D-009

Webコンテンツ作成支援のためのリンク目的を意識したリンク先推薦手法

Reference Page Recommendation Based on Intention of Adding a Hyperlink for Web Content Creation

武吉 朋也† 服部 元† 小野 智弘† 滝嶋 康弘†
Tomoya Takeyoshi Gen Hattori Chihiro Ono Yasuhiro Takishima

1. まえがき

ブログの登場により、HTMLタグ等の記述上の専門知識がなくともWebコンテンツを作成し、Web上に公開することが一般のユーザにも容易になった。コンテンツ作成者はWebコンテンツの作成過程において、他のコンテンツへのリンクを作成することも多く、その目的(以降、リンク目的)としては、キーワードの解説や、自ら過去に作成したコンテンツの参照等、複数挙げられる。この場合、コンテンツ作成者はそのリンク目的に対応するWebページを自ら探す必要があり、リンク先コンテンツの検索や、URLの転記等の煩雑な作業が必要となる。コンテンツ作成者に対するこのような負担を軽減する従来技術としては、特定のキーワードに自動的にキーワード解説ページへのリンクを生成するキーワードリンク[1]や、コンテンツ中に記述された対象(例えば店舗名)の属性(例えば営業時間)とその値をWebから自動取得し、提示する技術[2]等が挙げられる。しかしながら、これらの技術を用いると、ある特定のリンク目的のみを満たす結果、あるいはリンク目的の区別のない結果が提示されてしまう。

そこで本稿では、Webコンテンツ作成過程におけるリンクの作成支援を目的とし、コンテンツ作成者のリンク目的に応じてリンク先候補となるコンテンツをWebから自動取得し、コンテンツ作成者に推薦する手法について提案する。

2. 課題の設定

本稿で想定するWebコンテンツのリンク先推薦システムの概要を図1に示す。まずコンテンツ作成者(以降、作成者)は、システムの画面左側に用意されたスペースにWebコンテンツのテキスト本文を記述する。本文作成中にリンクを作成する場合、リンクを作成すべき文字列の範囲をマウスで指定し、「リンク作成」ボタンを押下する。システムは指定された文字列に対するリンク先候補をWebから自動収集し、画面右側のスペースに推薦度合いの高い順に提示する。作成者がリンク先としたいリンク先候補の「選択」ボタンを押下すると、指定した文字列にリンクが自動的に生成される。

上記のリンク先推薦を実現するためには、(ア)作成者のリンク目的の推定、(イ)作成者のリンク目的に合致するコンテンツの推薦、の2種類の課題がある。

それぞれ重要な課題であるが、本稿ではリンク目的は作成者から明示的に与えられるとし、作成者のリンク目的に合致するコンテンツを推薦する手法を検討する。

3. リンク目的を意識したリンク先推薦手法

3.1 手法の概要

本稿では、作成者のリンク目的に合致するコンテンツを推薦する手法として、(1)検索クエリをリンク目的に応じて自動生成してリンク先候補を検索し、(2)取得した各リンク先候補のリンク目的への合致度を算出し、合致度順に並び替えて提示する手法を提案する。(1)ではリンク先の対象となるコンテンツはリンク目的に応じて異なるため、リンク先候補を取得するために用いる検索クエリをリンク目的に応じて自動生成する。また、(2)については、情報の鮮度や信頼度といった尺度は、リンク目的によって重視される度合いが異なるため、リンク目的に応じて尺度の重み付けを変更し、リンク目的への合致度を算出することとする。以下、提案手法の処理手順の概要を述べる。

