C-13

スニペットを利用したWeb検索システムの一提案
A proposal of Web search systems using the snippets

橋口　昂矢† 　長行　康男‡
Takaya Hashiguchi 　Yasuo Nagayuki

1. はじめに

インターネット上に存在する大量の情報の中から必要な情報を見つけ出すためのツールとして Web 検索エンジンがあり，現在，Google[1]，Yahoo![2]，百度[3]など，様々な Web 検索エンジンが広く利用されている．

これら既存の Web 検索エンジンは非常に便利なツールであり，多くのユーザが情報検索時にその恩恵を受けているが，ユーザが必要とする情報を『常にすぐ』見つけ出してくれるわけではない．

例えば，『日本で最も円定期預金の金利が高い銀行』の Web サイトを探したいとする（このような検索の目的を，以下では検索意図と呼ぶこととする）．2009 年 6 月 30 日現在，日本で最も円定期預金の金利が高い銀行は『新生銀行』であるが，検索クエリとして，［銀行　円定期預金　金利　日本一］，［銀行　高金利　定期預金］など，思いつく様々な検索キーワードの組み合わせを使って既存の Web 検索エンジンで検索を行っても，検索結果の上位に新生銀行のサイトは現れない．

その理由は，現在の Web 検索エンジンの仕組みにある．現在の Web 検索エンジンは，Web サイト内の文章から単語を拾い，その単語とその Web サイトの関連付けを行っている．そして，Web サイト内の文章に含まれる単語が，検索クエリ内に含まれる単語と一致した場合にのみ，その Web サイトが検索結果に表示される．さらに，検索クエリ内に含まれる，より多くの単語と一致することが検索結果の上位に表示されるための必要条件となっている．すなわち，新生銀行のサイトが，上述のような検索クエリで検索結果の上位に表示されるためには，新生銀行のサイト内に「日本で最も円定期預金の金利が高い銀行」といったことに関連した記述（正確には，より多くの検索クエリ内の単語との一致）がなければならない．しかし，新生銀行のサイトにはこのようなことは記述されていないため，検索結果の上位に新生銀行のサイトは表示されないのである．その他にも，例えば，『日本で一番学生数の多い大学』は現在，日本大学であるが，日本大学のサイト内には日本で学生数が一番多いことに関連する記述はない．また，『世界シェア 1 位の検索エンジン』は現在，Google であるが，Google のサイト内には世界シェア 1 位であることに関連する記述はない．したがって，［学生数日本一　大学］，［世界シェア 1 位　検索エンジン］などの検索クエリを使って Web 検索を行っても，日本大学のサイトや Google のサイトは検索結果の上位に表示されないのである．このような事例は他にも沢山にあると思われる．

ところで，「日本で最も円定期預金の金利が高い銀行が新生銀行である」ことや，「日本で一番学生数の多い大学が日本大学である」こと，「世界シェア 1 位の検索エンジンが Google である」ことなどは，電子掲示板，Wikipedia，比較サイト，ニュースサイト，解説・紹介サイト，blog サイトなどに記述されていることが多い．そして，上述のような検索クエリで Web 検索を行った場合，検索結果の上位にはこれらの Web サイトが多く表示される．その結果，検索結果画面上のスニペットには，『新生銀行』や『日本大学』，『Google』といった単語がいくつか出現する．我々は，この点に着目し，検索結果画面上のスニペットから単語をうまく抽出して，それらを検索クエリとして再検索を行うことにより，ユーザの検索意図を満たす Web サイト（上述の例の場合，新生銀行，日本大学，Google の Web サイト）をピックアップできるのではないかと考える．

そこで本研究では，まず，既存の Web 検索エンジンにおける検索結果のスニペットから単語を抽出する手法を提案する．そして，その抽出した単語を検索クエリとして再検索を行い，その検索結果を既存の Web 検索エンジンにおける検索結果に付加して出力する Web 検索システムを提案する．そして，既存の Web 検索エンジンでは見つけ出すことのできないユーザの検索意図を満たす Web サイトを，提案した Web 検索システムにより見つけ出す（検索結果に出力する）ことが可能であることを示す．

2. 既存のWeb検索エンジンにおける検索結果画面

本研究では，既存の Web 検索エンジンにおける検索結果の情報を利用する．そこで，既存の Web 検索エンジンにおける検索結果画面についてここで簡単に紹介しておく．

図 1 は，現在世界で最も広く利用されている Web 検索エンジンである Google で，［銀行　高金利　定期預金］という検索クエリを用いて検索を行った検索結果画面の一部である．検索結果画面には，デフォルトで１０件分の Web サイト情報が表示される．検索結果 1 件分の Web サイト情報は，タイトル，スニペット，URL などで構成される．ここでスニペットとは，その Web サイトの要約文のことで，その Web サイト中に存在する検索キーワードの周辺の文章などが表示される[4].

