社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

類似性を考慮したスニペットの再生成による検索結果のパーソナライズ

高見　真也†　　田中　克己†

† 京都大学大学院 情報学研究科 社会情報学専攻
〒 606-8501 京都市左京区吉田本町
E-mail: †{shie,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　ウェブ検索エンジンが返す検索結果のパーソナライズを行うために，ウェブページやスニペットの内容をもとにクラスタリングを行う研究はこれまでいくつか行われてきている．しかし，ウェブページに複数の話題が存在したり，偏ったセグメントがスニペットとして抽出されている場合は，精度上の問題が発生する．我々はユーザの意図に適したスニペットが提示されることで，検索結果の把握がすばやく行えるだけではなく，クラスタリングなどの精度も向上させることができると考えている．本論文では，ウェブページに関する視覚化された定量的評価を付加し，動的再生成が可能な改良型スニペット「Rich-Snippet」を紹介し，特定のスニペットに類似するように他のスニペットを再生成し，その類似度でリランキングを行うことにより，検索結果のパーソナライズを行う手法を提案する．
キーワード　パーソナライズ，スニペット

Personalizing Search Results by Re-generating Web-snippets by Similarity

Shinya TAKAMI† and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
E-mail: †{shie,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　For personalizing search results returned by web search engines, many researches focussed how to classify the search results by analyzing the context of each web page or web-snippet. There are, however, still quality problems in such approaches because some web pages have two or more topics and web-snippets are fragmentally extracted. We believe that we can guess the characteristics or the whole content of the web page quickly if the systems provide proper web-snippets for each user. It will also help to improve clustering quality. In this paper, we introduce a dynamically re-generatable web-snippet called 'Rich-Snippet' which visualizes some quantitative evaluation of the web page. We also propose a new generation method of web-snippets for personalizing search results which utilized similar web-snippets for re-ranking.
Key words　personalization, web-snippet

1. はじめに

インターネットを利用する人口の増加とウェブ上に散在する情報の多様化により，Google や Yahoo!に代表されるウェブ検索エンジンが返す結果は，ウェブ上で何らかの情報を探す際に大変重要な情報と見なされるようになってきた．我々は通常ウェブ検索エンジンに単語の組み合わせを検索語（以下，クエリ）として入力し，返された検索結果のうちごく限られた上位のものだけを対象に，目的とする情報が含まれていそうなウェブページを探す作業を繰り返し行っている．しかし，クエリの組み合わせでウェブページの重要度ランキングが一意に決定される検索結果では，多様な検索目的を持つすべてのユーザを満足させることは難しい．そこで，検索結果をパーソナライズすることにより，ウェブ情報検索支援を実現しようとする研究が行われている．

検索目的に適したウェブページ群のうち，どれか一つが発見できると目的が達成されるような場合は，正解ページのうち少なくとも一つがより上位にランクされることが重要である．一方，比較や調査目的でウェブ検索を行う場合のように，複数の正解ページを発見することが求められる場合は，正解ページか

どうかをすばやく認識できることが重要である．後者の場合，ランキングの精度を向上させるだけではなく，検索結果のクラスタリングやリランキングといった検索結果のパーソナライズを行うことで，ウェブ情報検索を支援することができる．ただし，ウェブページやその抜粋（以下，スニペット）の内容をもとに検索結果のクラスタリングを行う場合，ウェブページに複数の話題が存在したり，偏ったセグメントがスニペットとして抽出されていると，精度上の問題が発生する．

そこで，我々はウェブ情報検索の支援を行うために，ウェブページの内容を推測する際に重要視されるスニペットの改良に着目した．スニペットはウェブページを評価するための重要な手がかりであるにも関わらず，現行のものではその役割を十分に果たせていないように思われる．ユーザの意図に適したスニペットが提示されることで，検索結果の把握がすばやく行えるだけではなく，クラスタリングなどの精度も向上させることができる．本論文では，ウェブページに関する視覚化された定量的評価を付加し，動的再生成が可能な改良型スニペット「Rich-Snippet」を紹介し，特定のスニペットに類似するように他のスニペットを再生成し，その類似度でリランキングを行うことにより，検索結果のパーソナライズを行う手法を提案する．

2. ウェブ情報検索支援

ウェブ情報検索を支援するための手法として，様々なものが提案されている．ここでは，クエリ拡張およびクラスタリング，スニペットの改良に着目した手法を紹介する．

2.1 クエリ拡張とクラスタリング

ウェブ情報検索に関する研究分野では，HITS [1] や PageRank [2] といった優れたランキングアルゴリズムがいくつか提案されている．それらは，ハイパーリンクの構造解析による客観的な評価基準をもとに，あるクエリを含む数千，数万のウェブページ群から多くの人々が求めるものを上位にランキングする手法としては，十分価値のある結果を提供している．しかし，多くの場合，検索の目的はウェブページの URL リストを取得することではなく，あるウェブページ上に存在する何らかの情報を見つけることにある．そのため，ウェブ検索エンジンが返す結果の上位に含まれるウェブページ群が目的にそぐわない場合，目的のウェブページがより上位にランクされるように，クエリを再考し再検索が行われることが多い．そこで，クエリに追加または削減すべき単語の提案などを行うことで，ユーザの意図に適した検索結果を提供しようとする研究が行われている．

