リンクパス情報に基づく自動フィルタリング手法の提案

3X−01

大森 信行　楢原 佳織　宮本 勝　杉崎 正之　牧野 俊朗　稲垣 博人
日本電信電話株式会社　NTTサイバーソリューション研究所

1　はじめに

現在、インターネットWWWサービスを利用することにより、容易に情報発信が可能となっており、数十億オーダのWEBページが存在すると言われている。その中には、多様な情報が発信されており、「見る必要がない」情報が含まれているために、設定されたレベルに応じて情報を選択的に受信する、すなわちコンテンツが一定の指標に一致するかどうか判定し、判定結果によってコンテンツの閲覧を制御するフィルタリングを行う必要が生じている。

例えば、教育分野向けの例では検索エンジンの検索結果から教育に関連するコンテンツだけを選択する場合などである。

フィルタリングの基準として、ルールやプロファイルを人手で設定や維持する必要があるタイプ[1]が多かったが、我々の提案するシステムでは、ハイパーリンクのパス情報に基づいて自動的にフィルタリングを行う。

2　システム構成

本システムの概要を図1に示す。本システムは、以下の機能から構成されている。

(1) ハイパーリンク抽出部
(2) リンク情報データベース
(3) ページ得点計算部

ハイパーリンク抽出部では、収集したWEBページからリンク情報を抽出する。抽出したリンク情報は、リンク情報DBに登録され、ページ得点計算部でのページ得点の計算に利用される。以下、詳細に説明する。

3　リンクパス情報に基づくフィルタリング

本稿で提案するリンクパス情報に基づくフィルタリングでは、製作者によってWEBページ内に設定されたハイパーリンク情報に基づいて、ページを得点によって格付け(レイティング)し、その結果に基づいてフィルタリングを行う。

ページの得点は、指定されたページとの関連度によって計算する。関連度はリンクを用いて、次のような考え方に従って計算する。

> あるページから直接リンクされているページは関連度が大きく、リンクをたどり離れたページほど関連度が小さくなる。またページ間の関連度は、2つのページをリンクで結ぶ経路数が多いほど、大きくなる

ページ得点の計算手順は以下の通りである。ページ得点は、そのコンテンツとあるドメインとの関連度を示す指標である。教育に利用する例では、得点が少ないほど教育に関連の大きなコンテンツであることを示す。ページの得点が予め定められた閾値を越えていれば、フィルタリングによる削除の対象となる。

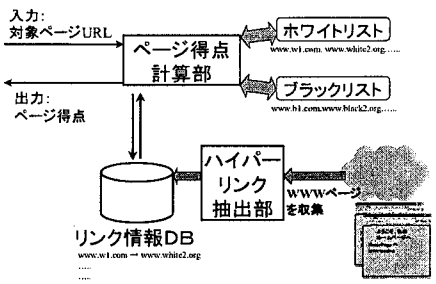

図1: リンクパスに基づくフィルタリングシステム

(1) シードURLの登録
(2) リンク情報の登録
(3) 対象URLとホワイトリストとの関連度計算
(4) 対象URLとブラックリストとの関連度計算
(5) ページ得点の計算

3.1　シードURLの登録

ユーザが対象ページとの関連度を計算するために指定したURLをシードURLといい、ページ得点計算部でのページの得点計算の基準となるURLである。今回は、ホワイトリスト、ブラックリストの2種類のシードURLを登録する。

A study on content filtering method based on hyperlink-path infomation.
Nobuyuki OHMORI, Kaori NARAHARA, Masaru MIYAMOTO, Masayuki SUGIZAKI, Toshiro MAKINO and Hirohito INAGAKI
NTT Cyber Solutions Laboratories,NTT Corporation

ホワイトリスト $G1$ には、教育に適すると判断され、閲覧を推薦されるサイトのURLが登録され、ブラックリスト $G2$ には、教育に関係がないと判断され閲覧が不要であるとされるサイトのURLが登録される。