ここで，本稿で多用する“検索クエリ”と“検索キーワード”について定義しておく．本稿では，Web 検索エンジンで検索するときに検索ボックス（本稿では，［ ］で表す）に入力する全情報を“検索クエリ”と呼び，検

† 神戸電子専門学校　IT エキスパート学科
‡ 神戸情報大学院大学　情報技術研究科

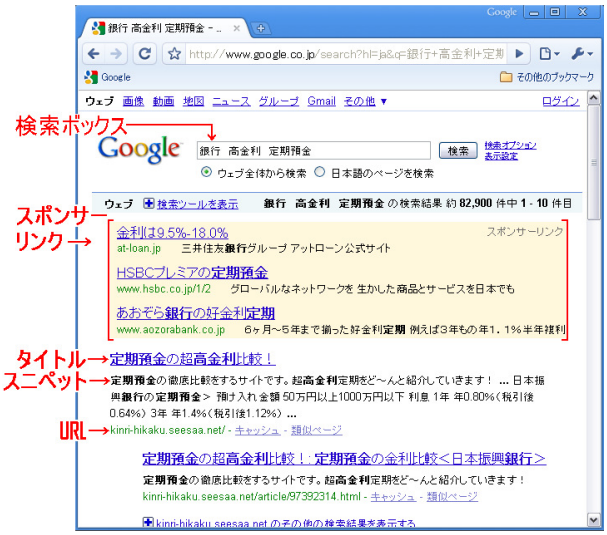

図１．Google の検索結果画面

索クエリ中の空白で区切られた個々の要素を“検索キーワード”と呼ぶことにする．例えば，検索ボックスに［銀行　高金利　定期預金］と入力した場合，検索クエリは『銀行　高金利　定期預金』で，検索キーワードは，『銀行』，『高金利』，『定期預金』のそれぞれである．

3. 提案システムの概要

本稿では，Web 検索結果のスニペットの情報を検索クエリとして利用した新たな Web 検索システムを提案する．提案する Web 検索システムにおける処理の手順を図２に示す．

図２中の（1）は，既存の Web 検索エンジンにおける検索クエリの入力と全く同様のものである．例えば，『日本で最も円定期預金の金利が高い銀行』を検索したい場合，［銀行　高金利　定期預金］などと入力する．

（2）では，（1）で入力された検索クエリを使用して既存の Web 検索エンジンで検索を行う．本研究では，既存の Web 検索エンジンとして Google を使用する．例えば，［銀行　高金利　定期預金］という検索クエリで検索を行った場合の検索結果は図１のようになる．

（3）では，検索結果のスニペットの情報をテキストデータとして取り出し，そのテキストデータから単語を抽出する．単語を抽出する方法としては，3.1 節で述べる，文字の種類の変わり目で単語に区切る手法を採用する．

（4）では，（3）で抽出した単語の中から検索キーワードを３つ選び出す．この検索キーワードを選出する方法は 3.2 節で述べる手法を用いる．（2）で，［銀行　高金利　定期預金］という検索クエリを用いて検索を行った結果から（3），(4）の処理を経由して選出される再検索用の検索キーワードは『キャンペーン』，『新生銀行』，『ネット』である．

（5）では，（4）で選出された３つの検索キーワードを用いて，既存の Web 検索エンジン（本研究では Google）により再検索を行う．ここで行う再検索は，３つの検索キーワードを組み合わせて行うのではなく，３つの検索キーワードそれぞれで行う．すなわち，３つの検索キーワードをそれぞれ独立の検索クエリとして扱う．例えば，３つの検索キーワードが『キャンペーン』，『新生銀行』，『ネット』の場合は，［キャンペーン］，［新生銀行］，［ネット］をそれぞれ検索クエリとして，独立に３回検索を行う．