しかし，我々が Google の検索結果をもとに調査を行ったところ，クエリの拡張が必ずしも検索結果の絞り込みを実現する訳ではないことが分かった．ここでは京都で有名な湯豆腐料理を出す店のホームページを対象とした調査結果を紹介する．まず，クエリとして「京都＋湯豆腐」を入力し，主要な 8 店舗のホームページの順位を確認した．京都では，南禅寺周辺と嵐山／嵯峨野周辺に有名な湯豆腐料理店が存在する．そこで，各地域の湯豆腐料理店のホームページの順位が上昇することを期待して,「南禅寺」「嵐山」「嵯峨野」をクエリに追加して再検索を行った．表 1 は各単語がクエリに追加された場合の順位の変化

表 1　湯豆腐料理店ホームページの順位
Table 1　Website Ranking of Yudofu Restaurants

店舗	地域	基準	＋「南禅寺」	＋「嵐山」	＋「嵯峨野」
No.1	南禅寺	4	1[↑]	-	-
No.2	嵐山／嵯峨野	7	-	-	96[↓]
No.3	南禅寺	8	15[↓]	-	-
No.4	嵐山／嵯峨野	13	-	9[↑]	1[↑]
No.5	南禅寺	14	5[↑]	-	-
No.6	嵐山／嵯峨野	24	-	8[↑]	-
No.7	南禅寺	34	120[↓]	-	-
No.8	嵐山／嵯峨野	57	-	-	65[↓]

を示している．この例では，およそ半数のウェブページの順位は上昇したが，残りの半数の順位は下降している．つまり，クエリの拡張がユーザの目的によっては必ずしも有効ではないことを示している．

このように，クエリ拡張による再検索では，適合率は上昇するが再現率が低下する可能性がある．そこで，再検索は行わず，検索結果上位 k 件を対象にして，クラスタリングやリランキングを行うことで，ウェブ情報検索の支援を行おうとする研究が注目されている [3] [4] [5]．検索結果のクラスタリングは，対象とするものがウェブページかスニペットかで二種類に分類することができる．

ウェブページを対象としたクラスタリングの場合，各ウェブページ毎に特徴ベクトルを生成し，その類似度を評価する方法などが用いられる．しかし，近年のウェブページは，複数のブロックに種類の違うコンテンツが配置されていることも多く，またページの単位で話題が区切られているとは限らない．そのため，ウェブページに複数の話題が存在すると類似度が低くなってしまう可能性がある．また，スニペットを対象としたクラスタリングの場合，特徴を評価するには情報量が少なすぎるといった問題や，スニペットがウェブページのどのセグメントが抽出されたものであるかによって，精度が左右されるという問題がある [6] [7].

スニペットの各要素がウェブページのどこから抽出されたセグメントであるかは大変重要な情報である．なぜなら，スニペットの各要素に含まれるクエリを構成する単語（以下，クエリ単語）のウェブページ内での位置が，その意味や重要性に深く関係しているからである．ほとんどのスニペットはクエリを構成する単語を少なからず含むが，スニペットとしては抽出されていなくとも，他にそれらクエリ単語を含むセグメントが対象のウェブページには存在している可能性がある．さらに，複数のウェブページに類似したセグメントが存在したとしても，それらがスニペットとして抽出されなければ，クラスタリング時に類似しているとは見なされない．つまり，検索結果のクラスタリングを行う際には，ウェブページの場合は対象とする範囲，スニペットの場合はその生成手法に注意する必要がある．

2.2 スニペットの改良

株式会社アイレップ SEM 総合研究所らの「インターネットユーザの検索行動調査」[8] によると，ウェブ検索エンジンの利

用者が検索結果の中から実際にウェブページを確認するかどうかを決定する判断材料として，クリックする場合はタイトル，スニペットの順に，クリックしない場合はスニペット，タイトルの順に内容を確認する傾向にあることが報告されている．米国における同様の調査では，その順序が反対になっているが，それは大きな問題ではなく，ウェブページの特性を実際に開いて確認する前に判断する材料として，スニペットが重要な役割を果たしているということが示されていることに注目したい．このように，検索結果におけるスニペットはウェブページの特性を判断する際に重要な情報と見なされており，我々はそれらを改良することで，ウェブ情報検索を支援できるのではないかと考えている．

スニペットはウェブページから断片的に抽出されたセグメントにより構成されるのであって，必ずしも意味的に抽出されているわけではない．つまり，スニペットはクエリを構成する単語の組み合わせにより動的に生成されるため，概要文として見た場合，ウェブページの全体を包括する内容ではなく，ほんの限定された一部の内容だけを示している可能性がある [9]．また，断片的に抽出されたセグメントをウェブページ内での出現順に単純結合しただけのスニペットは，意味的なつながりを持たず一貫性に欠ける概要文となることが多い．

現存するウェブページの多くは，文字情報だけではなく，画像を含むマルチメディアコンテンツを含んでいる．HTML や XML の構造は，ときに文脈における重要性や意味に影響を与える場合がある．例えば，ウェブページ内での意味や重要性は，その単語がタイトル部分に存在するか，本文に使用されているかによって違うため，その特性を利用して詳細度の違う 2 つの単語の関係を抽出しようとする研究もある [10]．一方で，HTML などの構造化テキストであるウェブページから生成されるにも関わらず，スニペットは文字情報だけからなる．そのため，スニペットは人間が読む事によってのみ理解され得るコンテンツである．