3.2 リンク情報の登録

ハイパーリンク抽出部では、収集したインターネットのWWWページから、ハイパーリンクを取り出し、リンク情報DBに登録する。ハイパーリンクは、リンクが設定されているページのURL(リンク元URL)と、リンクが指し示すURL(リンク先URL)で表すことができる。DBには、リンク一つにつき

- リンク元URL：リンク先URL

という一行が登録される。

登録時には、リンク情報の重複登録のチェックを行う。

3.3 ホワイトリストとの関連度計算

ある対象ページのURL URL_{obj} が入力される。ホワイトリストに登録されているURL群 $G1$ と、対象ページ URL_{obj} との関連度 $R(URL_{obj}, G1)$ を計算する。

関連度 $R(URL_{obj}, G1)$ は、URL_{obj} と $G1$ 内の各ページとを結ぶリンクのパスの数、およびそのリンクパスの距離によって計算する。

具体的には以下の計算式で求める。

$$\text{関連度} R(URL_{obj}, Gi) = \frac{\text{パス数}}{\text{パス距離合計}}$$

パス数は、URL_{obj} と $G1$ 内各ページとを結ぶリンクのパスの数を合計したものである。

ここで、リンクパスは始点URLから終点URLまでの連続した複数のハイパーリンクによって表現された経路である。リンクパスに含まれるリンクを始点URLから順次たどることで終点URLに到達する。

リンクパスの距離は、リンクパスの始点から終点までに含まれるリンク数である。例えば、URL_{obj} からリンクを1回たどって到達する場合は距離1、2回たどる場合は距離2、となる。

関連度計算に必要なパス数、パス距離合計は以下のように計算する。

$$\text{パス数} = \sum_{i=1}^{n} route(URL_{obj}, URL_i)$$

ここで、n は $G1$ に含まれるURL $(URL_1, URL_2, \ldots, URL_n)$ の数を表す。$route(URL_{obj}, URL_i)$ は URL_{obj} と URL_i を結ぶリンクパスの数である。

$$\text{パス距離合計} = \sum_{j=1}^{n} \sum_{k=1}^{route(URL_{obj}, URL_j)} dist(path_k)$$

$dist(path_k)$ は、あるリンクパス $path_k$ の距離を表す。$path_k$ は、URL_{obj} と URL_i を結ぶ k 番目のリンクパスである。

パス数およびパス距離合計は、リンク情報DBからハイパーリンクを順次検索し2URLを結ぶ全てのリンクパスを探索することにより求める。この際に、距離が探索上限値を越えるリンクパスは探索結果からは除く。

3.4 ブラックリストとの関連度計算

次にブラックリスト・グループに登録されているURL群 $G2$ と、対象ページ URL_{obj} との関連度 $R(URL_{obj}, G2)$ を計算する。関連度は上記と同様に計算する。

3.5 ページ得点計算

入力されたURLのページがフィルタリングの対象であるかどうかを判定するために、ページのスコアを計算する処理である。対象ページ URL_{obj} の得点 $score$ を以下の式で計算し、出力する。

$$score(URL_{obj}) = R(URL_{obj}, G2) - R(URL_{obj}, G1)$$

4 評価実験

ホワイトリスト、ブラックリストにそれぞれ300程度のURLを登録し、リンク情報データベースには、収集した数千万ページから抽出したリンクを登録した。距離の探索上限値を5程度と設定したとき、1ページの得点計算にかかる時間は、数分から数十分であった。

また、実験の結果、ブラックリストに含まれるページからのリンクの距離が大きくなるほど、閲覧に適すると判断されたページが増えてくる傾向が得られ、リンクパスによってページ得点を計算する本手法の有効性が確認できた。

5 まとめ

リンクパス情報に基づいてページの得点を計算し、フィルタリングする手法を提案した。今後は、リンクパス探索速度の向上により実時間処理に向け手法を検討する。また、定量的な評価、他の手法との比較を行っていく予定である。

参考文献

[1] 森田 他, 情報フィルタリングシステム, 情報処理学会ジャーナル Vol.37, No.8, 1996.