情報処理学会論文誌　Vol. 49　No. 4　1648–1656 (Apr. 2008)

ウェブ検索結果における検索目的に応じたスニペット生成

高　見　真　也[†1]　　田　中　克　己[†1]

ウェブ検索エンジンは，ウェブページを発見するためだけではなく，知識やサービスにアクセスするための道具としても使われるようになってきている．そのため，利用者が入力した検索語に応じて，広告コンテンツやサービスへの誘導リンクなども検索結果に表示されるようになった．しかし，検索語だけでますます多様化する利用者の検索目的を把握することは難しく，検索語で表示内容，表示順序が一意に決定される検索結果では，求める情報までの経路が最適化されているとはいえない．本論文では，内容の包括要約性および対象の集合依存性を軸として，検索結果として示すべき概要文（スニペット）を検索目的に対応したI型からIV型までのタイプに分類し，それぞれのスニペットを4種類の重要語により生成する手法を提案する．そして，最適化された検索結果を実現するためにスニペットの動的再生成を行う「Private View」を開発し，スニペット・タイプの違いが検索精度に与える影響について評価を行った．

Web-snippet Generation Suitable for Search Purpose in Web Search Results

Shinya Takami[†1] and Katsumi Tanaka[†1]

Web search engines are used as a tool not only to find web pages but also to access some knowledge and services. Therefore, the links to advertising contents and services etc. came to be displayed in the search results according to the search query that the user had input. However, it is difficult for such systems to know user's search purpose because it is more and more diversified. The route to target information is not necessarily optimized in the search results when the search query defines both the ranking and the content in the search results. In this paper, we classified the outline (Web-Snippet) in the search result into the type from the type I to the type IV by two criteria that are inclusive summary of content and group dependency of object. We propose the generation method of Web-Snippet by four kinds of important terms. And we developed "Private View" that can re-generate Web-Snippets dynamically and we evaluated the influence that the difference of the Web-Snippet types gave.

1. はじめに

インターネット利用者の増加にともない，一般の利用者が容易にコンテンツを制作できるようになったことで，膨大な情報がウェブ空間にあふれるようになった．本来，ウェブページを探すために利用されていたウェブ検索エンジンは，ウェブページの爆発的な増加により，登録制ディレクトリ検索型からロボット自動収集制キーワード検索型へと移行してきた．また，ウェブページを構成する素材が高度化したことで，利用者の検索目的も多様化し，ウェブページそのものではなく，京都にお寺はいくつあるのか，納豆はダイエットに効果的なのか，といった知識を得る目的でも使われるようになってきている．

マーケティング理論の分野でも，消費者の行動プロセスを5つのステージに分類したAIDMA（Attention → Interest → Desire → Memory → Action）モデルから，近年ではインターネットを意識し，購買前に検索を行い，購買後にブログなどのCGM（Consumer Generated Media）で情報を共有するAISAS（Attention → Interest → Search → Action → Share）モデルが提唱されている．このように，ウェブ情報検索はインターネット利用者にとって大変重要な利用目的の1つとなっている．

膨大な情報があふれる情報爆発時代では，ウェブ情報の断片化・再構成による新たな知識の創成はイノベーションの創出において重要な役割を果たす．なぜなら，そのような知識の創成技術の確立により，ウェブ情報検索がウェブ知識検索へと進化する可能性を秘めているからである．本研究では，ウェブ検索エンジンにより返される検索結果を，膨大なウェブ情報を断片化・再構成することによって生み出される新たな知識の1つととらえ，利用者の検索目的に応じた検索結果の最適化により，ウェブ情報の再構成による新たな知識創成を実現することで，ウェブ知識検索につながるウェブ情報検索が行えるようになることを期待している．

ウェブ上で何らかの情報を探す場合，我々は通常ウェブ検索エンジンに検索語の組合せを検索質問として入力し，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．一般にデータベースへの問合せを検索質問と呼ぶが，本論文では情報検索の場合に限定し，ウェブ

