DEIM2020 P2-9(day2 p82)

Web IndeXにおける学習応用のためのパターンマッチング機構の導入

新里　匠† 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
E-mail: †shinzato@db.ics.keio.ac.jp, ††toyama@ics.keio.ac.jp

あらまし　Web Index (WIX) とは，Web ページの閲覧者が閲覧中のページに存在するキーワードに対し，それに関連する他の Web ページへのアクセスを容易にするため，キーワードをハイパーリンクへと変換するシステムである．本システムでは，キーワードとそれに関連する URL を XML 形式で記述した WIX ファイルを用いることにより，WIX ファイル作成者の意図したキーワード集合に基づいたハイパーリンクへの変換を実現する．すなわち WIX は任意のキーワード集合を対象とできるため，例えば Web 上に存在する英語のコンテンツに対して辞書機能として WIX を使用するなど，教育・学習目的での利用も可能である．従来の WIX では，サーバーサイドでキーワード集合と入力文章との文字列マッチングを行っており，固定文字列のみが変換対象となるが，英語学習での利用を想定した場合，文法や熟語等の影響で固定文字列のマッチングでは対応できないケースが考えられる．そこで任意の Web コンテンツを利用した英語学習を目的とし，WIX にパターンマッチャーを導入する．

キーワード　Web 情報システム, Web IndeX, e-learning

1　はじめに

近年，インターネットの普及により，Web 上での情報検索のニーズが著しく増加し，ユーザーは検索エンジンを使用することで必要な情報を得ることが可能となった．それは教育や学習の現場においても言えることであり，学習時に疑問点が生じた場合などに Web 上のコンテンツを検索，閲覧することにより解決を図ることが可能であり，ある程度体系的にまとまっているコンテンツであれば副教材としての利用も可能である．Web コンテンツベースでの学習の利点として，ハイパーリンクにより関連するコンテンツへの参照が容易であるという点が挙げられる．すなわち，リンク化によって学習コンテンツとしての利便性が生かされることになる．

著者らは，リンク化により Web の利便性を向上すべく，Web IndeX（以下 WIX と呼ぶ）システムという情報資源形式の提案，開発を行っている [1] [2]．このシステムは，任意の Web ページに対して利用可能なシステムである．Web ページを閲覧している際，閲覧中のページ内に存在する単語に対してさらなる詳細情報を得たい場合，新たに検索エンジンで再検索をし，選択をしなければならない．Web ページの作成者は，ハイパーリンクを用いて Web ページに関連する複数の情報をあらかじめ結合することが可能であるが，これは Web ページ作成者が意図した関連のみで結合されるため，必ずしもその Web ページを閲覧するユーザーのニーズを満たしているとは限らない．WIX ではこのニーズを満たすべく，ユーザ主導でページ内のキーワードをハイパーリンクへと変換し，Web における情報資源結合を実現する．

WIX システムは任意のキーワード集合，任意の Web コンテンツの文章を対象とできるため，教育・学習システムとして応用することも可能である．その際，従来の WIX で行っていた固定文字列に対するマッチングでは対応できないケースが生じる．本論文ではそのようなケースに対応するため，パターンマッチング機構を導入した WIX システムを提案する．

本論文の構成は以下の通りである．まず 2 章で従来の Web IndeX システムの概要について述べる．3 章では本論文の関連研究を述べる．4 章では学習応用のためのパターンマッチングの導入の提案について述べる．5 章で評価，6 章で結論を述べる．

2　Web IndeX システム

2.1　システムの概要

Web IndeX システムとは，キーワードと URL の組み合わせであるエントリの集合を，XML 形式で記述した WIX ファイルを用いて，Web ページ内の文章に出現するキーワードを対応する URL へのハイパーリンクに変換する，Web における情報資源結合を実現するシステムである．この結合操作のことをアタッチと呼ぶ．アタッチによりハイパーリンクへの変換操作を行う前後の Web ページを図 1，図 2 に示す．

2.2　アーキテクチャ

図 3 に従来の WIX のアーキテクチャを示す．

WIX は，Google Chrome の拡張機能として実装されたクライアントサイドと，そこから受け取った文章に対して実際にキーワードをハイパーリンクへと変換するサーバーサイドから構成される．ユーザが拡張機能導入後にブラウザ下部に現れるボタンをクリックすると，閲覧中の Web ページの文章がサーバーサイドへと送信される．その文章とあらかじめ構築されたオートマトンを用いて文字列マッチングを行い，見つかったキーワードに対してタグ付けを行った後にクライアントサイドへとその結果が返される．

