Vol.2010-HCI-136 No.13
2010/1/22

ユーザに対する検索語喚起の支援を目指したWeb検索システムの開発

福島裕† 平川正人††

今日，インターネット利用において，キーワード検索は目的の情報を得るための手段として欠かせないものとなっている．ユーザにとってインターネットは検索から始まるといっても過言ではないだろう．しかしながら，従来のキーワード検索では入力したキーワードに極めて合致した結果を返すだけであり，目先の目的には効率はよいが，目的に合ったキーワードの選定を必要とする．本研究では，Web検索を行う上で的確なキーワードを持ちあわせていない，または発想出来ていないユーザ対して提供するべく，ユーザに新しい検索ワードの発想を促すと共に興味を喚起するような連想語句を選出するシステムを提案する．システムが提示するキーワードを利用し，従来のキーワード検索では達成出来ない，知的好奇心を高めるようなWeb 検索システムの開発について述べる．

Development of a web search system helping the user to have search keywords

Hiroshi Fukushima† and Masahito Hirakawa††

Today, keyword search on the web is indispensable to us as a means to get the desired information. However, in the traditional keyword search, words or web pages which match the exact keywords are retrieved. This means that the user is requested to choose appropriate keywords to have satisfactory results. In this paper we present a web search system, which helps the user who may not have a good knowledge of search keywords, by presenting associated words to step further into knowledge discovery.

1. はじめに

今日のインターネット利用において，キーワード検索は目的の情報を得るための手段として欠かせないものとなっている．ユーザにとってインターネットは検索から始まるといっても過言ではいないだろう．実際，Web検索には多くの利用手段が存在し，ユーザの好みにあった検索エンジンや検索サイトを選択し，利用することが出来る．その多くがキーワード検索という形式を取っており，何かしらユーザの目的にあったキーワードを入力することで，検索結果を返す形となっている．

しかしながら，一般的なキーワード検索では入力したキーワードを含んだ極めて限定された結果を返すだけであり，はっきりとした目的を持って検索作業を進める場合には都合がよいが，適切な検索結果を得るために目的に合致したキーワードを必要とする．適切なキーワードを喚起出来ないユーザにとっては前に進むことが出来ない．Web検索が日常的になってきている中，その問題は軽視出来ない．そこでユーザの入力したキーワードとの相違を少なくするため，検索サイトの多くでは検索結果と同時に様々な情報がユーザに提示されるようになっている．しかしながら，入力するキーワードの重要度は依然として極めて高く，検索におけるキーワードの重要性は変わっていない．

そこで最近では，関連語に重点を置くサービスが増えてきている．関連語とは，入力したキーワードに対して間接的に結びつくキーワードのことで，サービスの多くはその語句をさらに結びつけて検索結果の選択範囲を狭めることを目的としている．これにより，検索手順の簡略化，検索キーワードの選択補助という面での補完は叶えられている．ただしこれも検索結果が最初のキーワードとの結びつきが強い場合は効率が良いが，キーワードが定まらない場合，適切な結果は得ることができず，最初のキーワード入力における負担の軽減には至らない．

また，日常生活において検索キーワードの思い付きと言う行為が，実は非常に困難なことと考えられる．インターネットが日常的になる中で，キーワードが思いつかないということは新しい情報を得る方法の一つを断念していることになる．そのため，ユーザが普段から持つ探究心を向上させるような Web 検索の必要性があるのではないかと考える．

本研究では，有効な検索結果を得るための的確なキーワードを持っていない，または発想出来ていないユーザに対して検索の手がかりとなるキーワードを提供するべく，ユーザに新しい検索キーワードの発想を促し，またユーザの好奇心を喚起するような

† 島根大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Shimane University
†† 島根大学総合理工学部
Interdisciplinary Faculty of Science and Engineering, Shimane University

連想語を選出するシステム開発を行う．システムが提示するキーワードを利用して，従来の一般的なキーワード検索では達成が難しい，知的好奇心を高めるような Web 検索システムの実現を目指す．