†1 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

© 2008 Information Processing Society of Japan

検索エンジンに与える検索語の組合せを検索質問と呼ぶことにする．多くのウェブ検索エンジンは，検索結果として，タイトル，URL および概要文（スニペット）を含むウェブページのリストを返す．そのようなシステムにおいて，検索質問によく適合するウェブページが検索結果の上位に順位付けられることはもちろん重要であるが，たとえまったく同じ検索質問が入力されたとしても，その目的により，システムが返すべきスニペットは同じであるとは限らない．そこで，我々は検索目的に応じたスニペットの提供により検索結果の最適化を実現することで，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．

2. 検索結果の最適化

2.1 質問修正・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，利用者により入力された検索語を含む数千，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順位付けする手法としては，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順付けされるように，検索質問を再考し再検索が行われることが多い．そこで，検索質問に追加または削減すべき単語の提案などを行うことで，利用者の検索目的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，検索結果における目的のウェブページの順位は検索質問の修正や拡張によって必ずしも上昇するわけではないことが分かった[3]．

また，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[4] 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[5),6]．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 2 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，各ウェブページごと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こから抽出された断片の組合せであるかによって，精度が左右されるという問題がある[7),8]．

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断片であるかは大変重要な情報である．なぜなら，同じ検索語を含む断片でも，各断片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検索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ても，他にそれら検索語を含む断片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断片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精度は，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内容に大きく依存している．

2.2 スニペットの改良

ウェブ検索エンジンにより提供されている現行のスニペットには，いくつかの問題がある．現行のウェブ検索エンジンにより生成されるスニペットの多くは，ウェブページから断片的に抽出された検索語を含むテキス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，スニペットは検索語の組合せに依存して生成されるため，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テキス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．たとえば，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 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かかわらず，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ことによってのみ理解されうるコンテンツである．

このように，現行のスニペットは検索質問が決定されると一意に決定されるため，利用者にとっては検索質問依存で静的な概要文である．また，ウェブページの特性を定量的には表現しておらず，人間がそれらを読むことでしか理解できない．しかし，各検索語がウェブページのどこにどれだけ存在しているのか，スニペットに含まれる各断片はウェブページのどこから抽出されたものなのか，などの情報を獲得することはそれほど難しいわけでは

© 2008 Information Processing Society of Japan

ない．そのような情報は，ウェブページの特性を定量的に表現することができ，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，ウェブページやスニペットに対する視覚化された定量的評価を提示したり，検索質問が同じ場合でも利用者の検索目的に適したスニペットを動的に提供したりすることで，検索結果の最適化が実現され，利用者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我々は考えている．

3. スニペットの生成方法による分類

本来，図書館における検索の目的は本を探すことであり，様々な動機があるにせよ検索対象は本であった．検索を行うための情報も本の内容すべてが対象になっているわけではなく，著者名や内容の一部などのメタデータを利用している．しかし，ウェブページはその内容すべてに容易にアクセスでき，情報提供以上の価値を発信するようになったことで，ウェブ検索エンジンはウェブページを探すためだけの道具ではなくなってきている．ほとんどの場合，その検索対象はウェブページであるが，多くの場合そのウェブページ上に存在する情報の一部やサービスを利用するために検索が行われている．検索目的は，知識の獲得とサービスの利用に大別されるが，我々は前者の目的のウェブ情報検索を支援することに注目した．

知識の獲得が検索目的の場合，目的の知識を含むウェブページを提示できれば最適な検索結果となる．しかし，目的の知識を含むかどうかをシステムが判断することはきわめて難しい．そこで，ウェブ検索エンジンは，目的の知識を含むかどうかを利用者の判断に委ねるために，タイトルやスニペットといった判断材料を検索結果として提供している．つまり，検索結果は利用者の主観的な基準により判断されており，そのような主観的な判断を支援することが本研究の目的である．

我々は利用者の検索目的に適した検索結果を提供するために，スニペットをその生成方法の違いにより 2 種類の軸で分類した．1 つ目の軸は，ウェブページからスニペットを生成する際に，内容の包括要約性を考慮するかどうかである．断片集約型は特定の単語，たとえば，検索語を含むといった何らかの基準で抽出された断片をまとめてスニペットにするタイプで，包括要約型は多種多様な単語を含み，全体の内容を包括する要約を意識したスニペットを生成するタイプである（**図 1**）．

もう 1 つの軸は，スニペットを生成する際に考慮するウェブページの数である．単体独立型は 1 つのウェブページから得られる情報だけで生成するタイプで，集合依存型は検索結果などに含まれる他の複数のウェブページ集合から得られる情報を考慮して生成するタイプである（**図 2**）．