伝統を守り
進化を続ける慶應義塾

慶應義塾は、1858（安政5）年に福澤諭吉によって創立された日本最古の近代的高等教育機関であり、民間有志の協力によって経営される義塾の伝統を守りながら発展してきました。

21世紀の基調であるグローバル化の中で生き残るためには世界標準に適合すると同時に個性をもつことが必要です。慶應義塾の個性とは、世の中の流行に惑わされず、主体的に世の行く末を考えることのできる独立自尊の人材を社会のさまざまな分野に送り出し、「民」の力による日本の近代化に貢献してきたことです。これからもその原理は変わりませんが、時代に適合した教育、研究の強化が求められています。学生と正面から向き合う教育支援、学問分野の個性を尊重しつつ連携・融合を強化する研究支援、現場の声に耳を傾け、人の和と働きがいを大切にする法人経営。慶應義塾は学問の府としての原点を忘れず、教育・研究・医療を強化し、それによって社会に貢献していかなければなりません。

図 1　アタッチ前の Web ページ

伝統を守り
進化を続ける慶應義塾

慶應義塾は、1858（安政5）年に福澤諭吉によって創立された日本最古の近代的高等教育機関であり、民間有志の協力によって経営される義塾の伝統を守りながら発展してきました。

21世紀の基調であるグローバル化の中で生き残るためには世界標準に適合すると同時に個性をもつことが必要です。慶應義塾の個性とは、世の中の流行に惑わされず、主体的に世の行く末を考えることのできる独立自尊の人材を社会のさまざまな分野に送り出し、「民」の力による日本の近代化に貢献してきたことです。これからもその原理は変わりませんが、時代に適合した教育、研究の強化が求められています。学生と正面から向き合う教育支援、学問分野の個性を尊重しつつ連携・融合を強化する研究支援、現場の声に耳を傾け、人の和と働きがいを大切にする法人経営。慶應義塾は学問の府としての原点を忘れず、教育・研究・医療を強化し、それによって社会に貢献していかなければなりません。

図 2　アタッチ後の Web ページ

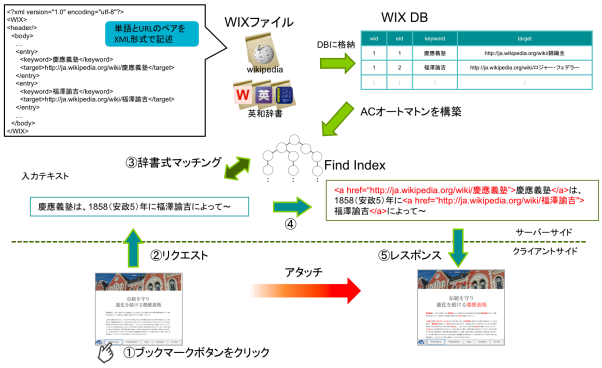


図 3　WIX のアーキテクチャ

2.3　WIX ファイルと WIX DB

上記に述べたオートマトンは，WIX ファイルというファイルをもとに構築される．WIX ファイルはキーワードと URL の組み合わせを XML 形式で記述したファイルである．見出し語となるキーワードは keyword 要素として，それに対応する詳細情報を示す Web ページの URL を target 要素として記述し，この 2 つを合わせてエントリと呼ぶ．また，header 要素にそのファイルの概要やメタデータ，作成者のコメント等を記述することも可能である．WIX ファイルは，「日本語版 Wikipedia の見出し語一覧」「プロ野球公式サイト」「Ameba ブログ」などのように，内容をグルーピングしたファイルが複数存在する．WIX ファイルの例を図 4 に示す．なお，本論文では主に英語学習の場での WIX 利用を想定し，keyword として英単語や英熟語，target として keyword に示された語句の詳細ページの URL が記述された WIX ファイルに基づくものとする．

