

リンクの半自動生成に向けたアンカーテキスト推定手法

Estimation Method of Anchor Text for Semi-Automatic Hyperlink Generation

武吉 朋也†
Tomoya Takeyoshi

服部 元†
Gen Hattori

小野 智弘†
Chihiro Ono

滝嶋 康弘†
Yasuhiro Takishima

1. まえがき

Web コンテンツ作成者(以降, 作成者)は, 他のコンテンツへのリンクを作成することが多い. 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.

作成者によるリンク作成の負担を軽減する既存サービスとして, 特定のキーワードに対して, そのキーワードの解説が掲載されている Web コンテンツへのリンクを自動的に生成するサービス[1]がある. しかしながら, キーワードの解説のリンク目的のみに対応しており, 他のリンク目的に適用できない問題がある.

この問題を解決するため, 筆者らはこれまでに, 作成者がコンテンツ作成中にアンカーテキストとリンク目的を指定するとリンク先候補を自動的に提示する, リンク先推薦手法を提案した[2]. 本稿では作成者の負担を更に軽減するために, リンク先推薦手法を拡張し, 作成者がコンテンツ本文を入力するだけで, アンカーテキストとリンク目的を推定し, リンク先を設定するまでを全て自動で行う手法(以降, 提案手法)を提案する.

2. 利用シーンと課題の設定

本稿では, 作成者がコンテンツ本文を入力すると, 入力されたコンテンツ本文中でリンクを作成すべきテキストの範囲と, その範囲でのリンク目的を推定し, 適切なリンク先へのリンクを自動設定する自動リンク作成システムを目指す.

自動リンク作成システムを実現するためには, (ア)リンクを作成すべきアンカーテキストの候補抽出, (イ)適切なリンク目的の推定, (ウ)適切なアンカーテキストとリンク先の推定という, 3つの課題を解決する必要がある. なお本稿では, リンク目的として以下の4つを扱う.

- (a) 作成者自身が過去に作成したコンテンツへのリンク
- (b) 言葉の意味や意義を説明するコンテンツへのリンク
- (c) 企業等の公式サイトへのリンク
- (d) 関連する事実や詳細を記載するコンテンツへのリンク

3. 提案手法

本稿では, 課題(ア), (イ), (ウ)を同時に解決する手法を提案する. (ア)については, 統計的にアンカーテキストに含まれやすい単語を抽出し(Step1), その前後の単語の組み合わせ方を変化させ, 4つのリンク目的ごとに複数のアンカーテキストの候補を抽出する(Step2)方針とする. (イ)については, 4つのリンク目的とそれらに対応するアンカーテキストの候補の組み合わせそれぞれに対応するリンク先候補を取得する. それらのリンク先としての適切さ(以降, 評価値)をリンク目的間で比較し, 最適なリンク目的を求める方針とする(Step3). (ウ)についてはアンカーテキスト

の候補間で, リンク先候補の評価値を比較し, 最適なアンカーテキストとリンク先を求める方針とする(Step4).

上記の方針に従い, 提案手法は以下の4つの処理ステップで実現する. 図1に処理の流れを示す.

- Step1. コンテンツ本文中からアンカーテキストを推定する際に基点となる単語を抽出する.
- Step2. Step1 で抽出した単語の前後の単語を加えることで, 4つのリンク目的ごとにアンカーテキストの候補(以降, アンカーテキスト候補)を複数生成する.
- Step3. Step2 で生成したリンク目的とアンカーテキスト候補の組それぞれについて, リンク先候補を検索して取得する. 次に, [2]で提案したリンク目的に共通の予測モデルで各リンク先候補を評価する. リンク目的間で評価値の平均値の比較を行い, 最も高いリンク目的を採用する.
- Step4. [2]で提案したリンク目的専用の予測モデルを用いて, リンク先候補を再度評価する. アンカーテキスト候補間で評価値の平均値の比較を行い, 最も高いアンカーテキストを採用する. また, 評価値が最も高いリンク先候補を採用する.