2. 関連研究

Web 検索における研究は数多く行われてきているが，今回は以下の２つの視点から関連研究を整理する．

(1) 関連単語抽出アルゴリズムを用いた Web 検索クエリの生成

Web 検索では，検索エンジンによって得られた結果がユーザの必要としている情報ではないことがあり[1]，その解決策として検索エンジンに与えるクエリを改善するクエリ拡張がある．この研究ではセンテンス間の距離に注目した関連単語抽出アルゴリズムを利用し，それを適合性フィードバックの一手法である RSV と組み合わせることにより，検索精度の改善に役立てる方法を提案している．

特に関連単語抽出にセンテンス間の距離という情報を利用し各単語の評価を行っているが，この研究は用意された文章に対しての各単語の重みを算出しており，システムとして利用するには問題が残されている。また，本研究で目指すような好奇心を高めるような視点は含まれていない．

(2) 単語の類義関係を利用したクラスタリングと Web 検索

単語の共起情報を利用し，類義関係を表すシソーラスを構築することでユーザの検索クエリの作成補助を目的とする研究がある[2]．ここではクラスタリングに共起クエリと呼ばれる検索サイトのクエリ情報から得た検索語ランキング上位の共起情報を利用して行う．取得した共起情報で単語間の距離を算出し，階層型クラスタリングを行うことでユーザに提示するためのシソーラスを構築する．また構築したシソーラスを使用し，類義語を階層的にリスト表示するアプリケーションを開発している．

この研究は，単語に分類番号を付属させ，階層型クラスタリングを行うことでより高速に処理を行うことに重点を置いている．さらにその分類番号を利用することで，検索結果の中でさらに階層を辿ることも可能になっている．しかしながら，ユーザの検索クエリ生成に対しては元々ある共起情報の範囲から出ておらず，本研究で重視している新規の単語という点は考察されていない．また，シソーラスを構築してしまうと Web 環境で重要な時間軸が無視されるのではないかという考えのもと，本研究ではシソーラス構築は行わず検索結果の単語のみでクラスタリングを行う手法をとっている．

3. システム概要

3.1 システムの目的

１章で述べたように，一般的なキーワード検索における初期キーワードの選定を促すとともに，知的好奇心を高めるような結果をユーザに提供する検索システムの実現を目標とする．従来の検索結果とユーザの希望の相違を軽減させる，または従来の検索でユーザの希望にかなった，よりよい結果を求めるための補助となるキーワードを抽出しユーザに提供する．

3.2 検索方法

ユーザが日常においてイメージしている何かしらのキーワードを入力することで検索が始まる．このシステムで必要な最初のキーワードは，通常の一般的な検索システムの場合とは異なり，何かしらのキーワードという曖昧なものでかまわない．最初の検索キーワードの敷居を極力低く設定することで，新たな検索の分野を開拓しようとするものである．

3.3 WebAPI の利用

本研究では，検索エンジンの開発を行うのではなく，WebAPI を用いてシステム実装を行った．検索エンジンの開発を避けた理由として，検索エンジンの開発は作業時間が膨大になることや開発環境の負荷等の問題点が挙げられるが，近年 Web 上で優れた WebAPI が多数存在し，これを利用することで同じ機能を持ちながら大幅に開発行程を削減することが出来る．

本研究では reflexa[3]という検索サイトの WebAPI を用いる．このサイトは入力したキーワードに対して，本論文で呼ぶところの「連想語」をユーザに提供する．一つのキーワードに対して単語数も一般的な検索サイトで挙げられる関連語よりかなり多く，その数は３０～６０ほどになる．

関連語と連想語の違いであるが，関連語は，そのキーワードに対して加えることで，より検索結果の内容を制限することを目的としている．そのため，関連語だけでは意味をなさないことが多い．例えば，「検索」と入力した場合に「方法」や「エンジン」などが関連語となる．これに対して連想語は，元のキーワードが連想出来る単語である．例えば「検索」の場合，「サーチ」や「Google」などが相当する．関連語と連想語は重複する部分も多少は存在するが，基本的な目的や意味が異なっている．本研究では，ユーザが想像出来ないもの，また知的好奇心を刺激することが目的のため，連想語の選出を行う reflexa を利用するに至った．

また，本研究ではユーザに提供するキーワードを抽出する際に，ある程度のデータ量を必要とする．この reflexaWebAPI を利用すれば，キーワードとしては大量のデー

タを簡単に入手出来ると考えた．実際，最初のキーワードから得られる「連想語」の中から更に「reflexa」WebAPI を利用することで，必要数の連想語を入手出来る．