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE WIX SYSTEM "http://wixdemo.db.ics.keio.ac.jp/wixfile.dtd">
<WIX>
  <header>
    <comment>コメント</comment>
    <description>説明</description>
    <language>ja</language>
  </header>
  <body>
    <entry>
      <keyword>大谷翔平</keyword>
      <target>http://npb.jp/bis/players/01305137.html</target>
    </entry>
    <entry>
      <keyword>斎藤佑樹</keyword>
      <target>http://npb.jp/bis/players/01905133.html</target>
    </entry>
    <entry>
      <keyword>石井裕也</keyword>
      <target>http://npb.jp/bis/players/71175110.html</target>
    </entry>
  </body>
</WIX>
```

図 4　WIX ファイルの例

これらのファイルは，WIX ファイルの管理システムにアップロードされることにより，WIX DB という WIX ファイルをエントリ単位に分割し，それぞれを RDB のタプルとして管理するデータベースに格納される．WIX DB 内のエントリ情報はメモリ上に展開され，高速なアタッチを実現するためのオートマトンが構築される．従来の WIX システムでは，Aho-Corasick 法に基づくオートマトンを構築し，辞書式マッチングを行う [3]．本論文においては，この文字列マッチングオートマトンを使用する前後の部分に変更を施し，新たにパターンマッチング機構を導入することを提案する．

3　関 連 研 究

Web Index の Web 上の単語をハイパーリンクに変換するという観点から，関連研究として山田育矢ら [4] による Linkify というアプリケーションが挙げられる．ニュースや文献を読んでいて知らない単語に出くわしたとき，テキストを選択，コピー＆ペーストして検索することの煩わしさを解消するために開発されたアプリケーションである．Linkify では，ユーザが検索したくなるようなキーワードを機械学習によって自動的に認識しリンクに変換することで，キーワードをタップするだけで Google や Wikipedia，Twitter などのさまざまな Web サイト上でキーワードを検索できる．

また，テキストを入力としてその中の単語を Wikipedia の記事と関連づける技術として Wikification が存在する．Wikipedia(http://ja.wikipedia.org) では，ページ内の単語に対し，その単語に対応するページが存在する場合に，そのページへのハイパーリンクが生成されており，ユーザーにとっての検

索の負担が軽減される．Rada Mihalcea ら [5] や David Milne ら [6] は Wikipedia へのリンクを自動生成する Wikify!のような Wikification システムを提案している．

4　学習応用のためのパターンマッチング

ここでは，従来の WIX のアタッチでは対応できないケースを示し，パターンマッチング機構を導入した提案システムのアーキテクチャを示す．

4.1　Web IndeX の学習応用

WIX では WIX ファイルに基づいたアタッチを行うため，任意のキーワード集合を基にしたハイパーリンク変換が可能である．したがって，英単語や英熟語を記述した WIX ファイルを用意すれば，Web 上の英文コンテンツに対してアタッチを行うことで，辞書機能あるいは e-learning システムとして WIX を利用することが可能である．

英語学習応用の際，課題の一つとしてあげられるのは英熟語への対応である．従来の WIX システムでは，サーバーサイドでキーワード集合と入力文章との文字列マッチングを行うが，ここで対象とされるものは固定文字列のみである．アタッチ対象を英熟語にも拡張することを考えたとき，文法や語形変化の影響で固定文字列に対するマッチングのみでは対応できないケースが考えられる．

英熟語では，動詞の語形や動名詞など形そのものに意味のあるものも存在する．例えば，be used to do と be used to ～ing では大きく意味が異なる．前者は受動態と不定詞の副詞的用法を組み合わせた「～するために使われている」のような意味となるが，後者は慣用表現として「～することに慣れている」という意味を持つ．

また，be 動詞のように単語単体で登場した際には簡単な単語であっても，熟語内で使われている場合にはこれの有無で大きく意味が変わることもあり，アタッチ時に不用意に無視することはできない．例えば used to は上記の通り be 動詞を伴って慣用表現として使われる他，be 動詞を伴わずに「よく～したものだ」のような意味で助動詞的に用いられることもある．

4.2　英熟語のパターン

本研究では，英熟語を最終的に以下のようなパターンにカテゴライズし，これらに対応することを目的とする．

4.2.1　語形変化のない熟語

もっとも単純なパターンであり，WIX ファイル内に記述されたキーワードと完全一致するか否かでアタッチ可能であるかを判断できるものである．したがって，従来の Web IndeX システムでアタッチ可能な熟語である．

