検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化

Customizing Web Search Results by Dynamic Re-Generation of Web-Snippet based on Search Purpose

高見 真也♥　　田中 克己♠
Shinya TAKAMI　Katsumi TANAKA

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，ユーザ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等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ユーザ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．我々はユーザの検索目的に適した検索結果を提供するために，検索結果として示すべき概要文（スニペット）を二種類の軸で分類した．そして，最適化された検索結果を実現するためのスニペットの動的再生成を紹介する．

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. We classified the outline (Web-Snippet) in the search result by two criteria to offer a suitable search result for user's search purpose. Then we introduce dynamic re-generation of the Web-Snippet to achieve the optimization of search results.

1. はじめに

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み合わせを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．一般にデータベースへの問い合わせをクエリと呼ぶが，本論文では情報検索の場合に限定し，ウェブ検索エンジンに与える検索語の組み合わせをクエリと呼ぶことにする．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（以下，スニペットと呼ぶ）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，クエリによく適合するウェブページが検索結果の上位にランクされることはもちろん重要であるが，たとえまったく同じクエリ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順位や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動的再生成を実現することで，検索結果を個人適応化し，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. 検索目的に応じたスニペット生成

2.1 スニペットの特徴

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」[1]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，検索結果におけるスニペットがウェブページの特性を推測する際に重要な情報と見なされていることに注目したい．

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ており，必ずしも意味的に抽出されているわけではない．つまり，既存のスニペットは，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．一方で，各検索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．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我々は考えている．

2.2 スニペットの分類

我々はユーザ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により二種類の軸で分類した．一つ目の軸は，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．単体独立型は一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，全体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 1）．

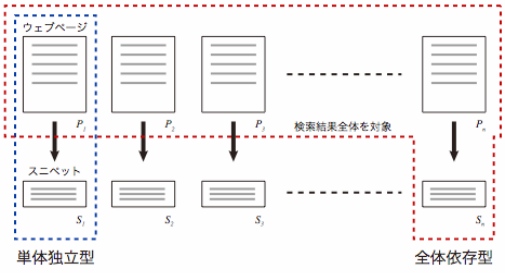

図 1　単体独立型と全体依存型

Fig.1　Single-Independent/Multi-Depedent Type

♥ 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

もう一つの軸は，ウェブページからスニペットを生成する際に，内容の包括性を考慮するかどうかである．断片集約型は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，包括要約型は検索語などには依存しない全体の内容を包括した要約としてのスニペットを生成するタイプである．このように，スニペットの生成方法を対象の集合依存性および内容の包括要約性で分類すると，スニペットは図 2 の I 型から IV 型に整理することができる．

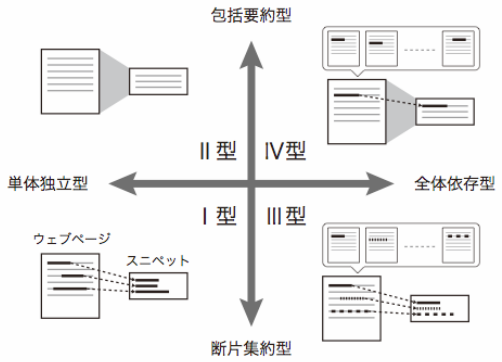

図 2　スニペットの生成方法による分類

Fig.2　Classification of Web-Snippet by Generation Method

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タイプを採用している．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．例えば，検索語として「京都」と「湯豆腐」が入力された場合，ユーザの検索目的は湯豆腐が食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，I 型を採用することで，「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」といった内容のスニペットを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．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．例えば，検索語として「京都」「湯豆腐」「食べる」が入力された場合，ユーザ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出来る．このようなスニペットは，必ずしも検索語を含むとは限らない．

ウェブ情報検索の目的が，検索結果に含まれる複数のウェブページを比較したり横断的に確認することである場合，それぞれのウェブページを独立に評価するのではなく，他との関係も考慮する必要があり，III 型または IV 型のスニペットが適している．III 型は全体依存型かつ断片集約型であり，他のウェブページには存在しない排他性の高い断片をスニペットとして抽出する場合，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．例えば，検索語として「京都」と「湯豆腐」が入力された場合，「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．また，全体依存型かつ包括要約型である IV 型のスニペットでは，検索結果全体に共通するような内容が示されることで，所在地や営業時間の比較を行うことができる．

2.3 スニペットの生成手法

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検索語を含む断片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[3]．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，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，図 2 における I 型に該当する．

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．このようなクエリ独立型要約手法により生成されたスニペットは，図 2 における包括性を重視した II 型となる．コンピュータによるクエリ独立な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[4][5][6]．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．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，重要文抽出によるアプローチを要約生成手法として採用することにする．

図 2 で示したスニペットの分類において，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は，III 型や IV 型のスニペットを生成すべきである．このような全体依存型のスニペットも，重要文抽出手法により生成することができる．断片集約型の場合，検索結果に含まれる上位 k 件のウェブページを対象にして，TF-iDF 値[7]の大きな単語に重みを与えることで，スニペットとして生成される内容がウェブページごとに独自性をもつことになる．また，包括要約型の場合は，要約を生成する手法に加えて，TF-DF 値の大きな単語等に重みを与えることで，他のウェブページとの関係を考慮したスニペットを生成することができる．