3.4 実装

システム実装にあたっては，WebAPI を利用するという観点から，内部処理の記述には PHP を用いた．またユーザに知的好奇心を与えることを目的としているため，GUI 部分はリッチインタフェースの作成が容易な Actionscript で実装している（図１）．

また本研究では，Web 検索システムという性質上，リアルタイム性を重視している．そのため，システム上にデータベースを持たず，都度 Web 上からデータを取得するという形を取ることでその性質を維持している．

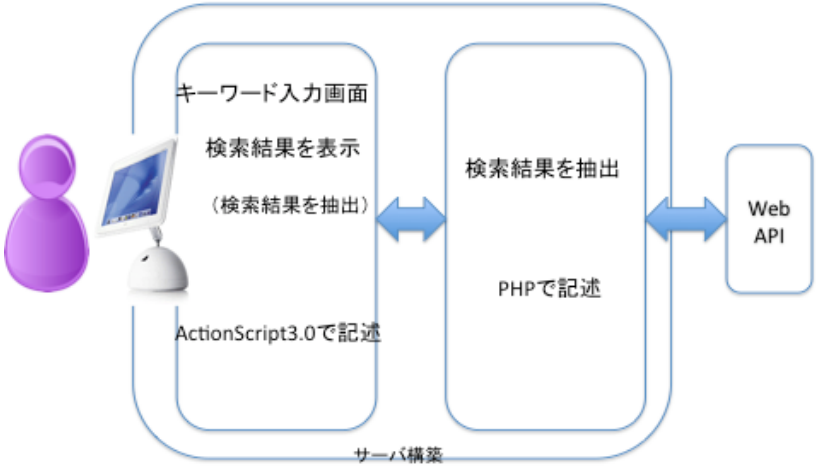

図 1　システム構成図
Figure 1　System diagram

3.5 画面構成

ユーザインタフェース設計にあたり，余計な画面遷移は極力避けることとし，検索結果をより対話的に，かつユーザの知的好奇心を刺激するようなインタフェースの実装を図る．キーワード入力部分と検索結果表示部分は一画面に集約し，表示することとした．実装にあたっては Flash を用いた(図２).

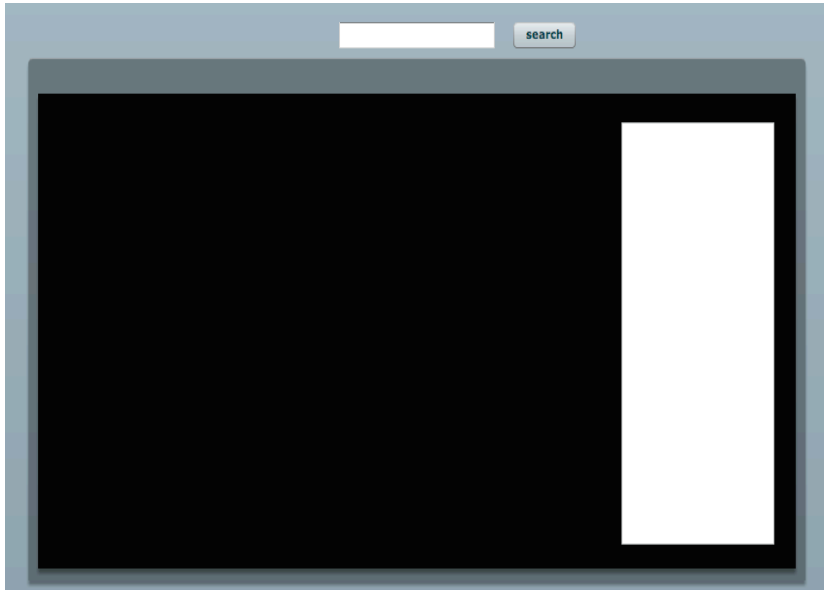

図 2　システムのメイン画面
Figure 2　Main screen of the system

4. キーワード抽出方法

4.1 キーワードの重要度設定

検索キーワードに対する単語や Web ページの抽出方法には様々な手法が存在する．特にページに対しての試みは数多くの方法が提案されてきている．しかし，本研究では検索キーワードのみに着目し，ページという概念をシステム内に取り入れることはしていない．そこで，reflexaAPI により Web から取得したデータを利用する．多数の単語からなる連想語のデータの集まりを大きなドキュメントの集合と見立て，そのデータの中から各単語の重要度を算出し集計することで，鍵となるキーワードの抽出を行う（図３）．