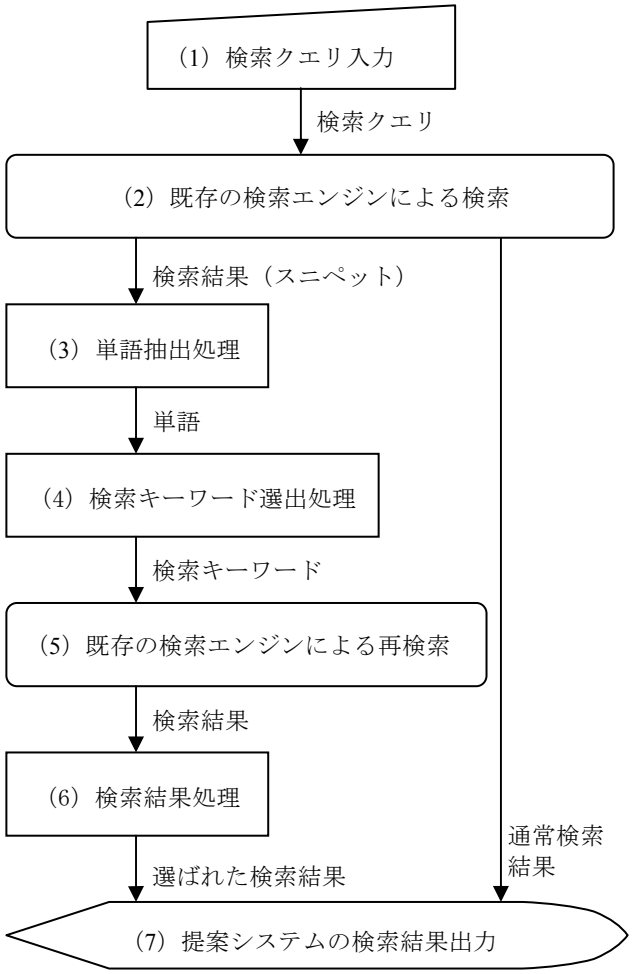

図２．提案システムにおける処理の手順

（6）では，（5）より得たそれぞれの検索結果から，それぞれの検索結果の最上位の１件分の情報（タイトル，スニペット，URL）を取り出す．そして，その取り出した情報を提案システムの検索結果（図２（7））に出力するかどうかを 3.3 節で述べる判断基準を基に決定する．上記の例の場合，［キャンペーン］，［ネット］の検索結果の最上位情報は出力しないと判断され，［新生銀行］の検索結果の最上位情報は出力すると判断される．

（7）では，（2）の通常検索結果のスポンサーリンクのデータを，（6）で提案システムの検索結果に出力すると判断した結果に置き換えたもの（スポンサーリンクがない場合は，最上位に（6）で出力すると判断した結果を付加したもの）を提案システムの検索結果として出力する．［銀行　高金利　定期預金］という検索クエリで検索を行った場合の提案システムの検索結果は図４のようになる．図１の Google の出力結果と比較すると，スポンサーリンクのデータが，ユーザが探している新生銀行のデータに置き換わっていることがわかる．

3.1　単語抽出処理

本研究では，検索結果のスニペットの情報から単語を抽出する（図２（3））ために，文字の種類（ひらがな，カタカナ，漢字，数字，アルファベット，句読点，特殊記号，空白など）を利用する．この文字の種類の変わり目で文章を区切り，その区切りで仕切られた要素を単語として認識する．例えば，「Google は多言語対応のサーチエンジンである。」という文章は，「(Google)(は)(多言語対応)(の)(サーチエンジン)(である)(。)」と分解され，それぞれの( )内の要素を単語として認識する．文章から単語を抽出する手法としては，その他にも形態素解析やYahoo!デベロッパーネットワークが提供しているキーフレーズ抽出などがあるが，本研究では，実験結果として安定した性能を示した，この文字の種類の変わり目で単語に区切る手法を採用した．

3.2　再検索用の検索キーワード選出処理

本研究では，再検索用の検索キーワードを選出する（図２（4））ために，まず，3.1 節の単語抽出処理により抽出された単語のリストを作成する．そして，この単語リストの中から，検索クエリとしては相応しくないと考えられる，助詞，接続詞，句読点，特殊記号などを取り除く．ここで特殊記号とは，「？」，「（」，「※」などの記号のことである．そして，残った単語の度数分布リストを作成し，そのリストを出現頻度の降順に整列する．そして，出現頻度の高いほうから３つの単語を再検索用の検索キーワードとして選出する．

3.3　再検索の検索結果処理

本研究では，Web 検索エンジンによる３回の再検索（図２（5））により得られた検索結果から，それぞれ最上位の１件分の情報（タイトル，スニペット，URL）を取り出し，その取り出した情報を提案システムの検索結果（図２（7））に出力するべきかどうかを次の判断基準で決定する．

まず，図２（1）で入力された検索クエリに対して形態素解析を行い，形態素を抽出する．そして，その形態素の集合から名詞のみを抜き出し，その名詞が，取り出した最上位１件分の情報のスニペットに含まれていれば提案システムの出力結果に出力し，含まれていなければ出力しない．

この処理を行う理由は，図２（3），（4）を経由して選出される検索キーワードには，もとの検索意図とは明らかに関係のないものが含まれることがあるからである．検索意図が「日本で最も円定期預金の金利が高い銀行を探したい」のときに選出された『キャンペーン』，『ネット』の再検索用の検索キーワード（3 章の（4）の説明文参照）などがそれにあたる．図２（1）で入力される検索クエリは，ユーザの求めている情報に関連する検索キーワードの組み合わせで構成されており，その検索キーワードの形態素の名詞がスニペット中に含まれていないということは，ユーザの求めている情報に適していない可能性が高いと考えられ，提案システムの検索結果に出力するべきでないと判断する．

