特定事業・分野の文書集合を利用したニュース記事収集システム

鈴木 康祐 岡本 東 堀川 三好 菅原 光政
岩手県立大学大学院ソフトウェア情報学研究科

1. はじめに

昨今，個人の嗜好や特定の目的に基づいて記事を収集して閲覧者に紹介するニュースサイトが注目を集めている．新聞社などが自社のニュースを提供するサイトはメディアニュースサイトと呼ばれるのに対し，個人が運営するサイトはパーソナルニュースサイトと呼ばれる．パーソナルニュースサイトの特徴として，メディアのニュース記事を引用するだけではなく，専門的なホームページやブログを紹介するケースが多く見受けられる．

しかしながら，ウェブ上には膨大な記事があるため，そのような記事の収集・整理を，運営者が手作業で行うことは負担となる．また，記事の選定は運営者が主観で行うため，特定の事業・分野における閲覧者にとっての重要な記事を網羅することは難しい．

そこで，本研究では，閲覧者から得られた書き込みなどの文書集合から特徴語を抽出し，検索活動に反映させる仕組みを備えたシステムを提案する．これにより，特定事業・分野向けニュースの記事の収集を効率化する．

2. パーソナルニュースサイトの問題点

(1)ウェブサイトの膨大さ

時事性の高いニュース記事は，メディアや他ニュースサイトから安定して得ることが容易だが，専門性の高いホームページやブログなどの記事は手作業で探す必要があり，運営者の負担となってしまう．

(2)閲覧者のニーズの明確化が困難

提供する記事の選択は，運営者の主観に依るものが大きい．そのため，閲覧者にニーズのある重要な記事を見落としてしまう可能性がある．

3. ニュース記事収集システム

3.1. システムの概要

本システムでは，Google や Yahoo といった既存のウェブ検索エンジンから特定事業・分野において価値のある記事を収集する．記事の収集には，運営者の目的・嗜好だけではなく(図 1)，特定事業や分野における情報共有基盤から得られた内部テキストを利用する(図 2)．内部テキストとは，CMS (contents management system) によって登録されたデータや掲示板の書き込み，ニュースに対するコメントといった閲覧者が登録した文書集合を指す．内部テキストを分析して得られた特徴語と閲覧者のニーズには深い関連があると考えることができる．

図 1. 既存の記事収集

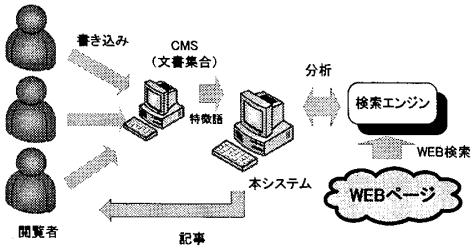

図 2. 提案するシステムの概要

また，特徴語を利用して得られた文書と内部テキストの内容の類似度を基に比較を行うことで，フィルタリングによる再現率の向上と共に，閲覧者の潜在的なニーズに合った記事を選定する．

3.2. 記事収集の流れ

本システムでは，閲覧者から得られた文書集合を内部テキストとし，特徴語 k を抽出する．その特徴語 k を基に検索を行い，得られた記事を内容の類似度に基づいて適合・不適合に分類する．具体的な分析の手順を以下に述べる．

Step1:対象分野・事業の内部テキストとなる文書集合を得る．文書集合は，特定事業・分野における電子掲示板やコミュニティサイトとのインタラクションを前提とし，自動的に取得する．

Step2:内部テキストから特徴語 k を抽出する[1]．抽出には $tf-idf$ の式(1)用いる．$tf-idf$ とは，テキスト中の単語に着目し，その出現回数や出現範囲などから重要度を設定する手法である．

$$t_{ij} = \log\left(1+f_{ij}\right)\cdot\log\left(\frac{n}{n_i}\right) \quad \ldots(1)$$

t_{ij} ：文書 T_j に対する単語 w_i の重み

f_{ij} ：単語 w_i の文書 T_j における出現頻度

n ：文書集合中の文書数

n_i ：単語 w_i を含む文書数

Step3:内部テキストから特徴語ベクトル，得られた記事から記事ベクトルを生成する．まず，Step2で抽出した特徴語 k と共起度 kw_i が一定以上である

Development of news article collection system using document set on specific business field
Kousuke SUZUKI, Azuma OKAMOTO, Mitsuyoshi HORIKAWA, Mitsumasa SUGAWARA
Faculty of Software and Information Science, Iwate Prefectural University