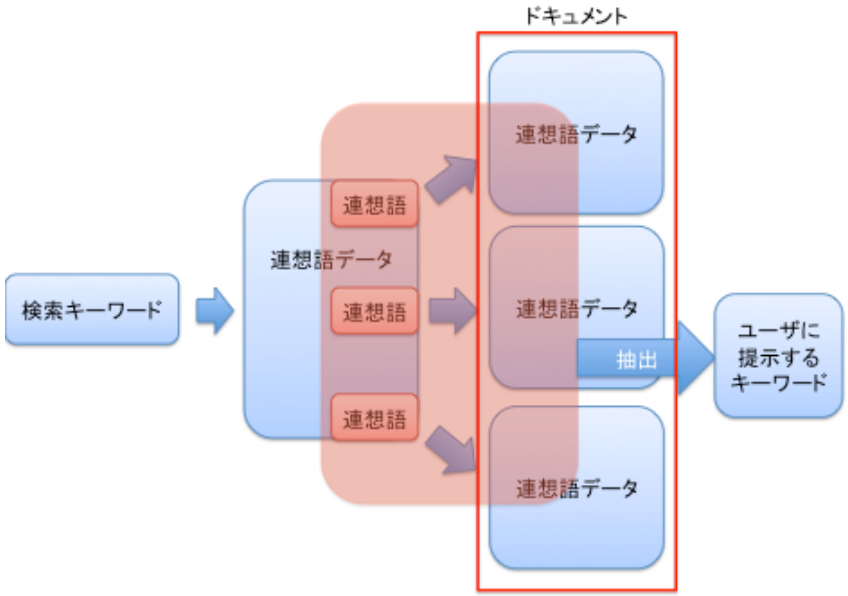

図 3　キーワード抽出方法
Figure 3　Keyword extraction method

本研究では TF-IDF 法に着目する[4]. TF-IDF 法[5]とは TF（単語の出現頻度）と IDF（逆出現頻度）の組み合わせで，文章中の特定の単語を抽出するためのアルゴリズムであり，以下の(1)のように定義される．

$$tfidf = tf \cdot idf \qquad (1)$$

TF-IDF 法では多数のドキュメントに出現する単語は重みが低く，出現率が低い単語は重みが高い値を算出する．TF-IDF 法において tf と idf は式(2)，(3)のように表される．式(2)は単語の頻度/出現する総単語数を表し，(3)は総文書数/単語の現れる度数となっている．n_i は単語 i の出現頻度，|D|は総ドキュメント数,d は総ドキュメント内の単語数である．

$$tf_i = \frac{n_i}{\sum_k n_k} \qquad (2)$$

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|} \qquad (3)$$

本研究ではドキュメントの代わりに最初のキーワードを TF で算出する.最初のキーワードで出てきた検索結果を複数のドキュメントと仮定し IDF によって算出し，最終的な重みを比較し分類する．

4.2　分類方法

検索キーワードから得た各単語の重みを算出後，重要度別に分類する．なお本研究では，重要度が高い単語の他に逆に低い単語も利用することで，ユーザに対して目新しさを誘発することを目指している．これは，日頃見慣れた物のわずかな部分をなじみのないものに置き換えることで，いつもとは違う目新しさを感じさせる手法[6]に依っている．

提案システムにおいては，通常の検索では除外する重要度の低い単語を，逆にユーザに提示するものとする．検索キーワードを探しているユーザにとって，重要度が低い単語は表示された直後は関係性がない無意味な単語と取られるが，他の検索結果と比較することで，意外性を含む単語としての利用を期待している．

5.　システム利用方法

5.1　利用手順

本研究で開発を行ったシステムについて，ユーザによる検索の流れは以下の通りである．

1．ユーザが入力フィールドに自由な検索キーワードを入力
　（最初のキーワードの重要度は低い）
2．このキーワードを基に検索を行い，データをシステムがクラスタリング
3．クラスタリング処理をした処理結果を画面に表示
4．さらに処理結果から，ユーザの操作により対話的に別の検索キーワードを表示