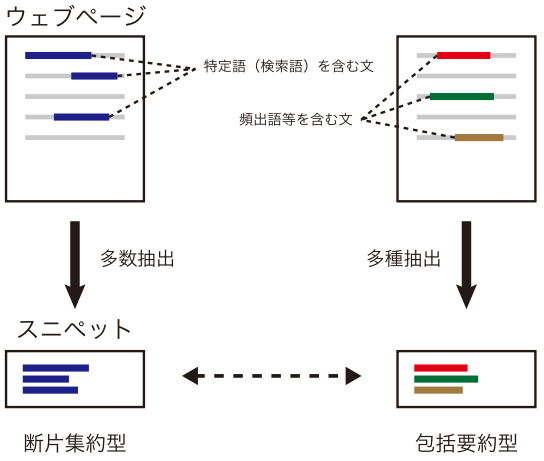

図 1　断片集約型と包括要約型
Fig. 1　Segment-collect and inclusive-summary type.

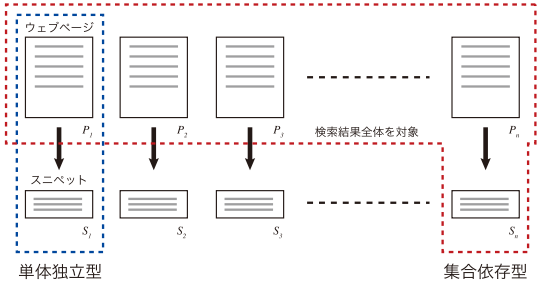

図 2　単体独立型と集合依存型
Fig. 2　Single-independent and multi-depedent type.

このように，内容の包括要約性および対象の集合依存性を軸として，生成されるスニペットのタイプは**図 3** の I 型から IV 型に分類することができる．

I 型は単体独立型かつ断片集約型で，既存のウェブ検索エンジンの多くがこの手法を採用している．検索対象についての知識が少ない場合，手がかりとして入力された検索語を含む周辺情報を提供すべきであり，このような場合は I 型のスニペットが適している．たとえば，検索語として「京都」と「湯豆腐」が入力された場合，利用者の検索目的は湯豆腐が食

© 2008 Information Processing Society of Japan

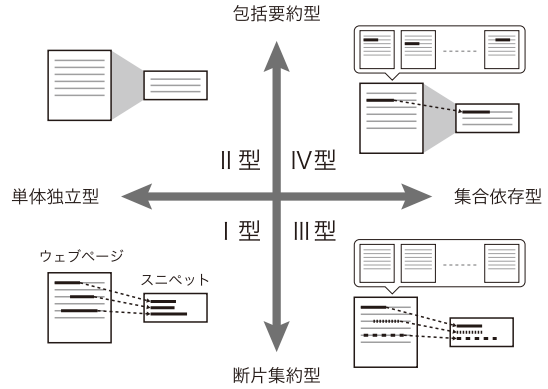

図 3　スニペットの生成方法による分類
Fig. 3　Classification of Web-snippet by generation method.

べられるお店を探すことかもしれないし，湯豆腐に使われる豆腐のことが知りたいのかもしれない．そこで，まずは検索語に共起しやすい話題を含む I 型のスニペットを採用することで，「京都で湯豆腐を食べるなら〜がおすすめ」「京都の〜では湯豆腐に嵐山の〜というお店の豆腐が」といった内容をスニペットとして提供することができる．

II 型は単体独立型かつ包括要約型で，検索質問適合度を評価基準として生成されたスニペットと比較して，検索質問に依存しない要約は，著者の意図を反映するものである．検索対象についての知識がいくらかある場合，特定の種類のウェブページを探している場合が多いため，その場合はウェブページの種類を推測しやすい包括的な情報を提供すべきであり，II 型のスニペットが適している．たとえば，検索語として「京都」「湯豆腐」「食べる」が入力された場合，利用者の検索目的はおそらく湯豆腐料理店を探すことだと思われる．このとき，「京都の湯豆腐料理店一覧」といった内容がスニペットとして含まれている場合，いくつかのお店が紹介されているウェブページであると容易に推測できる．また，「今日は，その南禅寺順正をご紹介」といった内容が含まれている場合は，特定の湯豆腐料理店についてウェブページであることが分かる．このようなテキストは，必ずしも検索語を含むとは限らない．