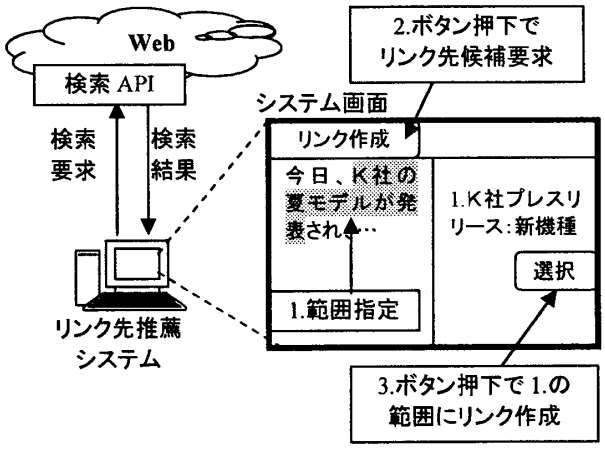

図1:リンク先推薦システムの概要

Step1. リンク目的に応じて検索クエリを自動生成する。

Step2. 検索クエリを汎用Web検索APIに入力し、検索結果のURLリストを取得する。次に、URLリストに含まれるコンテンツをダウンロードする。ここで得られたコンテンツが、リンク先候補となる。

Step3. ダウンロードした各リンク先候補を解析し、情報の鮮度や信頼度等の尺度ごとのスコアを算出する。

Step4. 各リンク先候補について、リンク目的に応じた尺度スコアの重み付き総和を算出し、リンク目的への合致度とする。

Step5. リンク先候補を合致度順に作成者に提示する。

作成者のリンク目的は様々考えられるが、本稿では以下の主要な4つのリンク目的を例として具体的な実現方法を説明する。

目的A： 指定した語句の意味や意義を説明するコンテンツの参照

目的B： 指定した語句に関連する事実・詳細を記載するコンテンツの参照

目的C： 指定した語句に関連する作成者自身が過去に作成したコンテンツの参照

目的D： 指定した語句に関連する企業やイベント等の公式サイトの参照

以降では、上記のリンク目的例を用い、提案手法の重要な処理であるStep1、Step3、Step4の詳細をそれぞれ3.2、3.3、3.4で述べる。

3.2 検索クエリの自動生成

表1に、リンク目的別の検索クエリ自動生成方法を示す。なお、表中の「作成箇所」は、リンクを作成するために作成者が指定した文字列の範囲を表し、「重要語」は作成中

† (株) KDDI研究所, KDDI R&D Laboratories Inc.

のコンテンツに含まれる単語(名詞やサ変動詞)のうち、TF-IDF 値が高い単語を表す。また「オプション」は検索時に指定する検索オプションである。OR 検索はいずれかの検索語を含むコンテンツ、AND 検索は全ての検索語を含むコンテンツの検索を行う。また、ドメイン指定は特定のドメインに限定した検索、更新日指定は最終更新日が指定期間内であるコンテンツのみの検索を行う。

表1：検索クエリの自動生成方法

目的	検索語	オプション
A	(i)作成箇所 (ii)作成箇所＋「とは」	OR 検索
B	(i)作成箇所中の重要語 (ii)作成箇所を含む文の重要語	AND 検索
C	目的 B と同じ	AND 検索 ドメイン指定 更新日指定
D	(i)作成箇所	

3.3 尺度スコアの算出

3.1 で挙げた4つのリンク目的において、重視される度合いが異なる以下の 5 つの尺度のスコアを、Step3 で取得した各リンク先候補に対して算出する。

尺度1) 情報の鮮度
コンテンツの新しさを表す尺度であり、最終更新日が新しいコンテンツほど高い値とする。

尺度2) 信頼度
信頼できると考えられるサイトに含まれるコンテンツほど高い値とする。具体的には、サイト URL のドメイン部分に”.org”、”.go.jp”等の特定組織を表す文字列が含まれている場合や、PageRank[3]が高い場合に信頼できると判定する。

尺度3) 時間表現の一致度
作成箇所に含まれる時間表現(「前日」等)をコンテンツ作成日時等の情報に基づき具体的日時に変換し、各リンク先候補の最終更新日と比較して、近いほど高い値とする。

尺度4) 内容の類似度
作成中コンテンツとリンク先候補の内容の類似度とする。ここでは、作成中コンテンツとリンク先候補に含まれる単語とその TF-IDF 値から特徴ベクトルを生成し、コサイン類似度により算出する。