3. スニペットの動的再生成

3.1 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，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．

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 がもつ重みを 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．

$$Rank(s) = \sum_{i=1}^{n} \left\{ E(w_i) \cdot v_{w_i} \right\} \quad \cdots(1)$$

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通りである．

- 特定語（索語）：　I 型
- 頻出語（TF 値が高い単語）：　II 型
- 独自頻出語（TF-iDF 値　〃　）：　III 型
- 共通頻出語（TF-DF 値　〃　）：　IV 型

重要語が検索語の場合，各検索語が持つ重みは通常同じになるため，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．頻出語の重みは，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ためである．

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数選択し，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．

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．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，クエリに依存しない，より包括要約的なスニペットが生成でき，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，クエリ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 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．

3.2 Private View の実装と今後の課題

我々はユーザ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．ユーザ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，入力したクエリ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．図3はあるウェブページのI型からIV型までのスニペットが順に再生成された様子を示している．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．

図3 スニペットの動的再生成

Fig.3 Dynamic Re-generation of Web-Snippet

再生成されたスニペットの評価を行うために，我々は各重要語がスニペットにどれだけ含まれているかを示す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．$E(w_i)$ 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．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，$idf(w_i)$ はその逆数である．なお，図3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

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

- 頻出語網羅度

$$E_{II} = \frac{\sum_{i=1}^{n} \{E(w_i) \cdot tf_p(w_i)\}}{\sum_{i=1}^{n} tf_p(w_i)} \quad \cdots(3)$$

- 独自頻出語網羅度

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

- 共通頻出語網羅度

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

図4は，検索語として「京都」と「湯豆腐」を入力した場合の Google の検索結果上位10件について，I型からIV型までの各スニペットにおいてそれぞれ4種類の網羅度を計算し，平均値を示した例である．I型の場合は，特定語網羅度が他のタイプに比べ比較的高く評価されているのが分かる．同様に，II型の場合は頻出語網羅度，III型の場合は独自頻出語網羅度，IV型の場合は共通頻出語網羅度が高くなる傾向がある．他のタイプのスニペットと比べ，高い網羅度を示す重要語同士は共起性が高いことが示されており，入力された検索語における検索結果の特徴を示す重要な情報である．また，別の検索語を入力した場合のものと比べると，各網羅度の比率がクエリに依存していることが確認できた．そのため，このような網羅度と検索語との関係を分析することで，ユーザが入力した検索語に適したスニペットのタイプをシステムが提示することができるようになると期待している．

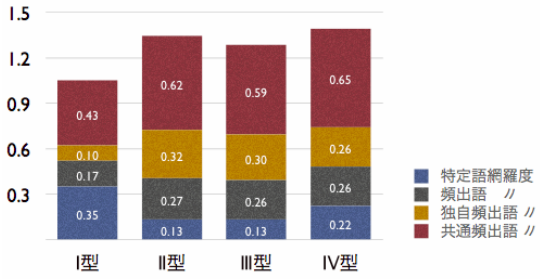

図4 各種網羅度によるスニペットの評価

Fig.4 Evaluation of Web-Snippet

検索目的に応じてスニペットを動的に変更できることで，従来の静的なスニペットに比べ，ウェブページの全容をより正確に把握することができる．また，各種網羅度で示されるスニペットの特徴評価を視覚化することで，スニペットの特徴と変更した場合の特徴量の変化を確認することができ，ウ

ェブページの特徴をより直感的に把握出来るようになった．今後は，各スニペットタイプと検索目的との適合度について評価を行う予定である．

4. 関連研究

ウェブ検索エンジンの普及に伴い，検索結果の個人適応化に着目した研究がいくつか行われている．Yahoo! Research は，「Yahoo! Mindset」[8]と呼ばれるユーザ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，ユーザ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，検索結果の個人適応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．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[9][10]．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個人適応化しようとする研究もある[11][12]．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．

5. 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超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，検索目的に適した検索結果を示すために，スニペットをその生成方法により二種類の軸で分類し，スニペットの動的再生成を行う手法を提案した．我々は，このように検索結果を個人適応化すべく検索目的に適した検索結果表示を行うことを SPO（Search Purpose Optimization）と呼んでいる．

[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．

[文献]

[1] 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.

[2] 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).

[3] Hu, Y., Xin, G., Song, R., Hu, G., Shi, S., Cao, Y. and Li, H.: “Title extraction from bodies of HTML documents and its application to web page retrieval”, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-2005), New York, NY, USA, ACM Press, pp. 250-257 (2005).

[4] Luhn, H.P.: “The automatic creation of literature abstracts”, IBM Journal of Research and Development, Vol.2, No.2, pp. 159-165 (1958).

[5] Paice, C.D.: “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).

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

[7] Salton, G. and Buckley, C.: “Term Weighting Approaches in Automatic Text Retrieval”, Technical report, Ithaca, NY, USA (1987).

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

[9] 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).

[10] 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).

[11] 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).

[12] 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).

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．