このように，現行のスニペットはウェブページの特性を定量的には表現しておらず，クエリが決定されると一意に決定されるため，ユーザにとってはクエリ依存で静的な概要文である．また，人間がそれらを読むことでしか理解出来ない．しかし，各クエリ単語がウェブページのどこに存在しているかなどの情報を獲得する事はそれほど難しいわけではない．また，クエリに依存しない要約との違いを評価することで，著者の意図とユーザの目的の適合度を示すことができる．そのような情報は，ウェブページの特性を定量的に評価しており，我々がウェブページの全容を推測する際に役立つ．そのため，ウェブページに関する視覚化された定量的評価をスニペットに付加したり，クエリが同じ場合でもユーザの目的に適したスニペットを動的に提供することで，ユーザがウェブページの特性を推測する作業を支援することができると考えられる．

3. スニペットの生成手法

検索結果において，各ウェブページのタイトルおよびスニペットは重要なメタ情報である．また，クエリが同じ場合でも，スニペットが動的に提供されることでウェブ情報検索を支援することができる．ここでは，スニペットを生成する際に考慮すべき二つの視点を紹介し，さらにそれら二つの視点を統合することにより動的再生成が可能となった新たなスニペット生成手法を紹介する．

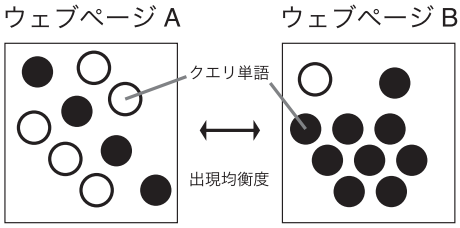

図 1　定量的評価：クエリの出現均衡度
Fig. 1　Quantitative Evaluation: Balance Level

3.1 クエリ依存型抽出

現行のスニペットの多くは，クエリがユーザの目的を表していると考えられることから，ウェブページから抽出されたクエリ単語を含むセグメントの組み合わせで構成されている．スニペットを生成する際には，キーワード抽出の手法を適用するために，各文を一つの単位として取り扱い，クエリに適合した重要文を抽出することが基本的なアプローチとされている．このような，クエリ適合度によって評価された重要文抽出によるアプローチは，ユーザの目的を表すクエリ依存型抽出手法によるスニペット生成と見なすことができる．

しかし，クエリ適合度だけを重み付け評価に用いた場合，いくつか品質上の問題が発生する．もしスニペットの各要素がウェブページ内の散在する箇所から抽出されたものである場合，それらを単純結合したスニペットは意味的に一貫性を持たない可能性が高くなる．また，ウェブページ内でクエリ単語が出現する箇所に偏りがある場合，クエリ単語を含むセグメントにより構成されたスニペットは，ウェブページ全体の特性を示すことが難しくなる．一方，Google などのウェブ検索エンジンのいくつかは，クエリ単語どうしの関係を考慮しないが，ウェブページとクエリの関係は定量的に評価されたとき，ウェブページの特性を示すことができる．ウェブページとクエリの関係を定量的に評価する基準はいくつか考えられるが，ここでは特徴的な三つの評価基準を紹介する．

図 1 は，ウェブページ内でのクエリの出現均衡度を示している．左側のウェブページでは，白丸と黒丸（これらはクエリ単語に相当）の数がほぼ均衡しているが，右側のウェブページでは，黒丸の方が圧倒的に多い．このような出現均衡度の違いは，それぞれのクエリ単語がウェブページ内でどのような位置づけにあるかに深く関係している．右側のようなウェブページの場合，白丸で表されるクエリ単語は，このウェブページの中では重要な扱いにはなっていない可能性が高い．このような関係は，単語の出現頻度（TF: Term Frequency）などの比較で評価することができる．

図 2 は，ウェブページ内でのクエリの出現分散度を示している．左側のウェブページでは，白丸も黒丸も散在しているが，

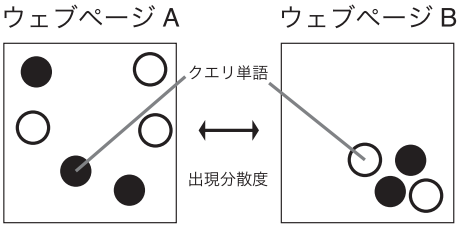

図 2　定量的評価：クエリの出現分散度
Fig. 2　Quantitative Evaluation: Decentralized Level

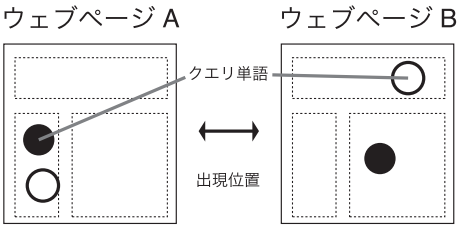

図 3　定量的評価：クエリの出現位置
Fig. 3　Quantitative Evaluation: Actual Location

右側のウェブページでは，一カ所に集中して出現している．このような出現分散度の違いは，ウェブページ全体に対する包括度を示している．もちろん，クエリ単語が局所的に出現しているからといって，そのクエリ単語が内容全体に関わっていないとは限らないが，クエリの出現分散度はこの後に説明するウェブページ構造との関係と併せて考察することにより重要な意味を持つ．クエリの出現分散度（$DL_{(}n)$）は，クエリ単語間におけるウェブページ内での距離を総和し，出現数で割る事により求められる（式 1）．ここでいう距離（d_{ij}）とは，ウェブページ内で二つのクエリ単語間（q_i, q_j）に存在する文字数と定義している．

$$DL_{(}n) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}}{{}_nC_2} \tag{1}$$

