5L-09

Web Index における個別カスタマイズ可能なライブラリ構築システムの提案

大島　拓也† 遠山　元道††

†, ††慶應義塾大学理工学部情報工学科　〒223-8522 神奈川県横浜市港北区日吉 3-14-1

E-Mail: †nanadama@db.ics.keio.ac.jp ††toyama@ics.keio.ac.jp

1. はじめに

インターネットを利用し情報を検索するにあたり，近年最も広く利用されているのが検索エンジンである．検索エンジンを用いた情報検索は，検索結果として得られた Web ページ内の単語の詳細情報を得たい場合には，新たに検索エンジンで再検索し，選択する必要がある．Web ページ作成者はハイパーリンクによって Web ページに関連する複数の情報を予め結合することが可能だが，これは Web ページ作成者が意図した関連のみで結合され，必ずしも閲覧するユーザのニーズに応えているとは限らない．そこで，著者らは Web における利用者主導による情報資源結合を実現するために，Web IndeX(WIX)システムという情報資源表現形式の提案，開発を行っている[1][2].

本研究は，WIX システムにおいて，ユーザがユーザ自身の興味のある単語について，それに関連するニュースサイトの記事のような日々追加されていく Web ページへアクセスすることを可能にするための，WIX システム用のライブラリの更新システムを提案するものである．

2. Web Index システム

2.1. WIX システムと WIX ファイル

キーワードと，ターゲットである URL の組み合わせをエントリと呼ぶ．エントリの集合を XML 形式でまとめて記述したものを WIX ファイルという．WIX システムは，Web ページの文章中に出現するキーワードを， WIX ファイル中のエントリに基づき URL へのハイパーリンクに変換するものである．ハイパーリンクへの変換をアタッチと呼ぶ． WIX システムのユーザは，ユーザ自身が利用したい WIX ファイルをブックマークとしてあらかじめ登録しておき，ページ閲覧中にアタッチすることができる．

Proporsal of Personal Customizable Library Construction System on Web Index

† Takuya Ohshima, †† Toyama Motomichi

†, †† Keio University Faculty of Science and Technology Department of Information and Computer Science

2.2 現在提供されている WIX ファイル

現在，WIX システムにおいて提供されている WIX ファイルの種類は以下の二種類に限られている．

(1) URL が，対応付けられるキーワードそのものを含んでいるエントリからなる WIX ファイル

例として，Wikipedia の見出し語をキーワードとし，当該単語の記事の URL をターゲットとした WIX ファイルが挙げられる．

(2) URL が，対応付けられるキーワードに対し変化の小さいエントリからなる WIX ファイル

例として，企業の名称をキーワードとし，当該企業の公式 Web サイトのホームページの URL をターゲットとした WIX ファイルが挙げられる．

現在の WIX システムで提供されているこれらの WIX ファイルでは，ニュースサイトの記事の Web ページのような，日々追加されていくコンテンツに直接アクセスすることができない．このようなコンテンツは Web 上にある情報資源の中でも，リアルタイム性，速報性があるという点で重要なものであるため，それらの Web ページへのアクセスを WIX システムにおいて可能にすることは必要性の高いことである．

3. 自動ライブラリ更新システム

3.1. システムの概要

自動ライブラリ更新システムは，主にニュースサイトから配信される記事ページのような，日々追加されていく Web ページに「結合」することが可能な WIX ファイルの生成，自動更新を可能にするシステムである．本システムの導入により，WIX システムのユーザは，ユーザ自身が興味を持つキーワードに関連する，ニュース記事のような Web ページのうち最新のものに簡単にアクセスすることが可能になる．

