URLの類似度に基づくWebページの信頼度算出手法

森祐樹† 鈴木優† 川越恭二†
†立命館大学 情報理工学部

1 はじめに

現在，Web 上に存在する大量の Web ページの中から信頼度の高い情報と低い情報を的確に判断するために，信頼性を推定する研究が行われている．信頼性を推定する手法の一つとして，各 Web ページの信頼度を算出する手法がある．既存研究では，リンク構造 [1] や文書内容の解析 [2] による信頼性の推定が行われている．リンク構造による研究では，blog の人気や更新頻度，支持率といった点からブロガーの信頼度を算出し，信頼できるブロガーのサイトから多くリンクを貼られたページは信頼できるページであるという考えにもとづいて，信頼性の推定を行っている．しかし，リンク構造による信頼性の推定では，リンクが貼られていないページに対して信頼性の推定が行われない．また，文書解析による信頼性の推定では，信頼性に影響する要因を決定する判断基準が困難である．この二つの点から十分な信頼度の算出が困難である．

そこで本稿では，Web ページが持つ URL に着目した．なぜならば，URL に含まれるドメイン情報やファイル名が類似している場合，その Web ページ同士の信頼性は類似しているのではないかと考えたためである．そこで，トライグラム・インデキシング [3] により抽出した URL の文字列の重みを要素とするベクトルから URL 間の類似度を算出し，算出した類似度から信頼度を算出する手法について提案を行う．その結果，Web ページのリンク構造や文書内容を考慮することなく高い精度で信頼度を算出することが可能となる．

2 URLの類似度に基づくWebページの信頼度算出

提案システムの処理手順を図 1 に示す．まず，あらかじめ他の利用者は，信頼できる Web ページ集合と信頼できない Web ページ集合を収集する．利用者は，問合せをシステムに入力する．システムは，問合せに対する検索結果の URL と，他の利用者が収集した Web ページの URL との類似度を算出する．問合せに対する検索結果の URL と他の利用者が収集した Web ページの URL との類似度を算出するために，図 1 の Step1 において，システムはトライグラム・インデキシングにより 3 文字単位で各 URL から文字列を抽出し，各文字列の重みを算出する．Step2 において，システムは Step1 で算出した URL の類似度に基づいた各 Web ページの信頼度を算出する．本稿では，信頼できる Web ページの URL に類似する Web ページは信頼度が高く，信頼できない Web ページの URL に類似する Web ページは信頼度が低いとみなす．Step3 において，システムは取得した検索結果を信頼度の高い順に再順位付けする．最後に，システムは再順位付けされた検索結果を利用者へ提示する．

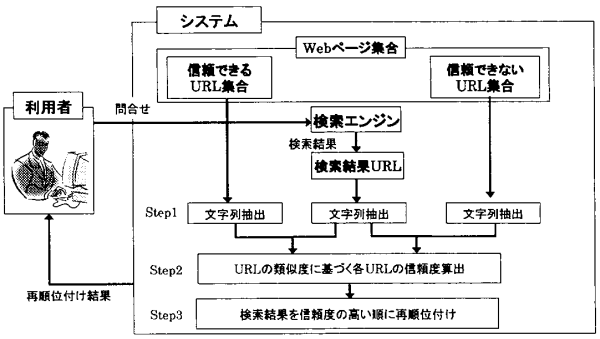

図 1: 提案システムの処理手順

2.1 トライグラム・インデキシングによる URL からの文字列抽出

トライグラムの考えに基づき，URL の “http://” 以降の文字列に対して，文字列の先頭から 1 文字ずつずらしながら，3 文字単位で文字列を抽出する．例えば，“http://www.yahoo.co.jp/” では，{ www, ww., w.y, .ya, ··· , co., o.j, .jp, jp/} のように 3 文字ずつの文字列を抽出する．ここで，各 URL の文字列に対する重みを算出するため，検索結果の URL に対して，文書検索で用いられる TF-IDF 法を利用する．信頼できる Web ページ集合と信頼できない Web ページ集合をそれぞれ

A Method to Calculate Credibility of Web Pages Based on the Similarity between URLs
Yuuki MORI†, Yu SUZUKI†and Kyoji KAWAGOE†
†College of Information Science and Engineering, Ritsumeikan University
†mori@coms.ics.ritsumei.ac.jp,
†{yusuzuki,kawagoe}@is.ritsumei.ac.jp

比較対象 Web ページ集合と呼ぶ．検索結果の URL と比較対象 Web ページ集合の URL が存在し，それぞれから文字列を重複なく抽出する．各文字列の重みを w を算出する式を (1) 式に示す．