図 3 は，ウェブページ内でのクエリの出現位置を示している．先に述べたように，HTML などの構造化テキストでは，クエリがウェブページ内のどこに出現しているかが，その意味や重要性に深く影響を与える可能性がある．一般的なウェブページでは，メニュー部分と本文部分が分かれている事が多く，ウェブページ全体をいくつかのブロックに分割することができる．左側のウェブページの場合，白丸と黒丸は同じブロック内に存在するが，右側のウェブページの場合は，別々のブロックに存在している．右側のウェブページで考えた場合，黒丸が存在しているブロックが本文ブロックだと仮定すると，白丸はメニューブロックや広告やニュースなど頻繁に内容が変更されるブロックに含まれている．その場合，黒丸に比べて，白丸はウェブページ内で重要な位置付けにはなっていない可能性が高い．このような構造的な出現位置を単純な計算式で評価することは難しいが，Ruihua Song らはこのブロック特性を検索ランキングの精度向上に利用しようと試みている [11].

ここで挙げた三つの評価基準は，組み合わせて利用されることでより効果が大きくなると考えられるが，スニペットの分散度を評価した実験では，ウェブ検索エンジンにより抽出されたスニペットの各要素の分散度の違いが，ウェブページの内容の違いを示す可能性が高いことが確認されている [12]．このように，クエリ依存型抽出手法によるスニペット生成では，意味的にも空間的にも偏った重要文抽出であることを補完するために，クエリの出現分散度などで示されるウェブページの定量的評価に関する情報を加えることで，ウェブページの特性をより正確に表現することができる．

3.2　クエリ非依存型要約

スニペットはウェブページの要約文の一種であるが，多くの場合，それはクエリに強く依存している．ユーザの目的を表すクエリ適合度を評価基準とした重要文抽出と比較して，クエリに依存しない要約は，著者の意図を反映するものである．コンピュータによる要約生成に関する研究は，古くから行われており，様々な手法が提案されている [13] [14] [15]．要約をその機能をもとに分類すると，以下に示す二種類のタイプに分けることができる [16].

- 指示的（Indicative）
- 報知的（Informative）

指示的な要約とは，原文が読むべきものかどうか，自分の関心に合うかどうかなど，目的への適合性を判断するために原文を参照する前の段階で利用されるものである．また，報知的な要約とは，原文の代わりとして利用されるものである．そのため，スニペットは指示的な要約としての役割を担っているが，クエリ依存による手法では指示的，クエリ非依存による手法では報知的な側面が強いスニペットが生成されると考えることができる．コンピュータによる自動要約の多くは，様々な観点から文の重要度に重み付けを行い，ランキング上位の重要文を選択し，その出現順に並べることで要約が生成されるアプローチをとっている．抽象化または言い換えによる読みやすさの向上や文短縮によるアプローチなども試みられているが，本研究では，クエリ依存型抽出手法との共存も考慮し，指示的な要約としてのスニペットを対象としていることから，重要文抽出によるアプローチを要約生成手法として扱うことにする．重要度評価に用いられる情報には様々なものがあるが，奥村らはそれらを以下の六種類に分類している [16].

- テキスト中の単語の重要度
- テキスト中あるいは段落中での文の位置情報
- テキストのタイトル等の情報
- テキスト中の手がかり表現
- テキスト中の文あるいは単語間のつながりの情報
- テキスト中の文間の関係を解析したテキスト構造

H. P. Edmundson の実験によると，これらは位置情報，手がかり表現，単語の重要度の順で効果的であると報告されており，またそれらを組み合わせることで精度が向上されるとの結果が示されている [17]．ただし，その後の様々な検証により，対象となるテキストの種類によって，効果的な手法にばらつきが出ることも知られている．

クエリ非依存な要約としてのスニペットを生成する際，これらの研究成果は大変役立つものである．また，近年では対象を

ウェブページに特化したり [9] [18] [19]，クエリに関連する文に大きな重要度を与えることにより，クエリに重み付けされた要約の生成手法も提案されている [20] [21]．従来型要約手法では，テキストの内容をもとに要約は静的に決定できるという考え方で実現されてきたが，それはクエリによる重み付けを加えた場合でも基本的には変わっていない．しかし，対象となるテキストとクエリが同じ場合でも，検索目的の多様性から，指示的な要約としての効果をより高める為には，何らかの基準で一意に決定されるものではなく，ユーザの意図によって動的に変化する要約の方がウェブ情報検索には有効であると考えられる．

3.3 Rich-Snippet

前節までに述べた二つの視点は，スニペットを生成する際に考慮すべき重要な考え方であるが，それらはどちらも重要文抽出によるアプローチによって生成される点では共通している．また，現行のスニペットはクエリ依存な重要文抽出により生成されることが多いと考えられるが，クエリが同じ場合は同じスニペットが生成されるといった静的な基準で評価された概要文では，ユーザの多様な目的を満足させることは難しい．

そこで，我々は二つの視点を統合し，ユーザの意図により再生成可能な重要文抽出によるスニペットの生成手法を提案する．スニペット生成のために各文の重要度を評価する特徴ベクトルは以下のように生成する．