III 型は集合依存型かつ断片集約型であり，比較や分類といった検索対象を集合的に評価することが目的の場合などは，他のウェブページには存在しない独自性の高い断片をスニペットとして抽出することで，検索結果全体を見れば特徴的な情報を把握することができる．たとえば，検索語として「京都」と「湯豆腐」が入力された場合，「銀閣寺から永観堂を越え，南禅寺まで」「京都で〜が食べられるのは当店だけ」といった内容を含むスニペットを提供することができる．

IV 型は集合依存型かつ包括要約型であり，検索結果に含まれる複数のウェブページに対して，共通属性における相違点を知りたい場合，要約を生成する手法に加えて，共通頻出語に重みを与えることで，他のウェブページと相関性の高い部分，つまり，検索質問に関連する共通の話題をスニペットとして抽出することで，所在地や営業時間の比較などを行うことができる．たとえば，検索語として「京都」と「湯豆腐」が入力された場合，湯豆腐料理店が集まる「南禅寺」という地名が共通頻出語として重み付けられ，「京都市左京区南禅寺草川町」といった内容を含むスニペットを提供することができる．

4. スニペットの動的再生成

4.1 重要語によるスニペット生成

我々は，各文を重み付けるために利用する検索語などの重要語を変化させることによって，利用者の検索目的に応じて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．

まず，ウェブページの HTML ソースからタグを取り除き，文単位に分解し，各文の重要度を求める．重要語 w_i が持つ重みを v_{w_i} とすると，文 s の重要度 $Rank(s)$ は以下のように計算する（式 (1)）．$E(w_i)$ は，文 s に重要語 w_i を含む場合は 1，含まない場合は 0 となる関数である．

$$Rank(s) = \sum_{i=1}^{n} \{E(w_i) \cdot v_{w_i}\} \tag{1}$$

利用する重要語と生成されるスニペットのタイプの関係は以下のとおりである．

- I 型：検索語
- II 型：頻出語（tf 値[11] の大きな単語）
- III 型：独自頻出語（$tf \times idf$ 値の大きな単語）
- IV 型：共通頻出語（$tf \times df$ 値の大きな単語）

※ tf: term frequency
※ $(i)df$: (inverse) document frequency

本研究では，検索語以外の重要語の抽出において，形態素解析器（茶筌）を使用し，代名

© 2008 Information Processing Society of Japan

詞・接尾詞・数詞を除く名詞のみを対象としている．重要語が検索語の場合，各検索語が持つ重みは通常同じ 1 となるため，検索語が同数含まれる文の重要度はすべて同じになってしまう．そこで，二次的な重み付けとして，我々は頻出語の重みを用いている．あくまでも二次的な重み付けとして利用するため，頻出語で重み付けた文の重要度は，最も高いものが 1 となるように正規化する．また，頻出語による重み付けは，検索質問に関係なく計算可能なため，事前にシステム側で用意しておくことができる．

スニペットは，このように計算した重要文ランキングにおいて，重要度の高い文を規定量選択し，出現順に配置することで生成する．このとき，重要語が検索語の場合は I 型のスニペット，頻出語の場合は II 型のスニペット，独自頻出語の場合は III 型のスニペット，共通頻出語の場合は IV 型のスニペットを生成することができる．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．なお，本研究ではスニペットの文字数が 300 バイトを超えない範囲で重要文を選択している．ただし，最も重要度の高い文がそれ以上の文字数を有する場合は，その 1 文のみをスニペットとして抽出している．

このように，従来の検索語を重み付けに利用したスニペットだけではなく，頻出語を利用することで，検索質問に依存しない，より包括要約的なスニペットが生成できる．また，検索結果内の他のウェブページの情報も考慮した独自頻出語や共通頻出語を重み付けに利用することで，従来型のスニペットにはない検索結果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，検索質問が決定された場合に一意に生成される現行のスニペットと比べ，Rich-Snippet を利用したウェブ情報検索モデルでは，利用者はその検索目的に応じてスニペットを動的に変更することが可能となる．この特徴は，同じ検索質問が入力された場合でも，利用者に適した検索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持つ．