以下，これらの手順を詳しく説明する．

図４はキーワード入力画面である．画面中央上部の入力フィールドにキーワードを入力しボタンを押すことで検索が始まる．これは一般的な検索エンジンと同じ手順で

ある．

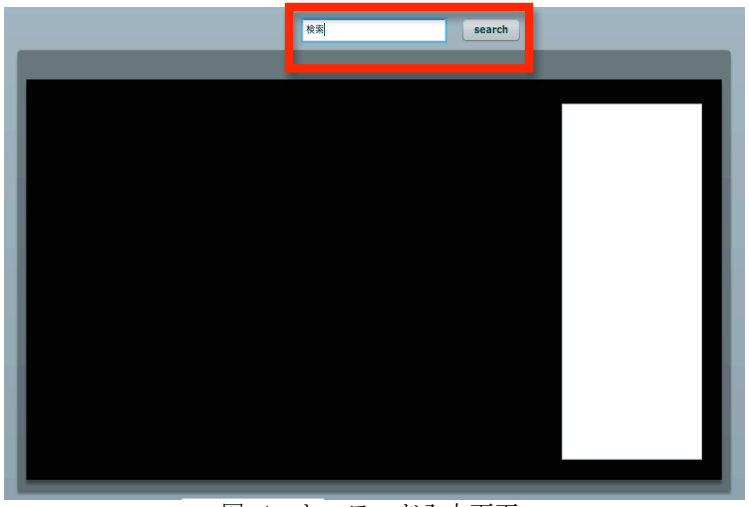

図 4　キーワード入力画面
Figure 4　Screen shot of entering keyword(s)

検索結果は図 5 のように表示される．表示結果は重要度に応じてフォントサイズが変えられる．また，結果の単語の関連によって画面空間での隣接関係が規定される．ここではユーザに対して視覚的に結果表示することで，知的好奇心を刺激することを狙っている．また，リスト表示も同時に行ない，クラスタリングを行った結果を重要順にソートし表示することで，ユーザの操作性を高めると同時に新しいキーワードでの検索を可能にする．

図 5　検索結果表示画面
Figure 5　Screen shot of displaying a result

6. おわりに

本研究では，ユーザが Web 検索を行う際，キーワードの想起・選択の補助を目的とした検索システムを提案した．

reflexaAPI を利用して多くの連想語を取得し，各単語の重要度を比較することで，ユーザに対して知的好奇心を刺激する検索結果を提供する可能性を示した．莫大な情報が Web 上に出回る現在，日常的に行われる Web 検索にとってキーワードは非常に重要な物であるが，不安定でもある．そのため，ユーザのキーワード喚起を補助出来れば非常に有用である．

また今後はユーザによる使用評価を行っていくことで，ユーザにとってのシステムの有用性を評価し得た内容からシステムの問題点や改良点について考慮し，より完成度の高いシステムを目指す予定である．

今回のシステム開発において，キーワードの抽出方法には DF-IDF 法を導入しているが，新規性のあるキーワードを更に信頼性高く提供出来る方法について検討が必要であろう． また表示方法においても不十分な点があると考え，検討を重ね改良していく．

謝辞　本研究を行うにあたって，議論および貴重な意見を頂いた平川研究室の諸氏

に深くお礼を申し上げるとともに慎んで感謝の意を表する.

参考文献

1) 大石哲也, 堀憲太郎:関連単語抽出アルゴリズムを用いた Web 検索クエリの生成,情報処理学会研究報告, Vol.2008,DBS-145,pp.33-40(2008).
2) 有田一平:検索語の共起情報を利用した単語クラスタリングと Web 検索への応用,電子情報通信学会技術研究報告,Vol.2007,NLC-107.pp.115-120(2007)
3) 連想検索エンジン「reflexa」:http://labs.preferred.jp/reflexa/ , 株式会社 Preferred Infrastructure.
4) 篷菜博哉,AdamJatown:Web からの文抽出と概念辞書を用いた概念間の関係発見支援,情報処理学会,DBS,Vol.146,pp.31-36(2008).
5) フリー百科事典「Wikipedia」:http://ja.wikipedia.org/wiki/.
6) 濱田芳治:新しさの作り方-形という記号が運ぶ意味-,多摩美術大学研究紀要,Vol.23, pp.61-69 (2008).