（1） ウェブページに存在する各文の重要度を計算し，特徴ベクトル v_t を生成（クエリ非依存型要約）

（2） クエリ単語を含む文に重み付けを与え，特徴ベクトル v_q を生成（クエリ依存型抽出）

（3） 特徴ベクトル v_t に対して，クエリ単語を含む文に α の重みを与えた特徴ベクトル v_t' を生成

（4） 特徴ベクトル v_t' および v_q を，β および $1-\beta$ の比率で統合（式 2）

$$v_{tq} = \beta \times v_t' + (1-\beta) \times v_q \tag{2}$$

まず，クエリ非依存な要約を生成するための特徴ベクトル（v_t）を生成する．文の重要度を評価する手法については，上で紹介した先行研究の成果を利用するものとして，ここでの説明は省略する．このプロセスは，クエリに関係なく生成可能なため，事前にシステム側で用意しておくことができる．次に，クエリが決定した後，クエリ依存な抽出を行うために，クエリ単語を含む文に重み付けを与えた特徴ベクトル（v_q）を計算する．このとき，出現数の多いクエリ単語や複数のクエリ単語を含む文の重要度を高くするなど，クエリの出現均衡度や分散度の評価も考慮して重み付けを行う．また，特徴ベクトル（v_t）に対して，クエリ単語を含む文に重みを与えた特徴ベクトル（v_t'）を計算する．最後に，二つの特徴ベクトル（v_t'，v_q）を統合し，総合的な特徴ベクトルを生成する．

スニペットは，統合された特徴ベクトル（v_{tq}）における重要度の高い文を規定数選択し，出現順に配置することで生成する．このとき，α および β を変動させることにより，各文の総合的な重要度が変化するため，生成されるスニペットが変化することになる．我々はこのように生成された改良型スニペットを「Rich-Snippet」と呼んでいる．

また，クエリ非依存な要約を生成するための特徴ベクトル（v_t）とクエリ依存な抽出を行うための特徴ベクトル（v_q）の類似度を計算することで，ウェブページに対するクエリの位置付けを評価することができる．本研究では，それをウェブページの主題を表すトピックに対するクエリの適合度と捉え，特徴ベクトル間のコサイン類似度をトピック-クエリ適合度（R_{tq}）として定義する（式 3）．

$$R_{tq} = \frac{v_t \cdot v_q}{||v_t|| \cdot ||v_q||} \qquad ||x|| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{3}$$

トピック-クエリ適合度は，ウェブページとクエリが決定されると一意に定義される評価値で，クエリで表現されたユーザの目的とウェブページ著者の意図とのずれを定量的に評価したものである．トピック-クエリ適合度が大きい場合は，クエリが含まれるセグメントがウェブページの主題に近いことを意味し，小さい場合は，複数の話題がウェブページに存在するか全体の内容としてはクエリを含むセグメントはあまり重要でない可能性が高いことを意味する．この評価値をもとに検索結果のリランキングを行うことができる．

我々が提案した手法により，システムは動的に再生成が可能なスニペットをユーザに提供することができる．二つの特徴ベクトルを統合する際の β 値を大きくすることで，よりトピック指向，つまり，クエリに依存しない要約としての側面が強いスニペットを生成することができる．逆に，β 値を小さくすることで，よりクエリ指向，つまり，クエリに依存したスニペットを生成することができる．そのため，クエリが決定された場合にクエリ依存型抽出手法により一意に生成される現行のスニペットと比べ，Rich-Snippet はユーザによる可変かつ動的なスニペットであるといえる．この特徴は，同じクエリが入力された場合でも，ユーザの意図する結果が同じであるとは限らないという問題を解決する可能性をもつ．

Rich-Snippet を利用することで，システムはトピック-クエリ適合度を評価基準に検索結果のリランキングを行うことができる．トピック-クエリ適合度は，ウェブページの著者が重要と考える部分とユーザが入力したクエリにとって重要な部分とのずれを定量的に評価したものである．また，検索結果のリランキングは，検索結果のパーソナライズにとって重要な要素である．ユーザによる検索結果の動的リランキングは，適合フィードバックによるクエリ拡張などを支援する．また，トピック-クエリ適合度を色表現で視覚化することにより，色の違いでウェブページの特性の違いを認識することができる．このように，Rich-Snippet はウェブページの定量的評価を視覚化することで，ウェブページの特性を推測する作業を支援する．

4. 類似性を考慮したスニペット生成

Rich-Snippet はユーザの意図に適したスニペットを動的に提供できるが，それぞれのウェブページの内容を把握することに主眼が置かれており，トピック-クエリ適合度の視覚化もクエリ

に依存した評価基準である．それらは，最初の正解ページをすばやく見つける場合には効果的であるが，複数の正解ページを見つける場合にはクエリに依存せず他との関係を考慮する手法も有効である．ここでは，特定のスニペットに類似するように他のスニペットを再生成し，類似度でリランキングを行うことで，ウェブ情報検索を支援する手法を提案する．