4.2　Private View の実装

我々は利用者の検索目的に応じてスニペットを動的に再生成させることができるウェブ検索インタフェースである「Private View」を開発した．利用者が生成手法を選択することで I 型から IV 型までのスニペットが動的に再生成され，入力した検索質問を変更することなくスニペットの内容を変化させることができる．本システムでは，検索結果のランキングに Google を利用している．**図 4** は検索語として「京都」と「湯豆腐」を入力した場合に上位に表示されたあるホームページにおいて，I 型から IV 型までのスニペットがそれぞれ順に再生成される様子を示している．本システムでは，直前のスニペットとの変化部分を青色で表示することで，どこが変化したかが一目で分かるように工夫されている．

再生成されたスニペットの特性の違いを視覚化するために，我々は各重要語がスニペット

図 4　Private View：スニペットの動的再生成
Fig. 4　Private View: Dynamic re-generation of Web-snippet.

にどれだけ含まれているかを示す，以下の 4 種類の示度を考案した．$tf_p(w_i)$ は重要語 w_i のウェブページにおける出現数であり，$tf_s(w_i)$ は重要語 w_i のスニペットにおける出現数である．また，$df(w_i)$ は検索結果内における重要語 w_i が出現するウェブページの数で，$idf(w_i)$ は $\frac{1}{df(w_i)}$ である．$E_s(w_i)$ は，スニペットに重要語 w_i を含む場合は 1，含まない場合は 0 となる関数である．なお，図 4 ではスニペットの右側に各網羅度がグラフ化され表示されている．

- 特定語網羅度（E_I）

$$E_I = \frac{\sum_{i=1}^{n} tf_s(w_i)}{\sum_{i=1}^{n} tf_p(w_i)} \tag{2}$$

© 2008 Information Processing Society of Japan

表 1　実験用の質問と目的
Table 1　Queries and purposes for experiment.

No.	検索質問	検索目的
1	felica	FeliCa（フェリカ）とは何か知りたい
2	CSS	ウェブにおける CSS とは何の略か知りたい
3	エネループ	エネループってどんな電池か知りたい
4	18 金+24 金	18 金や 24 金の「18」や「24」は何を意味するのか知りたい
5	湯豆腐	湯豆腐とはどんな料理か知りたい.
6	エルメス+マグカップ	エルメスのマグカップが買えるサイトを探したい
7	オーストラリア+お土産	オーストラリアのお土産をいろいろ紹介しているサイトを探したい
8	液晶+プラズマ+電気代	液晶テレビとプラズマテレビの電気代の違いについて書かれたサイトを探したい
9	光ファイバー+インターネット	光ファイバーインターネットが契約できる会社のサイトを探したい
10	京都+湯豆腐+お寺	京都で湯豆腐が食べられるお寺について書かれたサイトを探したい
11	夏+カレー	変わったカレーについて知りたい
12	マイル+海外旅行	マイレージ（マイル）を活用した変わった海外旅行について知りたい
13	電子マネー	マイナーな電子マネーが知りたい
14	兵庫+たこ焼き	明石焼（たまご焼き）が食べられる店が知りたい
15	京都+ステーキ	京都の祇園でステーキが食べられる店が知りたい
16	マンション+アパート+違い	マンションとアパートの一般的な基準の違いについて知りたい
17	からし+マスタード+違い	からしとマスタードの一般的な原料の違いについて知りたい
18	携帯+メモリカード	携帯の代表的なメモリカードについて知りたい
19	梅干し+アルカリ性+酸性	梅干しはアルカリ性食品？それとも酸性食品？が知りたい
20	京都+観光+嵐山	京都嵐山の代表的な観光名所が知りたい

- 頻出語網羅度（E_{II}）

$$E_{II} = \frac{\sum_{i=1}^{n} \{E_s(w_i) \cdot tf_p(w_i)\}}{\sum_{i=1}^{n} tf_p(w_i)} \tag{3}$$

- 独自頻出語網羅度（E_{III}）

$$E_{III} = \frac{\sum_{i=1}^{n} \{E_s(w_i) \cdot tf_p(w_i) \cdot idf(w_i)\}}{\sum_{i=1}^{n} \{tf_p(w_i) \cdot idf(w_i)\}} \tag{4}$$

