く切り出されてしまったのではないかと考えられる。

4．考察

今回の研究では、システムを一通り完成させて、簡単なテスト試行を行うところまで進めることができた。また、システム評価の結果から、Web 情報検索における、この種の補助システムの有効性が確認できたと考えられる。

しかし、3-3 で示したように、現在の関連性の強さの指標では、提示される単語の精度に、重大な欠点がある場合があることが分かっている。入力したキーワードと関連性のある単語を、自動的な機械処理で、いかに精度よく提示できるかが本質的な問題点である。

これに関しては、以下の２点を、現システムの重大な問題点として挙げておく。

(1) 単語の切り出し

現在のシステムでは、漢字・カタカナ・英字で二文字以上連続したものを単語として認識している。そのため、ひらがなを含む単語や漢字一字の単語が抽出できない。

(2) 単語の重みづけ

単語同士の関連を考慮せずに、検索上位の限られたページから単純に出現回数で単語を抽出しているので、

・　出現回数が多いだけでキーワードと関連の低

い単語が提示される。
- 類似した単語同士がいくつも提示される。
- 限られたトピックの単語ばかりが提示される。

などの問題が起こることがある。

５．まとめと今後の課題

Web 情報検索において、絞り込み検索で有用な単語を、自動的に提示してくれる補助システムを提案し、そのプロトタイプシステムを開発した。12 人の利用者に対して実施したシステム評価において、ほぼ良好な評価を得ることができ、この種の補助システムの有効性を確認することができた。

しかし、4 で議論したように、提示する単語の精度については、重大な改良の余地が残されていることが示された。そこで示された２つの問題点について、今度の課題として、以下の方法で改良を進めたいと考えている。

(1)単語の切り出し[3]

これらの改善として、まず単語の切り出しには、既存のフリーソフトである形態素解析システムを組み込む。これにより、ひらがなを含む単語や漢字一字の単語の切り出しが行える。

(2)単語の重みづけ[3]

次の方法により、提示する単語の精度向上を目指す。

(a)TF・IDF 法の利用

出現回数と逆出現頻度を利用した TF・IDF 法を使用する。TF・IDF は、値が大きいほど数少ないページ中で集中的に出現していることになるため、より絞り込みに適した単語が得られると思われる。あるページ Di に対する単語 Wk の重みωは以下の式で求められる。

ω=TF・IDF

TF=(ページ Di における単語 Wk の出現頻度)

IDF=log(前ページ M/単語 Wk が現れるページ数)

(b)キーワードと単語の距離を求め、重みに加える。

(c)ページの分類

検索結果のページを、含まれている単語の類似度で分類し、分類されたグループを代表する単語を提示することで、多様なトピックスの単語を提示できるようにする。

謝辞

本稿の「まとめと今後の課題」において、議論にご協力いただいた、平成 21 年度卒業研究学生の堂前友貴さんに感謝の意を表します。

参考文献

[1]Ben Shneiderman:「ユーザインタフェースの設計」、(東基衛，井関治 共訳)，日経 BP，1987

[2] ユーザインタフェースに関する解説，URL:http://lab.mgmt.waseda.ac.jp/sdev-info/files/07_08_02_text_Usability.pdf

[3]林一成:「文章情報の可視化による検索絞込み支援」、奈良先端科学技術大学院大学　情報科学研究科修士論文、2000 年 3 月