本システムでは，以下の 3 ステップによって WIX ファイルの自動更新を行っている．

(1) キーワード候補の登録

(2) RSS ファイルの登録と RSS アイテムの抽出

(3) キーワード候補単語と RSS アイテムからエントリ生成

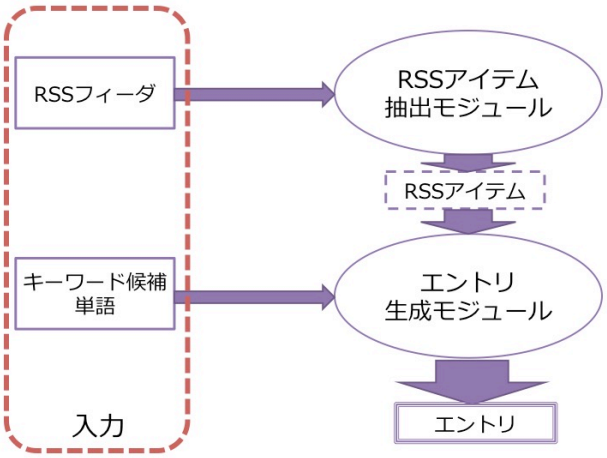

図 1 システム概要図

システムの概要とデータの入出力を図 1 に示す.

3.2. キーワード候補の登録

エントリのキーワードの候補となる単語をキーワード候補としてデータベースに登録する. キーワード候補の単語群には, 助詞や代名詞などのストップワードが含まれないようにする必要がある. 本システムでは, キーワード候補の単語群として, 日本語版 Wikipedia の見出し語約 200 万語から, 一部ストップワードや数字のみの見出し語などを取り除いて登録している.

3.3. RSS ファイルの登録と RSS アイテムの抽出

本システムでは, 日々追加されていく Web ページを収集する方法として, ニュースサイトなどから提供されている, RSS を利用している.

本システムでは, RSS フィーダから配信される RSS ファイルの URL を登録し, RSS アイテム抽出モジュールにて RSS アイテムを取得する. RSS アイテムには, Web ページの URL, Web ページのタイトルや更新日時が含まれている. 本システムでは, データベースに登録した RSS ファイルに更新がないかどうかを定期的に問い合わせ, 追加された RSS アイテムをデータベースに登録する.

3.4. エントリ生成

RSS アイテムの取得が終わると, 次にエントリ生成モジュールにおいて, WIX ファイルの更新に用いるエントリの生成を行う. このモジュールでは, RSS アイテムに含まれる Web ページのタイトルに, キーワード候補の単語が含まれているかどうかを調べる. タイトルにキーワード候補の単語が含まれていれば, 含まれていた単語をキーワード, RSS アイテム中の Web ページの URL をターゲットとしたエントリを生成し, WIX ファイルに追加し, WIX ファイルを更新する. RSS アイテムが削除されていれば, WIX ファイル中の対応するエントリを削除し, 古い Web ページにはアクセスできないようにする.

4.評価

評価に当たって, Yahoo!JAPAN ニュース (http://headlines.yahoo.co.jp/rss/list) で提供されている RSS ファイル 443 個を登録したところ, 2015 年 12 月 28 日 20 時に, 9,879 個の RSS アイテムを取得できた.

エントリ生成時において, キーワード候補の単語が含まれているかどうかの探索範囲を, Web ページのタイトルのみとした時と, Web ページの本文全体とした時の, エントリの生成数を比較した結果を表 1 に示す.

表 1 探索範囲とエントリ数の関係

探索範囲	エントリ数	キーワード数 *1	キーワードあたりのエントリ数
タイトル	45,266	10,888	4.16
本文全体	3,276,577	59,903	54.70

*1 キーワード数は重複を省いた数である.

探索範囲がタイトルのみの場合, キーワードごとに結合されるハイパーリンクの数が平均 4 つ程度で, ユーザが興味のある Web ページのアクセスの手助けとなりうると考えられる. しかし, 探索範囲が本文全体の場合, キーワードごとに結合されるハイパーリンクの数は平均 50 以上となってしまうため, 多数のハイパーリンクの中からユーザが興味のあるページにアクセスすることが困難になっていると考えられる.

5.まとめ

本論文では, RSS を利用して, 日々追加されていく Web ページへのアクセスを可能にする WIX システム用ライブラリの生成, 自動更新を可能にするシステムを実装し, 提案した.

文献

[1] 林昌弘, 青山峻, 朱成敏, 遠山元道.KeioWIX システム (1) ユーザインターフェース. データ工学ワークショップ, DEIM2011. 2011

[2] 森 良介, 藪 達也, 朱 成敏, 遠山 元道. Keio WIX システム (2) サーバーサイド実装. データ工学ワークショップ, DEIM2011. 2011.