尺度5) URL/タイトルの一致度
作成箇所に含まれる文字列が、リンク先候補の URL、またはタイトルと一致するほど高い値とする。

3.4 リンク目的への合致度算出

各リンク先候補のリンク目的への合致度を算出するため、リンク目的別に重視する尺度を以下のように仮定する。

目的A： 信頼度、情報の鮮度を重視する
目的B： 内容の類似度、信頼度を重視する
目的C： 時間表現との一致度、内容の類似度を重視する
目的D： URL/タイトルの一致度を重視する

これらの仮定に基づき、各尺度スコアの重みをリンク目的に応じて変更する。また、各リンク先候補のリンク目的への合致度は、尺度スコアの重み付き総和により算出する。リンク目的が a の時、i 番目の尺度の重みを $Wa(i)$、コンテンツ D の尺度 i の値を $Vi(D)$ とすると、コンテンツ D のリンク目的への合致度 $E(D)$ は次式により算出する。

$$E(D)=\sum_{i} Wa(i)\times Vi(D)$$

4. 評価実験

提案手法の有効性を検証するため、評価実験を行った。まず、リンク目的の分類の妥当性について検証するため、CNET Japan [4]のコラム記事(44 記事)に含まれるリンクを調査した。表 2 に、調査したリンク目的の分布を示す。3.1 で述べた 4 つのリンク目的は、全リンクの 92%を占めることから本稿で提示した 4 つのリンク目的はほぼ妥当であると考えられる。

表2：リンク目的の分布（CNET Japan コラム記事）

A	B	C	D	その他
7	44	12	38	9

次に、リンク目的への合致度の算出方法の妥当性を検証するため、CNET Japan の記事を利用して評価した。「依頼」という単語に目的 B でリンクが作成されていた記事 Ds[1]を例として挙げる。

まず「依頼」という単語のみを検索クエリとして Yahoo! 検索[2]に入力すると、この例での実際のリンク先 Dt[3]は上位 10 件には出てこなかった。ここで作成箇所を含む文章中に出てくる「警察庁」を追加し、「依頼」と「警察庁」との AND 検索を行うと検索結果 4 位として実際のリンク先 Dt が得られた。

続いて、記事 Ds はモバイルフィルタリング問題の経緯に関する記事であり、実際のリンク先 Dt はフィルタリング問題に関する総務省の報道発表であり、内容の類似度は高い。これに対し、「依頼」と「警察庁」との AND 検索の結果上位 3 件のコンテンツ(それぞれ脚注[4, 5, 6])の内容はフィルタリング問題に関連がなかった。よって、内容の類似度を重視することで、実際のリンク先 Dt を第1位として推薦することができる。以上の結果から、提案手法が有効に作用する見通しを得た。

5. まとめと今後の課題

本稿では、Web コンテンツ作成者に対し、リンクを作成する目的に応じて、そのリンク先候補となるコンテンツを Web から自動収集し、推薦する手法を提案した。提案手法では、リンク目的に応じて検索クエリを自動生成し、取得したコンテンツから算出される尺度スコアに対してリンク目的に応じて重み付けを行い、リンク目的への合致度を算出することとした。また、評価実験により、提案手法が有効に作用する見通しを得た。今後は、検索クエリの拡張方法および尺度スコアの算出方法の詳細化を進めるとともに提案手法を実装し、詳細な評価実験を実施する予定である。

参考文献

[1] はてなダイアリー　キーワードリンクとは：http://d.hatena.ne.jp/keyword/キーワードリンク
[2] 中村，吉永，鳥澤：Web から動的に獲得した属性/値を利用する文章作成支援環境，DBWeb (2007).
[3] L. Page, S. Brin, R. Motwani, T. Winograd : The PageRank Citation Ranking: Bringing order to the Web , Stanford Digital Library (1998).
[4] CNET Japan : http://japan.cnet.com/

[1] http://japan.cnet.com/mobile/filtering/story/0,3800085738,20370261-2,00.htm
[2] http://search.yahoo.co.jp/
[3] http://www.soumu.go.jp/s-news/2007/070216_5.html
[4] http://www.npa.go.jp/police/zoom/zoom02/kig01.htm
[5] http://www.nben.or.jp/11_seimei/20060519.html
[6] http://www.npa.go.jp/pressrelease/souni/furikomejusho.htm