4.1 スニペット依存型生成手法

現行のスニペットの多くは，クエリ依存型抽出手法により生成されている．そのため，スニペットは少なくともクエリの一部を含んでいることが多い．しかし，我々が提案する Rich-Snippet による動的再生成が可能なスニペットでは，クエリ非依存な要約としてのスニペットも提供することができるため，ユーザの操作または意図によっては必ずしもクエリを含むとは限らない．例えば，京都の観光についてウェブ情報検索を行う場合，クエリに「京都」と「観光」を入力したとしても，ユーザが求める情報は「金閣寺」や「八坂神社」といったより具体化された情報である．このような場合，ユーザに適したスニペットとは，「観光」という単語を含んだセグメントではなく,「金閣寺」や「八坂神社」を含んだセグメントである．

複数の正解ページを求める必要がある場合，最初の正解ページが見つかったと仮定すると，そのウェブページと類似しているものは別の正解ページである可能性が高いと考えられる．しかし，ウェブページには不要な情報が混在していたり，複数の話題が含まれている可能性が高いことから，ウェブページ同士の類似度を評価するのは精度が悪い．そこで，スニペット同士の類似度を評価することになるが，クエリに依存したスニペットではスニペットに含まれるクエリ以外の単語によりその精度が左右される．このとき，Rich-Snippet を利用した場合でも，クエリ以外の重要な単語とは，ウェブページの要約を生成する際に主題を表すと判断されたものであり，その重要度はユーザの意図とは関係なく静的なものである．

そこで，我々はクエリ以外の重要な単語をユーザが指定するための方法として，特定のスニペットをユーザが選択することで，スニペットに含まれる単語を重要語とみなす方法を考案した．そして，選択されたスニペットに含まれる重要語をもとに，他のスニペットを再生成することにより，スニペット同士の類似度を向上させる手法を提案する．このようなスニペット依存型スニペットは以下のように生成する．

（1） 選択されたスニペットを形態素解析し，名詞および名詞化した形容詞または動詞だけを抽出

（2） ウェブページの各文を単位とした変形 TF-IDF 法により，抽出された単語 t の重要度 $w(t)$ を計算（式 4）

$$w(t) = \frac{\text{ウェブページに含まれる単語 } t \text{ の数}}{\log(\text{単語 } t \text{ を含む文の数})} \tag{4}$$

（3） 上記の重要語を含む文に重み付けを与え，特徴ベクトル v_s を生成

（4） 特徴ベクトル v_s から，重要文抽出手法にてスニペットを生成

スニペットが選択されると，スニペットとなるテキストを形態素解析し，重要語となる単語候補を抽出し重要度を計算する．各単語の重要度は，選択されたスニペットの元になるウェブページにおいて，各文を一つのドキュメントと見なした変形 TF-IDF 法により評価する．この手法により，出現数の多い単語だけではなく，局所的に出現している単語の重要度も高くすることができる．このように生成した重要語リストを，他のウェブページにおけるスニペットを生成する際に，クエリの代わりに重み付けとして利用する．このように，特定のスニペットに含まれる重要語に依存する重要文抽出手法で生成されたスニペットは，元になるスニペットに近い内容を含むセグメントが抽出されやすくなる．そのため，選択されたスニペットに類似するように，他のウェブページのスニペットを再生成することができる．

表 2　スニペットの分類
Table 2　Classification of Web-snippet

	Google	Rich-Snippet	提案手法
検索目的	考慮しない	主に到達型	主に収集型
定量評価	なし	トピック-クエリ適合度	類似度
生成手法	クエリ依存型	ハイブリッド型	非クエリ依存型
提示手法	静的	動的（単独）	動的（複数）

Daniel E. Rose らは，ウェブ情報検索の目的を，Informational/Resource/Navigational に分類している [22]．正解ページが一つ見つかれば目的が達成されるような場合を到達型（Navigational），複数の正解ページが求められる場合を収集型（Informational）に分類すると，Google によるスニペット，Rich-Snippet，類似したスニペットは表 2 のように分類できる．

4.2 スニペットの類似度とリランキング

類似したスニペット同士の類似度は，それぞれのスニペットから単語の特徴ベクトルを生成し，コサイン類似度等を計算することによって評価することができる．このような類似度評価は，Rich-Snippet におけるトピック-クエリ適合度を計算する場合と基本的には変わらないが，自動要約の研究分野においてシステムにより生成された要約の評価を行うために提案されているいくつかの指標も利用することができる [16]．ここでは，N-gram 適合率で評価する BLEU（式 5）とその改良版である ROUGE（式 6）について，その計算方法を以下に示す．

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right) \tag{5}$$

$$ROUGE(i,j) = \exp\left(\sum_{n=i}^{j} \frac{1}{(j-i+1)} \log C_n\right) \tag{6}$$

また，Rich-Snippet のトピック-クエリ適合度のように，選択したスニペットとの類似度を色表現で視覚化することにより，各ウェブページに対象となる情報が含まれている可能性をすばやく把握することができる．本研究における類似度評価は，クラスタリング等を行うために実施するものではなく，検索結果における他の正解ページのランキングを向上させることを目的としたものである．そのため，スニペットに含まれる単語をもとに作成した特徴ベクトル同士の類似度計算よりも，文章表現

表 3　スニペットの類似度評価
Table 3　Evaluation of Web-snippet Similarity

店舗	地域	基準	提案手法	＋「南禅寺」
No.1	南禅寺	1	-	-
No.3	南禅寺	0.000009934	0.000147170	0.000009934
No.5	南禅寺	0.000058588	0.000119231	0.000002636
No.7	南禅寺	0.000009659	0.000013599	0.000009164