- 共通頻出語網羅度（E_{IV}）

$$E_{IV} = \frac{\sum_{i=1}^{n} \{E_s(w_i) \cdot tf_p(w_i) \cdot df(w_i)\}}{\sum_{i=1}^{n} \{tf_p(w_i) \cdot df(w_i)\}} \tag{5}$$

4.3　スニペット・タイプと検索精度に関する評価

次に，我々は本論文で提案した 4 種類のスニペット・タイプが検索精度に与える影響を調べるために，様々な検索質問と検索目的を用意し，合計 20 人の被験者を対象に評価実験を行った．本実験では，様々なジャンルのキーワードをもとに準備した数十種類の中から検索結果が適当な 20 問を検索質問として用意し，Google の検索結果上位 20 件までに表示されるウェブページに対して，検索結果のみを確認する形で各ウェブページが検索目的に適していそうかどうかの判定を行ってもらった．各被験者に対して，I 型から IV 型までのスニペットを各 5 問ずつ表示した．つまり，被験者 A の第 1 問は I 型のスニペット，第 2 問は II 型のスニペットが表示され，被験者 B の第 1 問は II 型のスニペット，第 2 問は III 型のスニペットが表示されるようにして実験を行った．なお，一般的な検索結果では表示されるタイトルの影響を評価するため，10 人には検索結果にスニペットのみを表示し，残りの 10 人にはタイトルとスニペットの両方を表示した．

表 1 は，本実験で使用した検索質問と検索目的のリストである．我々が提案した I 型から IV 型までのスニペットがどのような検索目的の場合に有効であるのかを調べることもまた，本研究の重要なテーマである．そのため，No.1〜5 までは，I 型のスニペット，No.6〜10 までは II 型のスニペット，No.11〜15 までは III 型のスニペット，No.16〜20 までは IV 型のスニペットに適した検索質問として問題を作成したが，被験者にはそれらの情報は伏

© 2008 Information Processing Society of Japan

表 2　総合適合率
Table 2　Integrated relevance.

タイプ	スニペットのみ	タイトル＋スニペット
I 型	0.6725	0.6700
II 型	0.6838	0.6988
III 型	0.6100	0.6738
IV 型	0.6788	0.7188

表 3　正解/不正解–適合率
Table 3　2 kinds of relevances.

タイプ	正解–適合率	不正解–適合率
I 型	0.5132	0.7410
II 型	0.5138	0.8108
III 型	0.4564	0.8924
IV 型	0.5670	0.8019

せ，どのタイプのスニペットが表示されているのかも分からない状態で実験を行った．なお，各ウェブページの正誤判定は，設定した検索目的に基づきそれぞれのウェブページの内容を確認することで，我々が事前に判断した結果（正誤判定表）をもとに行った．

表 2 は，各スニペット・タイプ別総合適合率を示している．総合適合率とは，正誤判定表に対する一致度であり，表 2 は各検索質問に対して計算した総合適合率の平均値である．この結果によると，I 型の場合はあまり差がないが，スニペットのみよりも，タイトルとスニペットの両方が提示される方が，検索精度が高くなることが分かる．また，検索語を含む I 型のスニペットよりも，包括要約性を考慮した II 型や IV 型のスニペットの方が平均的に検索精度が高いという結果が示されている．

表 3 は，タイトルとスニペットの両方を表示した被験者に対して，正解または不正解ページに対して，正しく判定した割合を評価した適合率の平均値を示している．この結果によると，独自頻出語を含む III 型は，正解ページを選択する精度は高くないが，不正解ページを正しく判断する精度が高いことが分かる．そのため，独自頻出語だけを重要語としてスニペット生成を行った場合，ノイズを多く含んだスニペットが生成されている可能性がある．また，検索語を含む I 型の場合，不正解ページを誤って正解ページと判断する可能性が高いことが分かる．

表 4 は，検索質問を 4 つに分け，総合適合率の平均値を求めた結果である．今回の実験

表 4　タイプ別総合適合率
Table 4　Relevance to Web-snippet types.

