様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月３１日現在

研究種目：若手研究（B）
研究期間：2008～2009
課題番号：２０７００２２７
研究課題名（和文）　部分文書検索技術を利用した検索エンジンの
スニペット構築に関する研究
研究課題名（英文）　Construction of Result Snippets on Search Engines
using XML Fragment Search Techniques

研究代表者
波多野　賢治（HATANO KENJI）
同志社大学・文化情報学部・准教授
研究者番号：８０３１４５３２

研究成果の概要（和文）：現在の検索エンジンでは，検索キーワードが出現している位置からスニペットを作成しているが，そのようなスニペットではユーザは検索された文書の内容を理解できないことが多いという問題が起こっていた．そこで，本研究では Web 文書の構造化文書である特徴を利用し，部分文書検索技術を用いてスニペットを構成すべき部分を特定することで，よりスニペットとしてふさわしい部分の抽出を目指した．評価実験の結果，既存手法と比較し，約 10%の精度向上を図ることができた．

研究成果の概要（英文）：Conventional search engines generate result snippets utilizing the positions of query keywords in an Web document. However, such result snippets have a bothersome problem that users cannot recognize their contents of Web documents. We tried to solve the problem using XML fragment search techniques, and report the effectiveness of our approach. Our approach can identify candidates of result snippets using XML fragment search techniques, so that they are coherent content of user's information need. As a result, our approach is more effective than existing methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,900,000	570,000	2,470,000
２００９年度	1,400,000	420,000	1,820,000
年度			
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：情報メディア，情報検索

１．研究開始当初の背景

現在，World Wide Web (WWW) 上には爆発的な量の Web 文書が存在している．このような情報過多の状況下では，インターネット利用者は情報検索技術を活用することで必要な情報を効率的に取得している．これら情報検索技術はそういったユーザに対してこれまで非常に大きな役割を果たしてきたということは紛れもない事実であるとともに，今後はこれまでとは異なる技術が求められているということもまた事実である．それはつまり，単にユーザの問合せに対して適合

する Web 文書を検索結果として提示するのではなく，ユーザにとって更に有益な情報を併せて提示する技術である．

Google や Yahoo! といった現在汎用的に利用されている検索エンジンにおいて，検索結果として文書のタイトルやリンクなどとともに文書の概要としてスニペットと呼ばれる 50～100 文字程度の短い説明文が提示される．過去に行われたアンケートから，これらスニペットはユーザが実際に訪問する文書を選択する際に重要視されているということが判明している．それにも関わらず，同アンケート結果によるとユーザは現在提示されている情報だけでは依然として不足であると感じている．つまり，現在の検索エンジンにおいて提示されるスニペットが，文書の内容を要約して提示するという本来の役割を十分に果たしていないということである．

このような現状を踏まえ，研究代表者はスニペット生成技術の問題に着目し，よりよいスニペットの構築に必要技術に関する研究を行うことにした．

２．研究の目的

前述の通り，現状のスニペットはユーザに対して満足な働きをしていない．なぜなら，これまでユーザに提示されているスニペットは，ユーザの問合せに含まれるキーワードが集中する箇所をある一定の文字数の範囲で抜き出してくるといったものであるからである．つまり，本来，人間が物事の解釈を行う際に必要不可欠である文章の文脈や文書中の要点を十分に考慮できていないことが大きな問題なのである．

これらスニペットは自然言語処理による重要部分抽出技術を利用することで作成されているが，Web 検索においてはこれらの技術を利用するだけでは不十分である．自然言語処理分野によって行われた研究の成果として得られている文書の自動要約技術は非常に高い精度を示しているが，Web 文書はくだけた内容のものが多く，自動要約がうまく機能しないことが多い．したがって，Web 文書が持つ別の特徴を用いてスニペットを作成しない限り，ユーザの満足のいくスニペットを作成できないのである．

