地理情報を用いたニュースのフィルタリング機構

7U-03

松本 知弥子[†]　馬 強[††]　田中 克己[††]
[†] 神戸大学 工学部　[††] 神戸大学大学院 自然科学研究科

1 まえがき

Web 情報検索や情報フィルタリングにおいては, 従来, 主にキーワードやユーザプロファイルを用いて検索・フィルタリングが行われている. 本稿では, これらのアプローチとは異なり,**Web 情報のローカル度**に基づいた検索・フィルタリング機構を提案する. これは,Web 情報がどの程度地域やコミュニティに特化したローカルな情報なのかという視点が, 情報検索や情報フィルタリングの精度向上に有効であると考えるためである. 関連研究として,NTT ソフトのモバイルインフォサーチのように, 位置情報をもつ Web 情報を統合する試み [4, 5] や,Clever[1] にようなリンク構造情報からのコミュニティ発見の試み [2, 3] があるが, 本稿のようなローカル度にもとづく情報フィルタリングという試みは存在しない.

Web 上のローカルな情報として, 地域情報とコミュニティでの話題を考えることができる. 地域名や組織名などの地理情報が多く含まれているページや, 一部のユーザの集まりであるコミュニティでの話題はローカルな内容である可能性が高い.Web 情報を取得する上で, ユーザがどの程度のローカル度をもつ情報に興味があるかどうかは, ユーザによって大きく異なる. いずれにしろ,Web 情報のローカル度をいかにして決定するか, その概念をどのように定義するかという問題自体, 曖昧であり明確な基準がないのが現状である. そこで, 本稿では, まず,Web 情報中にどの程度地理的用語が含まれるかどうかにしたがってローカル度を定義する. さらに,Web ページの有するリンク構造を考慮し, そのページと他のローカルなページとの間の参照関係を考慮した形に拡張してローカル度を定義する. 地理的用語の含有割合によるローカル度の定義に関しては, 実験結果についても触れる. 最後に, 現在作成中の, ローカル度を用いたニュース情報のフィルタリング機構についても報告する.

"A News Filtering Mechanism by Geographical Infomation"
[†]Chiyako Matsumoto
Faculty of Engineering, Kobe University
[††]Qiang Ma, Katsumi Tanaka
Graduate School of Science and Technology, Kobe University

2 地理用語の割合によるローカル度

2.1 定義

ローカルに依存した Web 資源は, 地理用語を多く含んでいることから, ページの内容を自然言語処理ツールを用いて品詞分解し, 記事 art_x に書かれている名詞の総数 $totalword(art_x)$ に対する, 地理用語 $geoword_i$ の割合から, ローカル度 $Lcl_g(art_x)$ を計算する.

計算式は,$totalword(art_x)$ を n とすると,

$$Lcl_g(art_x) = \frac{\sum_{i=1}^{n} weight(geoword_i)}{totalword(art_x)} \qquad (1)$$

である.

$totalword(art_x)$ に代名詞は含まれない. このとき, 式 (1) では, 各地理用語に重み $weight(geoword_i)$ をつけて, ローカル度を求める. 重みは, 地域国名 < 組織名 < 地域一般名の 3 段階でつけている. 地域一般名の中でも県名とその他の重みは変えている. 組織名は, 会社名・学校名・団体名などである.

2.2 評価

式 (1) を用いて,`http://iij.asahi.com/` の記事の 1 日分約 100 ページのローカル度を計算し, 筆者の判断で正解記事を選び, 適合率と再現率を求めた. その結果, 再現率は 0.26 で, 適合率は,0.33 という結果になった.

地理用語の重み付けにより, 地理用語の数は同じでも内容によってローカル度が異なり, 正しい判定ができることが多い.

この方法の問題点は, 記事の話題が組織名に関するものの場合, ローカルに依存した記事と, そうでないも

のがあり, 判定が失敗することがあることである. また, 長い記事の場合, 名詞の総数が多く, 相対的に地理用語の割合が小さくなり, ローカルに強く依存していると判断されるページのローカル度が小さくなることである.

3 リンク構造を考慮した地理用語の割合によるローカル度

2の結果から, 新たなローカル度の定義について述べる. Web資源のリンク構造 [2, 3] を考慮すると, 図1のように, ページ A は, 2.1で定義したローカル度の高いページから参照されており, ローカル度が高いと考えられる. ローカル度の低いページはその反対である. そこで, ローカル度をリンク構造を利用して定義する.

本稿では, ページ作者により無関係な参照リンクが意図的に張られて, ローカル度が左右されるのを防ぐため, 参照されているリンクのみを考慮する.

記事 art_x が, ページ $p_1, p_2, ..., p_n$ の n 個のページからからリンクを貼られているとき, ローカル度 $Lcl_l(art_x)$ を,

$$Lcl_l(art_x) = \frac{Lcl_g(p_1) + Lcl_g(p_2) + ... + Lcl_g(p_n)}{n} \quad (2)$$

と定義する.

つまり, 参照されているページのローカル度の平均を求めている.

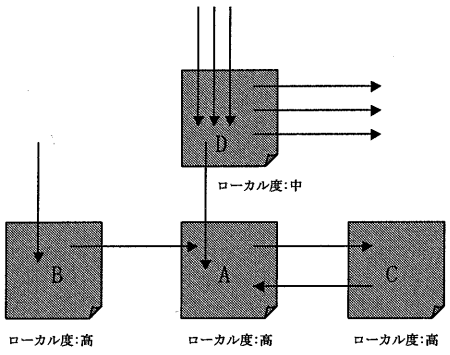

図 1: ページのリンク構造とローカル度

4 ニュースのフィルタリング

ニュースページのローカル度を上記の定義を用いて計算し, ユーザの興味に応じてフィルタリングを行う.

4.1 記事の詳細度

ニュースには, 全国版と地方版が存在し, 各々に役割がある. そのどちらにも取り上げられているニュースがある場合, 地方版のほうが詳細度が高い. 全国版の中のローカル度の高いニュースから, 地方版の同じ話題のニュースにリンクを生成すると, より詳細な記事をたどることができる.

4.2 ローカルな情報の排除

検索エンジンなどでページを探して, 一般的な情報を得たい場合, ローカルに依存した情報を排除したいことがある. その場合, ローカル度の高いページは検索結果から外す.

4.3 複数記事を比較

地理用語を多く含んで, ローカル度が高いと判定されても, 全国的な話題になることもある. このことから, 記事内の地理用語以外の単語を複数記事間で見比べて, 出現回数が低いものは, 全国的な話題になる可能性が高いものであり, 出現回数が高いものは, どこにでもある情報であると考えられる.

5 まとめ

Web資源のローカル度について定義し, ニュースのフィルタリングへの応用について述べてきた. 今後は, ローカル度の定義をさらに見直し, 実験による評価を行い, ローカルな情報を個人のニーズに合わせて配信する方法などを検討していく予定である.

参考文献

[1] Clever Project: http://www.almaden.ibm.com/cs/k53/clever.html,1999

[2] Sergey, Brin. and Lawrence, Page.: Google Web Search Engine, http://www.google.com/,2000

[3] Sergey, Brin., Lawrence, Page.: The Anatomy of a Large-Scale Hypertextual Web Search Engine,1998

[4] NTT ソフトウェア研究所: モーバイルインフォサーチ2実験, http://www.kokono.net/,2000

[5] 三浦 信幸, 高橋 克巳, 横路 誠司, 島 健一.: 位置指向の情報統合 ～モーバイルインフォサーチ2実験～, 情処 第57回 全大 (第3分冊 pp.637～638) ,1998