$$w = -tf \cdot \log\left(\frac{df}{N}\right) \tag{1}$$

ここで，URL における文字列の出現頻度を tf，文字列が出現する URL 数を df とする．

2.2 ベクトル空間モデルを用いた URL の類似度算出

2.1 節において求めた文字列の重みを要素としたベクトルを生成する．x 件存在する検索結果 URL の k 番目 $(1 \leq k \leq x)$ における U_k のベクトル $\boldsymbol{u_k}$ は，検索結果 URL U_k の各文字列の tf 値と比較対象 Web ページ集合 URL の idf 値を掛け合わせた値から生成される．また，y 件存在する比較対象 Web ページ集合の URL の l 番目 $(1 \leq l \leq y)$ における U_l のベクトルを $\boldsymbol{u_l}$ とする．ベクトル $\boldsymbol{u_l}$ は，比較対象 Web ページ集合の URL から抽出された各文字列の tf 値と比較対象 Web ページ集合 URL の idf 値を掛け合わせた値から生成される．そして，生成されたベクトルを用いて，URL のベクトル間類似度を算出する．本稿では，類似度を算出する方法としてコサイン尺度を利用する．検索結果 URL U_k のベクトル $\boldsymbol{u_k}$ と比較対象 Web ページ集合の URL U_l のベクトル $\boldsymbol{u_l}$ とのコサイン尺度を $\cos(\boldsymbol{u_k}, \boldsymbol{u_l})$ とする．コサイン尺度を算出する式について，(2) 式に示す．

$$\cos(\boldsymbol{u_k}, \boldsymbol{u_l}) = \frac{\boldsymbol{u_k} \cdot \boldsymbol{u_l}}{\|\boldsymbol{u_k}\|\|\boldsymbol{u_l}\|} \tag{2}$$

ここで，a 件存在する信頼できる Web ページ集合 URL の m 番目 $(1 \leq m \leq a)$ である U_{lm} のベクトルを $\boldsymbol{u_{lm}}$ とする．また，b 件存在する信頼できる Web ページ集合 URL の n 番目 $(1 \leq n \leq b)$ である U_{ln} のベクトルを $\boldsymbol{u_{ln}}$ とする．$\boldsymbol{u_k}$ と $\boldsymbol{u_{lm}}$ のコサイン尺度を求める場合は $\boldsymbol{u_l}$ に $\boldsymbol{u_{lm}}$ を代入する．また，$\boldsymbol{u_k}$ と $\boldsymbol{u_{ln}}$ のコサイン尺度を求める場合は $\boldsymbol{u_l}$ に $\boldsymbol{u_{ln}}$ を代入する．

2.3 Web ページの信頼度算出

2.2 節において算出した類似度に基づいて Web ページの信頼度を算出する．比較対象 Web ページ集合の l 番目である URL と検索結果 URL U_k 間の信頼度を $\alpha_{kl} = \cos(\boldsymbol{u_k}, \boldsymbol{u_l})$ とする．信頼できる Web ページ集合の URL と比較した際の検索結果 URL U_k の信頼度を S_k とする．そして，信頼できない Web ページ集合の URL と比較した際の検索結果 URL U_k の信頼度を V_k とし，S_k と V_k の算出式を (3) 式に示す．

$$\begin{aligned} S_k &= \Sigma_{m=1}^{a} \alpha_{km} \\ V_k &= \Sigma_{m=1}^{b} \alpha_{kn} \end{aligned} \tag{3}$$

(3) 式より算出された値を要素とした信頼度ベクトル $\boldsymbol{c_k} = [S_k\ V_k]$ を生成する．S_k の最大値を $\max S_k$ とする．そして，信頼度ベクトル $\boldsymbol{c_k}$ と信頼できるベクトル $\boldsymbol{s_k} = [\max S_k\ 0]$ との類似度をコサイン尺度を用いて，検索結果 URL U_k の信頼度 C_k を (4) 式に示す．

$$C_k = \frac{\boldsymbol{c_k} \cdot \boldsymbol{s_k}}{\|\boldsymbol{c_k}\|\|\boldsymbol{s_k}\|} \tag{4}$$

3 評価実験の方法

本章では，提案手法の信頼性の精度を測定するために行う評価実験の方法について説明する．Yahoo! Japan デベロッパーネットワーク[1]の Web 検索 API を利用する．まず，研究室の学生 10 名に協力をしてもらい，信頼できる Web ページと信頼できない Web ページを 100 件ずつ収集する．また，Yahoo! Japan の検索結果上位 100 件を取得する．検索結果上位 100 件を取得する問合せとして，本実験では「内閣 解散日」,「サッカー 移籍」,「消費税 引き上げ」という 3 件の問合せを用いる．そして取得した検索結果を対象として実験を行う．信頼性の精度を測定する方法として，実際の Yahoo! Japan における検索結果および提案手法で再順位付けされた結果における 11 点平均精度の比較を行う．

4 おわりに

本稿では，URL の類似度に基づいて，Web ページの信頼度を算出する手法について提案した．今後は 3 章の実験を行い，実験結果から考察することによって，本提案手法の信頼性の精度を測定する．また，信頼できる Web ページと信頼できない Web ページの判断基準は個々によって異なるため，個々の判断基準に適した信頼度の算出方法を検討する予定である．

参考文献

[1] 中島伸介, 竹原幹人, 舘村純一, 日野洋一郎, 原良憲, 田中克己："blog 解析に基づく Web 情報検索の信頼性向上技術", 人工知能学会第 6 回セマンティックウェブとオントロジー研究会, SIG-SWO-A401-05, **11**, (2004).

[2] 福島隆寛, 内海彰："Web ページの信頼性の自動推定", 知能と情報, **19**, 3, pp. 239–249 (2007).

[3] E. Baykan, M. Henzinger and I. Weber: "Web Page Language Identification Based on URLs", In Proc. of VLDB, pp. 176–187 (2008).

[1] http://developer.yahoo.co.jp/