単語 w_i の重みを q_i としたとき，特徴語ベクトルは $q=(q_1,q_2,q_3 \ldots q_m)$ と表される．なお，共起度 kw_i は式(2)の jaccard 係数によって求められる．

$$kw_i = \frac{|k \cap w_i|}{|k \cup w_i|} \ldots (2)$$

Step4:対象事業・分野名および特徴語によるウェブ検索を行い，記事 $D_1, D_2 \ldots D_n$ を得る．

Step5:次に，得られた記事 D_n に対し形態素解析を行い，m 個の単語 $x_1, x_2 \ldots x_m$ を得る．単語 x_m の文書 D_j における重みを d_{mj} としたとき，ある文書 D_j の記事ベクトルは $d_j=(d_{1j}, d_{2j}, d_{3j} \ldots d_{mj})$ と表される．

Step6:内容の類似度に基づいた適合・不適合の判定を，得られた特徴語記事ベクトル d_j とベクトル q との比較によって行う．類似度の算出には式(3)のベクトル空間モデルにおけるコサイン尺度を用いる．類似度の値 $\cos(d_j, q)$ が大きいものを，特徴語 k における適合記事とする．

$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{j=1}^{m} d_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}} \ldots (3)$$

4. 数値実験

4.1. 目的

提案した手法の有効性の確認のため，特徴語を用いて記事を収集し，その内容の類似度に基づいて適合・不適合を判定する数値実験を行った．事例として，学童保育事業を対象とした記事の収集を行った．

4.2. 対象

分析の基となる特徴語は，岩手県学童保育情報サイト[1]から抽出する．岩手県学童保育情報サイト（以下，情報サイトと呼ぶ）は，岩手県内の学童保育における情報共有の発信・共有の支援を目的に作られた CMS 機能を持つポータルサイトである．情報サイトには，事業におけるお知らせ，イベント，掲示板の書き込みといったテキストデータが，各保育所の利用者によって登録されている．

4.3. 特徴語の抽出

実験では 2008 年 12 月の登録データを内部テキストとして扱い，特徴語を抽出した[2]．特徴語には「クリスマス」や「冬休み」といった 12 月を表す単語が得られた．

4.4. 共起語を用いた分類

事業における特徴語に関連する記事を，既存の検索エンジンを用いて検索した結果 100 件をデータセットとした．特徴語は 2008 年 12 月に重みの最も高かった「クリスマス」とし，共起語は共起度の高かった「サンタクロース」「親子」「ゲーム」といった 48 単語を利用する．学童保育事業に関連する記事でかつ「クリスマス」の話題であるものを適合とし，さらにその中でも，事業関係者に共通して有益であると思われるニュースを興味文書として分類した（図 3）．適合記事の中には，特定層に向けた記事・掲示板も多く見受けられたが，一方で「学童保育のクリスマス会の運営に関わる相談」や「クリスマスに関する地方のニュース」また「クリスマスで行われた具体的な出し物」など，事業関係者が共通して興味を持ちやすい内容の興味文書も得ることができた．また，不適合記事は 0.01 以下に集中しており，フィルタリングとしての効果が期待できる．

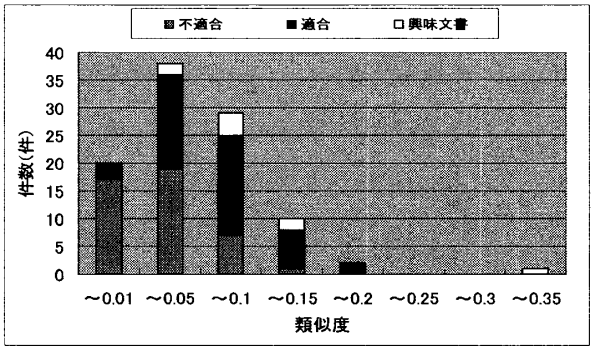

図 3. 類似度の分布

4.5. 考察

システムを用いることで，記事収集の負担が軽減できることを確認した．しかし，事業関係者に共通で価値のある記事は，適合文書のうちの 18%程度であった．今後は特徴語を複数設定し，幅広いジャンルの記事を収集していく必要がある．

5. おわりに

本研究では，パーソナルニュースサイト運営者が効率的に記事を収集する仕組みを提案した．閲覧者から得られた特徴語を検索に反映し，さらに内容の類似度を考慮したフィルタリングを行うことで，一般的な検索では得がたいブログやホームページ，幅広いジャンルのニュース記事などの入手が容易になる．しかしながら，特徴語とその共起語の抽出の際には，基となる文書集合が充分になければならないといった問題点もある．今後は，閲覧者の登録データ以外の文書集合を利用することを考慮する．また，ニュースサイト運営における記事の収集だけではなく整理といった観点からアプローチしていきたい．

参考文献

[1]舘澤千尋，岡本東，堀川三好，菅原光政：「学童保育を対象としたコンテンツ管理システム」，情報処理学会第 69 回全国大会講演論文集，分冊 4，pp.151-152 (2007)

[2]鈴木康祐，岡本東，堀川三好，菅原光政：「学童保育情報サイトにおけるテキストマイニングの活用」，情報処理学会第 70 回全国大会講演論文集，分冊 4，pp.721-722 (2008)