以降, Step1 から Step4 の具体的な処理について, 3.1 節から 3.4 節でそれぞれ説明する.

3.1 Step1 基点となる単語の抽出

まず入力されたコンテンツ本文に形態素解析を適用し, 単語を抽出する. 抽出対象は, 名詞や動詞とし, 助詞や助動詞などの機能語は含めない. 次に, 統計的なアンカーテキストへの含まれやすさ(以降, 単語スコア)が所定の閾値以上である単語を抽出し, 次の Step2 に進む. 図1では, 「K社」が Step1 で抽出される単語としている.

単語スコアの算出には, 予めリンクを含む既存コンテンツを Web から収集し, 各コンテンツに含まれる単語, およびアンカーテキストに含まれる単語を集計する. それらを用いて, 対象の単語を含むコンテンツ数に対する, その単語をアンカーテキストに含むコンテンツ数の割合を算出し, 単語スコアとする.

3.2 Step2 アンカーテキスト候補の抽出

Step1 で抽出した単語それぞれについて, リンク目的ごとに以下の処理を行い, 複数のアンカーテキスト候補をコンテンツ本文から抽出する. ここでは, ある単語 W を基点とする場合について述べる.

- Step2-1. 単語 W をアンカーテキスト候補に追加する.
- Step2-2. 追加単語数を s とし, $(s+1)$ 個の単語からなるテキスト上の範囲で, t 番目の単語が W である範囲をアンカーテキスト候補に追加する.
- Step2-3. Step2-2. の処理を 1 以上 $(s+1)$ 以下の t について行う.