そこで本研究では，Web 文書のもつ特徴の一つである文書構造を利用して，抽出されるべき文書部分の候補を絞り込み，その中から最終的にスニペットを決定する方法を提案した．この方法により，Web 文書の特性を文書構造から把握し，その特性を生かしたスニペット生成が可能となり，ユーザの情報要求を満足するスニペット生成につながると考えた．

３．研究の方法

(1) 研究代表者は，これまでに Web 文書をはじめとする構造化文書に対し，文書単位よりも細かい粒度，すなわち構造化文書中の開始タグと終了タグで囲まれたそれぞれの部分を一つの文書として検索対象とする部分文書検索技術に関する研究を続けてきた．部分文書検索はそもそも，ユーザの問合せに対して文書丸ごとを検索結果として提示するのではなく，文書中から情報要求を満たす部分のみを検索結果として提示することを目的として研究がなされている．したがって，本研究ではこれらの経験を活かすだけではなく，さらに重要部分抽出技術の知見を取り入れることで，文書中の概要と成り得る部分文書，すなわち文書構造を考慮したスニペットの抽出を試みている．

研究代表者が部分文書検索技術の適用を考えた理由は，構造化文書内の部分文書それぞれは，タグを用いて特定の意味を付与されており，それぞれのタグで囲まれたひとまとまりは一つ話題を表すと考えられたからである．そのため，部分文書単位でスニペットの作成を行うことで，従来は達成できていなかった文脈や文書全体のうちの重要度を考慮したスニペット作成を実現でき，これまで以上にユーザの情報要求を満足するスニペットが生成できるようになる．

一方，重要部分抽出に研究も，Web の利用が拡大するようになってから大きく発展している．この研究によって得られた知見として，重要部分は 1) ユーザの問合せに対して最大限の情報を盛りこまなければならない，2) 一読して理解できるように十分に要約を行わなければならない，3) スニペットは文書中の要約であるため本文中の用語をそのまま用いて作成しなければならない，という制約があり，スニペット生成もこれら制約を満たす必要があるのではないかと研究代表者は考えた．これら三つの制約を精査した結果，部分文書検索技術において要件 2)，3) を満たすための試みは存在したが，要件 1) を満たすための取り組みはされていなかったことが判明したため，要件 1) を考慮した部分文書検索技術を考案することでスニペットの作成を行うことができると考えた．

具体的には，これまでの部分文書検索技術では，文書検索技術に則り，索引語の生起頻度や文書長から得られる統計量によってスコアリング手法が考案されていた．これに対し本研究では，より問合せに対して適切な部分文書を抽出するために，問合せによって得られる統計量をも用いてスコアリング手法を提案した．提案手法によってそれぞれの部分文書に対して得点付けを行い，最も高い得点を付与された部分文書が，文書中のクエリに対して必要十分な最適部分，すなわちスニ

ペットであるとすることで，新しいスニペット生成方法の提案を行っている．

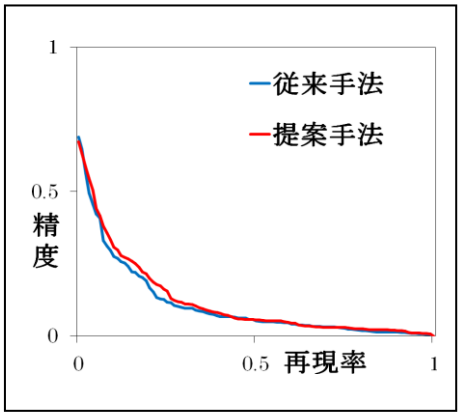
(2) (1) の研究が完遂されれば従来とは異なるアプローチによるスニペットを作成することが可能となるが，これだけではスニペットの目的である文書の要約を行うということにおいて不十分である．なぜなら，文書中の問合せに適合する部分を一つの部分文書で表すことが必ずしも適当であるとは限らないという問題が存在するためである．例えば，問合せに対する適合部分が文書中の離れた箇所に複数個存在する場合に，これら全てを一つの部分文書に含まれるように抽出すれば不要な部分も抽出してしまうという問題が起こる．これでは，要件 2) に反するためにスニペットとしては不適である．それに対して，もしいずれかの適合部分のみを抽出した場合には他の適合部分を抽出することができず，これでは要件 1) を満たすことができていないために，やはり有益なスニペットとは考えられない．