タイプ	No.1–5	No.6–10	No.11–15	No.16–20
I 型	**0.685**	0.675	0.625	0.695
II 型	0.735	**0.690**	0.695	0.675
III 型	0.790	0.660	**0.655**	0.590
IV 型	0.780	0.655	0.740	**0.700**

では，II 型と IV 型のスニペットを使用した場合は，想定したとおりの結果が得られたが，I 型と III 型のスニペットが適していると想定した検索質問の場合は他のスニペット・タイプの方が検索精度が高いという結果になった．

III 型のスニペットが想定した検索目的に適さなかった理由は，今回の実験のように日本語が多いウェブページでは，英単語や記号類の *idf* 値が高くなり多くがノイズとしてスニペットに混在してしまうことが原因と考えられる．ただし，誤った HTML 記述によるウェブページからスニペットを生成する際に，除去しきれなかったスクリプトやタグの一部が高い *idf* 値を示し，スニペットに混在してしまう問題については対策が難しい．また，意外性のある結果を想定しなければならないという意味では，III 型のスニペットに適した検索質問を用意すること自体が難しいが，スニペットの特性から，網羅的に情報を収集するようなタスクの方が向いていたのかもしれない．

I 型のスニペットについては，想定した検索目的の場合には他のスニペット・タイプと比較して最も低い総合適合率を示しているが，I 型におけるすべての問題に対しての平均値よりは高い．そのため，他の検索目的の場合よりは，想定した検索目的の場合に I 型のスニペットを使用することは有効であるが，I 型のスニペットに適していると想定した検索目的については，他のスニペット・タイプの方が適していたと考えることができる．

さらに，検索質問を個別に確認したところ，たとえば，No.8 の検索質問は II 型のスニペットを使用した場合に最も総合適合率が高いといったように，一部は想定どおりの結果が得られているが，II 型のスニペットが適していると想定していた検索質問は IV 型のスニペットを使用した場合に最も総合適合率が高いといったように，最適なスニペット・タイプの選択は検索目的だけではなく，検索語や検索結果にも依存していることが分かった．また，この調査結果から II 型と IV 型のスニペットに適した検索目的どうしには，何らかの近親性が存在する可能性も考えられる．

© 2008 Information Processing Society of Japan

4.4 今後の課題

今回の実験により，提案したI型からIV型までのスニペットのうち，II型のスニペットは検索対象についての知識がいくらかある場合，特定の種類のウェブページを探している場合に，IV型のスニペットは検索結果に含まれる複数のウェブページに対して，共通属性における相違点を知りたい場合に適していることが分かった．しかし，I型とIII型のスニペットは想定した検索目的の場合に必ずしも他のスニペット・タイプを使用した場合よりも検索精度が高くなるわけではなかった．また，最適なスニペット・タイプの選択は，検索目的だけではなく，検索語や検索結果にも依存していることが分かった．そのため，全体的に検索精度が低かったI型やIII型のスニペット生成手法の改善を行い，検索目的に加え各種網羅度の比率や検索語の特性から，検索精度が最も高くなるスニペット・タイプを判断することが今後の課題である．

5. 関 連 研 究

ウェブ検索エンジンの普及にともない，検索結果の最適化に着目した研究がいくつか行われている．Yahoo! Researchは，「Yahoo! Mindset」[12]と呼ばれる利用者の検索意図や検索目的に適した検索結果を表示する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，ウェブページの種類をcommercial（商品購買）とnon-commercial（商品情報）とに分類しており，利用者が購買目的または情報収集目的の度合いを選択することで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，検索結果の最適化を実現しているが，スニペットの機能は拡張されていない．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[7),8)]．しかし，既存のウェブ検索エンジンにより提供される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最適化しようとする研究もある[13),14)]．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．

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は図書検索と違い，検索対象はウェブページの枠を越えたウェブ空間に存在するすべての情報である．そのため，検索対象は複雑化し，検索目的は多様化してきている．本論文では，内容の包括要約性および対象の集合依存性を軸として，検索目的に対応したI型からIV型までのタイプにスニペットを分類し，それぞれのスニペットを4種類の重要語により生成する手法を提案した．また，スニペット・タイプの違いが検索精度に与える影響について評価を行い，包括要約性を考慮したスニペットについては，想定した検索目的の場合に適していることを示した．