としての類似性を評価することができるこれらの指標の方が本研究の目的には適している可能性が高い．また，正解ページのものと類似するように再生成されたスニペット同士の類似度をリランキングの基準に利用することによって，ウェブページ全体を対象とする場合よりも，より細かな基準でウェブページを評価することができる．そのため，我々はリランキングを行うことによって，複数の正解ページがより上位に再ランクされることを期待している．

4.3　評価と今後の課題

我々が提案する手法の妥当性を調べるために，先に紹介した湯豆腐料理店のホームページのうち，南禅寺周辺に存在する店舗を対象に類似度評価実験を行った．今回の実験では，類似度評価に特徴ベクトルによるコサイン類似度ではなく，自動要約の評価に用いられる BLEU を用いた．南禅寺周辺に存在する湯豆腐料理店のうち，順位が高い「No.1」のウェブページのスニペットを基準に，クエリが「京都＋湯豆腐」の場合に Google が出力するスニペット，我々の提案手法による類似したスニペット，クエリに「南禅寺」を追加した場合の Google スニペットについて，それぞれ類似度を計算した．その結果を表 3 に示す．この結果により，クエリ拡張により再検索された場合に再生成されるスニペットよりも，提案手法によるスニペットの方が類似度が高くなることが確認できた．

ウェブ検索エンジンにおいて同じクエリが入力されたとしても，検索結果において正解となるウェブページ群はその目的により異なるはずである．検索結果において指示的な要約としての役割を果たすスニペットが，各ユーザの目的に適したウェブページの摘要として提示されることで，検索結果の中から正解ページをすばやく発見することができると我々は考えている．近い将来，類似度でリランキングを行った際に正解ページを見つける時間が短くなるかどうかなど，ウェブ情報検索における支援効果に関する評価 [23] も行う予定である．

複数の正解ページを求めるウェブ情報検索において，比較等を目的としたものの場合，内容が若干異なっているものを集めたい場合がある．そのような似て非なる情報は，ある製品に対する評価を知りたい場合や，あるカテゴリに属する他社製品との比較を行いたい場合に重要である．このような検索目的の場合，類似したスニペットではなく，相違部分がスニペットとして提示されることで，ユーザはウェブページの内容をすばやく把握でき，分類されたものをさらにクラスタリングするような場合にも有効ではないかと考えている．このような高度な類似性をスニペット生成に取り入れ，より柔軟な検索結果のパーソナライズを実現することが今後の課題である．

5. 関 連 研 究

近年，ウェブ検索エンジンにより返された検索結果に着目した研究がいくつか行われている．それらのほとんどは，検索結果のパーソナライズを目的としたもので，関連研究としていくつか紹介する．Paolo Ferragina らは，スニペットの内容をもとに検索結果のクラスタリングを行おうとしている [6] [7]．しかし，既存のウェブ検索エンジンにより提供される現行のスニペットを扱うために，いくつか精度上の問題が報告されている．また，検索結果を類似度やコミュニティベースのスニペット・インデックスを利用してパーソナライズしようとする研究もある [24] [25]．これらは検索結果を分類するには有効な手法であるが，スニペットの再生成は考慮されていない．Yahoo! Research は，「Yahoo! Mindset」[26] と呼ばれる一種の意図指向型ウェブ検索インタフェースを提供している．彼らのシステムでは，調査または購買目的により検索結果をリランキングすることができる．このシステムもまた，検索結果の内容指向型動的リランキングを実現しているが，スニペットの機能は拡張されていない．

6. お わ り に

ウェブ検索エンジンを利用したウェブ情報検索において，検索結果のランキングがそのまま検索目的の適合度を表しているわけではないため，検索結果におけるスニペットは，ウェブページの特性を推測するための手がかりとして重要な情報である．しかし，既存のウェブ検索エンジンにより提供される現行のスニペットは，その役割を十分に果たせているとは思えない．そこで我々は，クエリに依存した現行のスニペットを改良する二つの手法を提案した．

一つは，クエリ依存型抽出手法とクエリ非依存型要約手法の統合により，ユーザが動的に再生成できる改良型スニペットである．さらに，ウェブページの定量的評価の一つであるトピック-クエリ適合度が色の違いで可視化されることで，対象となるウェブページの内容把握をすばやく行うことができ，その基準は検索結果のリランキングにも利用出来る．我々はこのようなスニペットを「Rich-Snippet」と呼んでいる．

もう一つは，「類似したスニペット」の生成である．複数の正解ページを発見することが求められるような場合，最初に見つけた正解ページと類似する内容のスニペットがあれば，そのウェブページも正解ページである可能性が高い．しかし，ウェブページ同士は類似した内容を含んでいても，それらがスニペットとして抽出されているとは限らない．そこで，ユーザが選択した特定のスニペットを基準に，そのスニペットと類似するようにスニペットを再生成する手法を提案した．最後に，提案手法により生成されたスニペットは既存のウェブ検索エンジンが出力するものよりも類似度が高くなっていることを実験で確認することができ，再生成されたスニペットをその類似度でリランキングすることにより，クラスタリングする場合と同様のウェブ情報検索支援効果が得られると期待している．

謝辞

本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己, 平成 14〜18 年度）ならびに，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウ ェア技術基盤の構築」における異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」における計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己, A01-00-02, 課題番号 18049041），計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073）によるものです．ここに記して謝意を表すものとします．

文　　献

[1] J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, J. ACM, **46**, 5, pp. 604–632 (1999).