4.2.2　語形変化する熟語

熟語中に含まれる語句が文章中で語形変化を起こし，WIX ファイル内に記述されたキーワードと完全には一致しない熟語を指す．代表的なものとして動詞を含む熟語があげられ，文章中で時制や主語の単複によって形が変わる他，動名詞や分詞など準動詞形に変化することも考えられる．また，代名詞を含む熟語（例 make oneself understood）も文章中で語形変化を起こすため，対応が必要である．

4.2.3　副詞の挿入のある熟語

熟語中で副詞 (句) の挿入の可能性がある熟語を指す．例えば，WIX ファイル内で pay attention to と定義された熟語に対して，文章中で pay more attention to という形で使用される場合，適切にアタッチを行うことができない．つまり，定義された熟語内に挿入語句がある場合にも対応する必要がある．

4.2.4　任意の目的語を含む熟語

動詞や前置詞の後ろに任意の目的語をとれる熟語を指す．例えば prefer A to B のようなものが挙げられ，A や B の位置には任意の名詞を目的語としてとることが可能である．

4.3　提案システムのアーキテクチャ

本論文で提案するシステムの全体像を図 5 に示す．

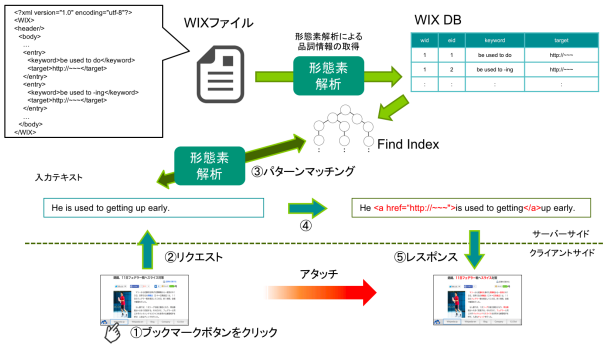


図 5　アーキテクチャ

従来の WIX システムでは，Aho-Corasick 法を用いたオートマトンを利用して固定文字列に対するマッチングを行っていたが，ここで提案するシステムにおいては，このマッチング部分に変更を施し形態素解析のステップを挟むことで，上記のような熟語に対するマッチングを行う．

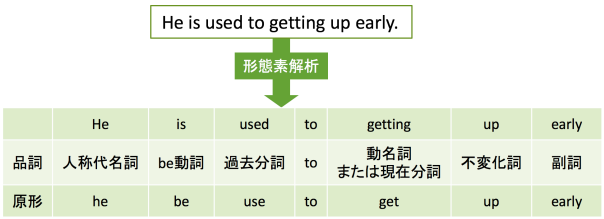
4.4　熟語に対するマッチング

熟語とのマッチングを行う際のアーキテクチャ案を示す．

WIX ファイルで記述された熟語に対し，WIX DB 格納前に形態素解析を行い，品詞や語形の変化に関する情報を取得する．WIX DB には各単語列を全て原形に直したものも一緒に格納しておき，これに基づいて高速文字列マッチングのためのオートマトンを構築する．入力文章に対しても同様に形態素解析を行い，品詞や語形の変化に関する情報を取得する．入力文章に対してこれらの情報を取得した結果の例を示したものが図 6 である．入力文章は一度全ての単語を原形に変換し，これに対して先に述べたオートマトンに基づいたマッチングを行った後，該当箇所の入力文章中での品詞と原形の情報が，WIX ファイルで定義された熟語においての品詞・原形の情報と一致した箇所をマッチングできた箇所とみなす．

5　評　　価

本論文で提案した手法では，4.2 節に示したパターンのうち，



	He	is	used	to	getting	up	early
品詞	人称代名詞	be動詞	過去分詞	to	動名詞 または現在分詞	不変化詞	副詞
原形	he	be	use	to	get	up	early

図 6　形態素解析による品詞情報の取得

語形変化のない熟語および語形変化する熟語に対してアタッチが可能となる. 副詞の挿入や任意の目的語をとれる熟語など，熟語中に任意の語句が挿入されるものに対しては，探索時のアルゴリズムへの改良が必要となる.

6 ま と め

6.1 結　　論

英語学習時に Web コンテンツを利用する際，任意のキーワード集合，任意のコンテンツを対象としたアタッチを可能にする Web IndeX システムを用いたリンク化によって，利便性の向上が期待できる．本論文では，従来のシステムで対応できない非固定文字列をアンカーテキストの対象とすることを目的とする英熟語対応のマッチング機構の導入を提案した.

6.2 今後の課題

英熟語表現を含んだサンプル文章およびそれらの熟語を記述した WIX ファイルを用意し，WIX システムによるアタッチを行なった結果とサンプル文章とを比較し，適切なアタッチが行われているかどうかを評価する. その上で，4.2 節に示したパターン全てに対応したアーキテクチャの改善が課題としてあげられる．また，設計変更による従来のシステムとの実行時間の差を小さくすることも対応すべき課題である.

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.

[3] 石崎文規，遠山元道. 大規模 Aho-Corasick オートマトンにおける追加更新手法の提案. データ工学ワークショップ，DEIM2012. 2012.

[4] Ikuya Yamada, Tomotaka Ito, Shinnosuke Usami, Shinsuke Takagi, Tomoya Toyoda, Hideaki Takeda, and Yoshiyasu Takefuji. Linkify: enhanced reading experience by augmenting text using linked open data. ISWC 2014 Semantic Web Challenge, 2014.

[5] Rada Mihalcea and Andras Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM ’07, pp. 233-242, New York, NY, USA, 2007.

[6] David Milne and Ian H. Witten. Learning to Link with Wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM ’ 08, pp. 509-518, New York, NY, USA, 2008.