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

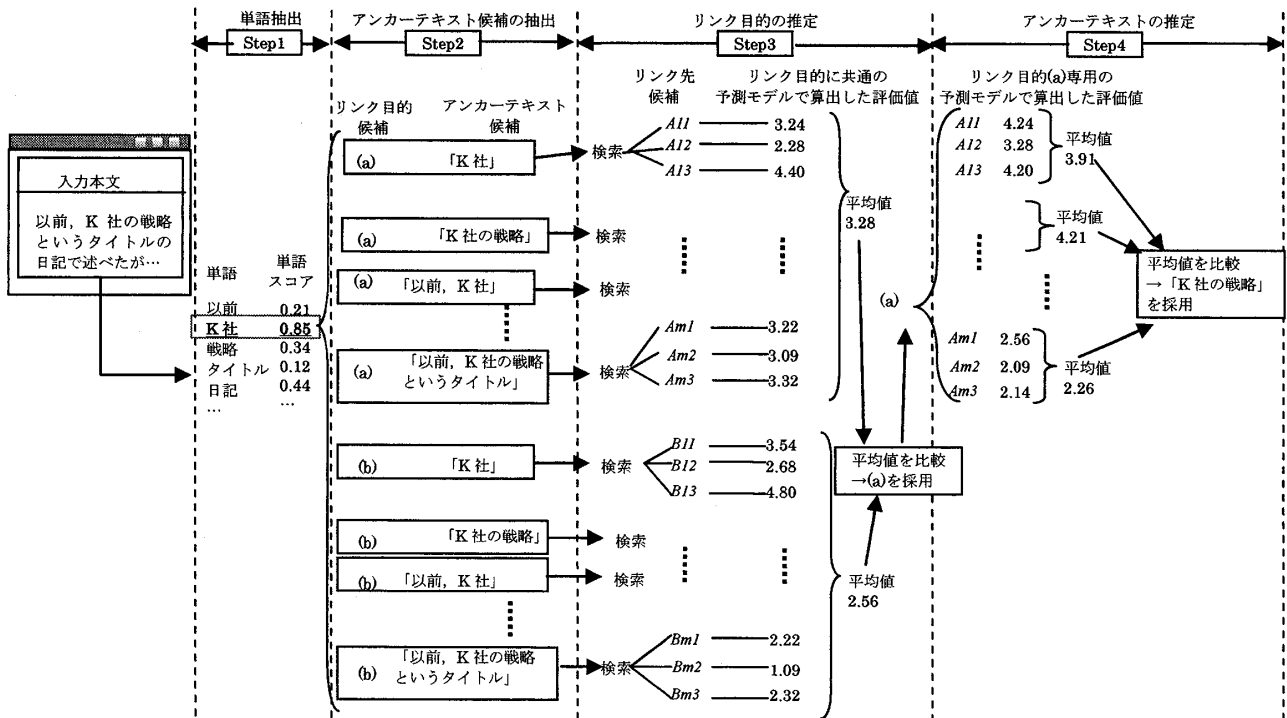


図1 提案手法の処理の流れ(図中の数値はダミー, リンク目的は(a)(b)の2つのみを記載)

Step2-4. Step2-3. の処理を1以上 n 以下の s について行う。

上記の n は, W に追加可能な単語数の上限とする。ただし, 追加できる単語は, 単語 W と同じ文に出現する単語までとする。図1では, 単語 W が「K 社」であり, 生成されるアンカーテキスト候補は, 「K 社の戦略」や「以前, K 社」などである。

3.3 Step3 リンク目的の推定

以降の処理は, Step2 で基点となった単語それぞれについて行い, 基点となった1つの単語に対し, 1つのリンク目的を決定する。まず, リンク目的とアンカーテキスト候補の組み合わせそれぞれについて, リンク目的に応じて検索クエリを拡張し, リンク先候補を取得する。図1では, リンク目的(a)でアンカーテキスト候補「K 社」に対するリンク先候補として $A11$ から $A13$ を取得している。

次に, 取得したリンク先候補のリンク先としての適切さ(評価値)を算出する予測モデルにより, 各リンク先候補の評価値を得る。ここで用いる予測モデルは, リンク目的に共通の予測モデルであり, 異なるリンク目的間の評価値の比較が可能である。この予測モデルは, 入力コンテンツの情報(アンカーテキストやコンテンツ本文など)とリンク先候補の情報(HTTP ヘッダやコンテンツ本文など)を入力として与えると, 評価値を算出するモデルである。リンク先候補の評価値をリンク目的ごとに集計し, その集計結果が最も高いリンク目的を採用する。図1では, $A11$ から $A13$ までの集計結果と $B11$ から $B13$ までの集計結果を比較し, リンク目的(a)を採用している。

3.4 Step4 アンカーテキストの推定

Step3 と同様に, 以降の処理は Step2 で基点となった単語それぞれについて行い, 基点となった1つの単語に対し, 1つのアンカーテキストを決定する。Step3 で取得した各リ

ンク先候補を, Step3 で採用されたリンク目的専用の予測モデルに入力し, 評価値を算出する。リンク先候補の評価値をアンカーテキスト候補ごとに集計し, 集計結果が最も高いアンカーテキスト候補を採用する。ここで用いる予測モデルは, 各リンク目的専用の予測モデルであり, 共通の予測モデルよりも厳密にリンク先候補を評価する。また, この予測モデルで算出した評価値に基づき, リンク先候補の提示順を決定する。図1では, $A11$, $A12$, $A13$ の集計結果から $Am1$, $Am2$, $Am3$ の集計結果までを比較し, アンカーテキスト候補「K 社の戦略」を採用している。

4. まとめ

本稿では, 他の Web コンテンツへのリンク作成を自動化するための手法を提案した。提案手法では, 入力されたコンテンツ本文中から, リンク目的とアンカーテキストを推定し, さらに適切なリンク先を推定する。この手法を応用することで, 一般のインターネット利用者が作成するコンテンツの質の向上が期待できる。今後の課題としては, 提案手法の各処理について評価実験を行い, 有効性を検証する。

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] 武吉, 服部, 小野, 滝嶋 : 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システムの実装と評価, IEICE SIG Notes, W12-2009-14 (2009.3).