[2] S. Brin and L. Page: “The anatomy of a large-scale hypertextual web search engine”, Proceedings of the seventh international conference on World Wide Web 7, Amsterdam, The Netherlands, The Netherlands, Elsevier Science Publishers B. V., pp. 107–117 (1998).

[3] M. A. Hearst and J. O. Pedersen: “Reexamining the cluster hypothesis: scatter/gather on retrieval results”, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-1996), New York, NY, USA, ACM Press, pp. 76–84 (1996).

[4] Y. Wang and M. Kitsuregawa: “Evaluating contents-link coupled web page clustering for web search results”, Proceedings of the eleventh international conference on Information and knowledge management (CIKM-2002), New York, NY, USA, ACM Press, pp. 499–506 (2002).

[5] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock and G. W. Flake: “Using web structure for classifying and describing web pages”, Proceedings of the 11th international conference on World Wide Web (WWW-2002), New York, NY, USA, ACM Press, pp. 562–569 (2002).

[6] P. Ferragina and A. Gulli: “A personalized search engine based on web-snippet hierarchical clustering”, Special interest tracks and posters of the 14th international conference on World Wide Web (WWW-2005), New York, NY, USA, ACM Press, pp. 801–810 (2005).

[7] F. Geraci, M. Pellegrini, P. Pisati and F. Sebastiani: “A scalable algorithm for high-quality clustering of web snippets”, Proceedings of the 2006 ACM symposium on Applied computing (SAC-2006), New York, NY, USA, ACM Press, pp. 1058–1062 (2006).

[8] 株式会社アイレップ SEM 総合研究所, 株式会社クロス・マーケティング：“インターネットユーザの検索行動調査”, Technical report (2006). Available as http://www.semirep.jp/info/20060626.pdf.

[9] E. Amitay and C. Paris: “Automatically summarising web sites: is there a way around it?”, Proceedings of the ninth international conference on Information and knowledge management (CIKM-2000), New York, NY, USA, ACM Press, pp. 173–179 (2000).

[10] S. Oyama and K. Tanaka: “Query modification by discovering topics from web page structures”, Proceedings of the 6th Asia-Pacific Web Conference (APWeb-2004), Vol. 3007 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 553–564 (2004).

[11] R. Song, H. Liu, J.-R. Wen and W.-Y. Ma: “Learning block importance models for web pages”, Proceedings of the 13th international conference on World Wide Web (WWW-2004), New York, NY, USA, ACM Press, pp. 203–211 (2004).

[12] S. Takami and K. Tanaka: “Quantitative evaluation of snippets returned by web search engines”, Proceedings of the Workshop of 1st Asian Semantic Web Conference (ASWC-2006) on Web Search Technology - from Search to Semantic Search, Jilin University Press, pp. 259–265 (2006).

[13] H. P. Luhn: “The automatic creation of literature abstracts”, IBM Journal of Research and Development, **2**, 2, pp. 159–165 (1958).

[14] C. D. Paice: “The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases”, Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR-1980), Kent, UK, UK, Butterworth & Co., pp. 172–191 (1981).

[15] G. Salton: “Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer”, Addison–Wesley (1989).

[16] 奥村学, 難波英嗣：“知の科学：テキスト自動要約”, オーム社 (2005).

[17] H. P. Edmundson: “New methods in automatic extracting”, J. ACM, **16**, 2, pp. 264–285 (1969).

[18] A. L. Berger and V. O. Mittal: “Ocelot: a system for summarizing web pages”, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-2000), New York, NY, USA, ACM Press, pp. 144–151 (2000).

[19] J.-Y. Delort: “Identifying commented passages of documents using implicit hyperlinks”, Proceedings of the seventeenth conference on Hypertext and hypermedia (HYPERTEXT-2006), New York, NY, USA, ACM Press, pp. 89–98 (2006).

[20] Y. Chali: “Generic and query-based text summarization using lexical cohesion”, Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence (AI-2002), London, UK, Springer-Verlag, pp. 293–302 (2002).

[21] H. Saggion, K. Bontcheva and H. Cunningham: “Robust generic and query-based summarisation”, Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL-2003), Morristown, NJ, USA, Association for Computational Linguistics, pp. 235–238 (2003).

[22] D. E. Rose and D. Levinson: “Understanding user goals in web search”, Proceedings of the 13th international conference on World Wide Web (WWW-2004), New York, NY, USA, ACM Press, pp. 13–19 (2004).

[23] T. F. Hand: “A proposal for task-based evaluation of text summarization systems”, Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization (ISTS-1997), pp. 31–36 (1997).

[24] M. Dontcheva, S. M. Drucker, G. Wade, D. Salesin and M. F. Cohen: “Summarizing personal web browsing sessions”, Proceedings of the 19th annual ACM symposium on User interface software and technology (UIST-2006), New York, NY, USA, ACM Press, pp. 115–124 (2006).

[25] O. Boydell and B. Smyth: “Community-based snippet-indexes for pseudo-anonymous personalization in web search”, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-2006), New York, NY, USA, ACM Press, pp. 617–618 (2006).

[26] Yahoo! Research: “Yahoo! Mindset”, http://mindset.research.yahoo.com/.