

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リランキング

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

高見 真也[†]

田中 克己[‡]

Shinya TAKAMI

Katsumi TANAKA

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、ユーザの検索目的に適したスニペットを動的に再生成することで、検索結果の把握がすばやく行えるだけでなく、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、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、その類似度で検索結果のリランキングを行う手法を提案する。

For personalizing search results returned by Web search engines, many researchers focused on 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's purpose. It will also help to improve clustering quality. In this paper, we propose a new re-ranking method of search results based on the similarity of Web-snippets.

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クエリとして単語の組み合わせ（以下、検索語と呼ぶ）を入力し、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、与えられた検索語に対して、ウェブページの重要度が一意に決定されるランキング方式では、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、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、検索解ページ群のうち少なくとも一つが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情報検索を行う場合のように、複数の検索解ページを発見することが求められる場合は、検索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なく、検索結果のクラスタリングやリランキングを行うことで、情報検索を支援する手法が提案されている。ただし、ウェブページやその抜粋（以下、スニペットと呼ぶ）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うまくいかない。

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に着目した。ユーザの検索目的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、ある検索解ページ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することで、他の検索解ページを発見しやすくする手法を提案する。

2. ウェブ情報検索支援

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、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ランクする手法としては、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、検索語を再考し再検索が行われることが多い。そこで、検索語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、検索語の拡張（クエリ拡張）が必ずしも検索解ページのランキング上昇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、検索語として「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」を検索語に追加して再検索を行った。表1は各単語が検索語に追加された場合の順位の変化を示している。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。

[†] 学生会員 京都大学大学院 情報学研究科 博士後期課程
shie@dl.kuis.kyoto-u.ac.jp

[‡] 正会員 京都大学大学院 情報学研究科
tanaka@dl.kuis.kyoto-u.ac.jp

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[↓]

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が多く、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ふ可能性がある。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、ウェブページのどのセグメントがスニペットとして抽出されたかによって、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、ウェブページ内における単語の出現位置が、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、他にそれらの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、ウェブページで行う場合は対象とする範囲、スニペットで行う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ための判断材料として、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、ウェブページの実態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実態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。

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。また、スニペットは与えられた検索語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、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実態を定量的には表現しておらず、検索語が与えら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。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、検索語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実態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. 類似性を考慮したスニペットの再生成

ウェブ検索エンジンを利用した情報検索において、検索結果として返される順位がそのまま検索目的への適合度を表しているわけではない。そのため、検索結果におけるスニペットは、ウェブページの内容や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スニペットは、検索目的ではなく検索語に依存した手法により生成されている。

比較や調査といった複数のウェブページを発見することが目的である場合、最初に見つけた検索解ページと類似する内容をもつものがあれば、そのウェブページも検索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他のスニペットを再生成する手法を提案する。さらに、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、情報検索支援を行う手法を提案する。

3.1 スニペット入力型生成手法

現行のスニペットの多くは、検索語を多く含むセグメントをウェブページからいくつか抽出する、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検索語の一部を含んでいることが多い。しかし、検索目的によっては、検索語そのものは重要ではない場合もある。例えば、京都の観光についてウェブ情報検索を行う場合、検索語として「京都八観光」が入力されたとしても、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。

複数の検索解ページを求める必要がある場合、最初の検索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検索解ページである可能性が高い。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ことが多く、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、クエリに依存した生成手法では、検索語以外の重要語が類似したウェブページ群のスニペットに含まれているとは限らない。

そこで、我々はスニペット同士の類似度を評価する前に、検索語以外の重要な単語ができるだけスニペットに含まれるようにスニペットを生成する手法を提案する。本手法では、特定のスニペットをユーザが選択することで、そのスニペットに含まれる単語を重要語とみなし、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ことができる。検索語の代わりにスニペットを入力とした、スニペット入力型生成手法では、以下のようにスニペットを生成する。

- 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
- ウェブページの各文を単位とした変形 TF-IDF 法 (式 1) により、抽出された単語 t の重要度 $w(t)$ を計算

$$w(t) = \frac{\text{ウェブページに含まれる単語}t\text{の数}}{\log(\text{単語}t\text{を含む文の数})} \dots (1)$$

- 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
- 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、スニペットを構成す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、検索語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、特定の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。

3.2 スニペットの類似度とリランキング

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。また、このような類似度評価は、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[11]。ここでは、N-gram 適合率で評価する BLEU (式 2) とその改良版である ROUGE (式 3) について、その計算方法を以下に示す。

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N \frac{1}{N} \log P_n\right) \dots (2)$$

$$ROUGE(i, j) = \exp\left(\sum_{n=1}^j \frac{1}{(j-i+1)} \log C_n\right) \dots (3)$$

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検索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と考えられる。また、検索解ページのものと同様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、複数の検索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。

3.3 評価

本手法により再生成されたスニペットの類似度変化を確認するために、先に紹介した湯豆腐料理店のホームページのうち、地域ごとの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 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、検索語が「京都八湯豆腐」の場合に Google が出力するスニペット、我々の提案手法により再生成されたスニペット、検索語に「南禅寺」を追加した場合に Google が出力するスニペットについて、それぞれ類似度を計算した結果を表 2 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、提案手法により再生成されたスニペットの方が高い類似度を示すことが確認できた。また、嵐山/嵯峨野周辺の店舗についても同様の結果が得られた。今後は、類似度でリランキングを行った際に、検索解ページを見つけるまでの時間が短縮されるかなど、ウェブ情報検索の支援効果に関する評価[12]も行う予定である。

表 2 スニペットの類似度評価
Table 2 Evaluation of Web-snippet Similarity

店舗	地域	基準	提案手法	八南禅寺
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

4. 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ある。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なっている[6][7]。しかし、既存のウェブ検索エンジンによって提供されるスニペットを扱っているために、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[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。Yahoo! Research は、「Yahoo! Mindset」[15]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、調査目的または購買目的の度合いを入力することで検索結果をリランキングすることができるが、スニペットの機能は拡張されていない。

5. おわりに

ウェブ検索エンジンを利用した情報検索では、ユーザは検索結果の上位数件程度にしか興味を示さないとされている。しかし、検索語として入力された情報だけで、ユーザがもつ多様な検索目的を満足させることは難しい。我々はランキングの精度を向上させるのではなく、検索結果をよりユーザの検索目的に適した形に動的に変化させることで、情報検索を支援できると考えている。

比較等を目的としたウェブ情報検索の場合、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「似て異なる情報」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、類似部分ではなく、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るのではないかと考えている。このように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。

[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. Tsioutsoulouklis, 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', <http://www.sem-irep.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] 奥村学, 難波英嗣: '知の科学: テキスト自動要約', オーム社 (2005).
- [12] 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).
- [13] 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).
- [14] 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).
- [15] Yahoo! Research: 'Yahoo! Mindset', <http://mindset.research.yahoo.com/>.

高見 真也 Shinya TAKAMI

京都大学大学院情報学研究科社会情報学専攻博士後期課程在学中。2003 年京都大学大学院情報学研究科社会情報学専攻博士前期課程修了。主にウェブからの知識発見、情報検索支援システムの研究・開発に従事。IEEE Computer Society, 情報処理学会, 日本データベース学会, 各学生会員。

田中 克己 Katsumi TANAKA

京都大学大学院情報学研究科社会情報学専攻教授。1976 年京都大学大学院修士課程修了。博士(工学)。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, 情報処理学会, 日本データベース学会等各会員。