図３．提案システムにおける検索クエリ入力画面

4.　提案システムの実装

図２で示した Web 検索システムを Java で実装した．検索クエリ入力画面（図３）は HTML で作成した．検索クエリ入力画面には，検索クエリを入力する検索ボックスと検索を実行するボタンがある．検索ボックスに検索クエリを入力し，検索ボタンをクリックすると Java サーブレットが起動し，図２に示した処理を実行し検索結果を HTML で出力する仕組みになっている．

図２（6）で使用する形態素解析には，形態素解析システム Sen[5]を使用した．なお，Web サーバには Apache HTTP Server を使用し，Java サーブレットを実行するためのサーブレットコンテナ（サーブレットエンジン）には Apache Tomcat を使用した．なお，Web サーバとサーブレットコンテナは連携させて稼動している．

5.　評価実験と実験結果

5.1　評価実験

提案した Web 検索システムを評価するための実験を行った．実験は，表１の３つの検索意図に基づいた検索クエリ（表２）を用いて行った．

表１．検索意図

実験	検索意図
実験１	日本で最も円定期預金の金利が高い銀行のサイトを閲覧したい
実験２	日本で一番学生数の多い大学を知りたい
実験３	世界シェア１位の検索エンジン知りたい

表２．検索クエリ

実験	検索クエリ
実験１	[銀行　高金利　定期預金]
実験２	[日本一　学生数　大学]
実験３	[世界シェア１位　検索エンジン]

5.2　実験結果

表２の検索クエリを，提案した Web 検索システムに入力した結果の出力画面は次の通りである．

図４．提案システムの検索結果画面（実験１）

図５．提案システムの検索結果画面（実験２）

図６．提案システムの検索結果画面（実験３）

１章で述べたように，実験１，実験２，実験３の検索意図を満たす Web サイトは，それぞれ，新生銀行，日本大学，Google の Web サイトである．

図４，図５より，実験１，実験２において，提案した Web 検索システムにより，ユーザの検索意図を満たす Web サイト（新生銀行，日本大学の Web サイト）をピンポイントに見つけ出し，通常検索結果（Google による検索結果）の上に表示できていることがわかる．これにより，通常検索結果を利用して，通常の Google の出力結果をチェックすることができ（通常の Google の出力結果に表示されるリンク先 Web サイトに，日本で最も円定期預金の金利が高い銀行が新生銀行であること，日本で一番学生数が多い大学が日本大学であることが記述されている），かつ，探している Web サイト（新生銀行，日本大学の Web サイト）へもすぐにアクセスできる．

図６より，実験３では，提案した再検索システムにより，ユーザの検索意図を満たす Web サイト（Google の Web サイト）は表示できているが，それ以外の Web サイトも表示されてしまっていることがわかる．これは，単語抽出処理（図２（3）），検索キーワード選出処理（図２（4））を経由して選出された再検索用の検索キーワードが，頻度が高い順に『位』，『Google』，『18』となり，これらを検索クエリとして再検索を行ったときの最上位に表示される Web サイト（「位階－Wikipedia」，「Google」，「青春 18 きっぷさかなのページ」）のスニペット中に，実験３の検索クエリ[世界シェア 1 位　検索エンジン] の形態素の名詞が含まれるからである．「位階－Wikipedia」のスニペット中には，形態素“位”が含まれ，「Google」のスニペット中には，“エンジン”，“検索”が含まれ，「青春 18 きっぷさかなのページ」のスニペット中には“1”が含まれる．これにより，３つとも提案システムの検索結果画面に出力する（ユーザが探している Web サイトかもしれない）と判断され，図６のような検索結果となっている．

6. おわりに

本研究では，Web 検索結果のスニペットの情報を検索クエリとして利用した新たな Web 検索システムを提案した．そして，既存の Web 検索エンジンでは見つけ出すことのできない，ユーザの検索意図を満たす Web サイトを，提案した Web 検索システムにより見つけ出すことが可能であることを示した．スペースの都合上，ユーザの意図を満たす Web サイトを見つけ出せない場合の結果等を取り上げていないが，（実験３を含め）うまくいかない例も多くあり，提案システムの性能を向上させることが今後の課題である．

参考文献

[1] Google：http://www.google.com/

[2] Yahoo!：http://www.yahoo.com/

[3] Baidu 百度：http://www.baidu.com/

[4] 小林竜己．“Web 検索エンジンの技術動向”，tokugikon，no.252，pp83-89 (2009)

[5] 形態素解析システム Sen：http://ultimania.org/sen/