これらを踏まえて，一つの文書から複数の文書を抽出することも踏まえたスニペット作成手法を考案する必要がある．その際に，本研究では一つの文書から単に複数の部分文書を抽出するのではなく，部分文書の統合や集約などの加工を行うこととする．そうすることで，より問合せに対して必要十分な箇所の抽出を試みる．

４．研究成果

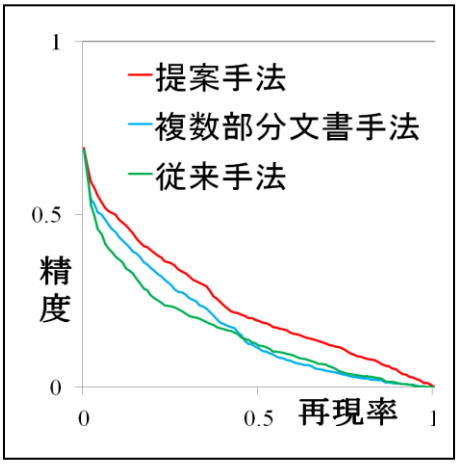
(1) 3. (1) で提案した手法の有用性を確認するために以下のような評価実験を行った．

実験には INEX プロジェクトによって構築された大規模 XML 部分文書評価用テストコレクションである INEX 2008 を利用した．評価実験の結果，1% から 100% までの各再現率点における精度の平均である MAiP が従来手法と比較して約 10% 向上した．



(2) 3. (2) で提案した手法の効果を確認するために，先程と同一のテストコレクションを用いて評価実験を行った．その結果，MAiP が約 46% 向上した．また，単純に一つの文書から複数の部分文書を抽出する手法と比較したところ，提案手法は 32% MAiP が高いという結果が得られた．

この結果から，部分文書に対して統合や集約などの加工を行うことで大きな効果が得られるということが判明した．



これらの結果が示している通り，研究代表者が行った研究は，この二年間で実用的なスニペットの作成に至ったということができる．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Atsushi Keyaki, Kenji Hatano, and Jun Miyazaki: A Query-oriented XML Fragment Search Approach on A Relational Database, Journal of Digital Information Management, Vol.8, No.3, pp.175-180, 2010.

〔学会発表〕（計 5 件）

① 欅惇志，波多野賢治，宮崎純: XML 検索技術を利用した検索結果の構成手法，情報処理学会第 72 回全国大会，東京，2010 年 3 月 11 日．
② 欅惇志，波多野賢治，宮崎純: XML 部分文書の再構成に基づく検索結果の提示手法，電子上通信学会第 2 回データ工学と情報マネージメントに関するフォーラム (DEIM 2010)，淡路島，2010 年 3 月 1 日
③ Atsushi Keyaki, Jun Miyazaki, and Kenji Hatano: A Method of Generating Answer XML Fragment from Ranked Results, INEX 2009 Workshop, Brisbane, Australia,7 December 2009.

④ Atsushi Keyaki, Kenji Hatano, and Jun Miyazaki: A Scoring Method of XML Fragments Considering Query-Oriented Statistics, 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2009), London, UK, 5 August 2009.
⑤ 欅惇志，波多野賢治，宮崎純: 索引語の統計量を用いた XML 部分文書検索法の組合せ利用とその効果，情報処理学会第 148 回データベースシステム研究会・第 95 回情報学基礎研究会 合同研究会，神戸，2009 年 7 月 28 日.

６．研究組織

(1)研究代表者

波多野　賢治（HATANO　KENJI）
同志社大学・文化情報学部・准教授
研究者番号：８０３１４５３２

(2)研究分担者

なし

(3)連携研究者

なし