我々は，ウェブ情報検索において検索質問のほかに，検索目的を入力する新たなウェブ情報検索モデルを提案した．利用者自身の手により加工，最適化される情報を提示することは，ウェブページよりも粒度の細かい情報や集約された情報を提供するためのインタフェースとしても新しい方法論である．本研究では，ウェブ情報の再構成による新たな知識の創成を，検索目的に応じたスニペットで構成される検索結果という形で実現し，検索目的に適したスニペットの提示により検索精度が向上したことで，本手法がウェブ情報検索を質的に改善させる可能性が高いことを確認できた．本研究の成果が，ウェブ知識検索へ向けた研究開発の発展に寄与し，ウェブ情報の新たな活用へのイノベーションを誘発するきっかけとなることを期待している．

謝辞　本研究の一部は，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」，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）ならびに，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」，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）および文部科学省グローバルCOE拠点形成プログラム「知識循環社会のための情報学教育研究拠点」（研究代表者：田中克己，平成19〜23年度）によるものです．ここに記して謝意を表すものとします．

参 考 文 献

1) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
2) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. 7th International Conference on World Wide Web 7*, Amsterdam, The Netherlands, The Netherlands, pp.107–117, Elsevier Science Publishers B.V. (1998).
3) 高見真也，田中克己：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，*DBSJ Letters*, Vol.6, No.1, pp.109–112 (2007).
4) Hearst, M.A. and Pedersen, J.O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR-1996*), New

© 2008 Information Processing Society of Japan

York, NY, USA, pp.76–84, ACM Press (1996).
5) Wang, Y. and Kitsuregawa, M.: Evaluating contents-link coupled web page clustering for web search results, *Proc. 11th International Conference on Information and Knowledge Management* (*CIKM-2002*), New York, NY, USA, pp.499–506, ACM Press (2002).
6) Glover, E.J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D.M. and Flake, G.W.: Using web structure for classifying and describing web pages, *Proc. 11th International Conference on World Wide Web* (*WWW-2002*), New York, NY, USA, pp.562–569, ACM Press (2002).
7) Ferragina, P. and Gulli, A.: 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, pp.801–810, ACM Press (2005).
8) Geraci, F., Pellegrini, M., Pisati, P. and Sebastiani, F.: A scalable algorithm for high-quality clustering of web snippets, *Proc. 2006 ACM Symposium on Applied Computing* (*SAC-2006*), New York, NY, USA, pp.1058–1062, ACM Press (2006).
9) Amitay, E. and Paris, C.: Automatically summarising Web sites: Is there a way around it?, *Proc. 9th International Conference on Information and Knowledge Management* (*CIKM-2000*), New York, NY, USA, pp.173–179, ACM Press (2000).
10) Oyama, S. and Tanaka, K.: Query Modification by Discovering Topics from Web Page Structures, *Proc. 6th Asia-Pacific Web Conference* (*APWeb-2004*), Lecture Notes in Computer Science, Vol.3007, pp.553–564, Springer Berlin/Heidelberg (2004).
11) Salton, G. and Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval, Technical Report, Ithaca, NY, USA (1987).
12) Yahoo! Research: Yahoo! Mindset. http://mindset.research.yahoo.com/
13) Dontcheva, M., Drucker, S.M., Wade, G., Salesin, D. and Cohen, M.F.: Summarizing personal web browsing sessions, *Proc. 19th Annual ACM Symposium on User Interface Software and Technology* (*UIST-2006*), New York, NY, USA, pp.115–124, ACM Press (2006).
14) Boydell, O. and Smyth, B.: Community-based snippet-indexes for pseudo-anonymous personalization in web search, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR-2006*), New York, NY, USA, pp.617–618, ACM Press (2006).

(平成 19 年 6 月 30 日受付)
(平成 20 年 1 月 8 日採録)

高見　真也（学生会員）

平成 15 年京都大学大学院情報学研究科修士課程修了．平成 19 年同博士後期課程指導認定退学．同年楽天株式会社に入社．現在，楽天技術研究所研究員．主にデータベース，広告の流通，ウェブ情報検索の研究に従事．IEEE Computer Society，日本データベース学会等各会員．

田中　克己（正会員）

昭和 51 年京都大学大学院工学研究科修士課程修了．博士（工学）．現在，京都大学大学院情報学研究科社会情報学専攻教授．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会等各会員．

© 2008 Information Processing Society of Japan