

Web Index における WIX ファイル作成支援のためのリンク解析ツール

青木 陽介[†] 遠山 元道[†]

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

E-mail: [†]aoki@db.ics.keio.ac.jp, ^{††}toyama@ics.keio.ac.jp

あらまし 著者らは、キーワードと URL の組み合わせであるエントリの集合が記述された WIX ファイルを用い、Web ページ内の文章に出現するキーワードに対して、それに対応するハイパーリンクを生成する Web Index(WIX) システムを開発している。利用価値の高いエントリで構成された WIX ファイルほど有用性が高い。本論文では、WEB ページ群内のアンカーテキストと URL の組合せの抽出し、解析することでアンカーテキストに対するリンク先の一意度を割合で提示するとともに分布図で可視化するツールを実装した。また、その一意度が閾値を越えたエントリから構成される WIX ファイルを生成することもできる。

キーワード Web Index, Web 情報システム, Web, データマイニング, 情報抽出

1. はじめに

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) と呼ぶ情報資源表現形式の提案、開発を行っている。

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 形式で記述したものを WIX ファイルという。WIX ファイルを用い、閲覧中の Web ページに結合することで、Web ページ内の文章に出現するキーワードをそれに対応する URL のハイパーリンクに変換する。現在の WEB 形態では、アンカーテキストとリンク先の指定は WEB ページ作成時に作成者が設定するため、作成時に存在する WEB ページへのリンクだけにしか飛び先を指定することができない。しかし、この WIX システムでは、WIX ファイル内に書かれたエントリ情報に基づいて、WEB 文書内の単語にハイパーリンクを貼り付けるため、古いページから新しいページへの遷移を可能にするという利点がある。そのため、WIX ファイル内のエントリ情報は WIX システムを利用するにあたり、非常に重要なものとなる。

本論文では、WEB ページ群内のアンカーテキストとリンク先の集合を解析することで、利用価値の高いエントリだけを抽出し、それらのエントリで構成された WIX ファイルを生成するツールを提案し、実装した。

本論文の構成は以下の通りである。まず、2 章で WIX システムの概要を説明する。3 章で提案システムについて説明する。4 章、5 章で評価・まとめを行う。

2. Web Index システム

2.1 WIX ファイル

WIX ファイルは XML 形式で記述されたキーワードと URL の組み合わせであるエントリの集合であり、記述例は図 1 のようになる。エントリ内には、キーワードとなる見出し語を keyword 要素に格納し、それに対応する詳細情報を示す参照先の URL を target 要素に格納する。また、WIX ファイルには、header 要素にファイル概要、作者コメントなど、その WIX ファイル全体に

についてのメタデータを格納することも可能である。

1 つの entry 要素内に複数の keyword 要素と 1 つの target 要素という組み合わせや 1 つの keyword 要素と複数の target 要素という組み合わせも可能である。ただし、データベースにこれらの entry 情報を格納する場合には、キーワードとターゲットは 1 対 1 となるように複数のエントリに分けて扱われる [1]。

```
<?xml version="1.0" encoding="utf-8"?>
<WIX>
<header/>
<body>
  <entry>
    <keyword>本田圭佑</keyword>
    <target>http://samuraiblu.jp/member/honda_keisuke.html</target>
  </entry>
  <entry>
    <keyword>香川真司</keyword>
    <target>http://samuraiblu.jp/member/kagawa_shinji.html</target>
  </entry>
</body>
</WIX>
```

図 1 WIX ファイルの記述例

2.2 ハイパーリンクの生成 (アタッチ)

WIX システムでは関係データベースの結合演算の考え方を Web の世界に持ち込むことによって、ユーザ主導の Web 情報源の結合を図っている。アンカーテキストとリンクを Web ドキュメントから独立した WIX ファイルという情報源として保存し、それを適宜ドキュメントに対して、結合を行うことによってハイパーリンクを生成を行う。この結合操作を“アタッチ”と呼ぶことにする。WIX システムのクライアントサイドは、Firefox add-on や Chrome Extension によって実装されている [2]。図 2 は Chrome Extension の一例である。図 3 の記事を閲覧中に、アタッチボタンをクリックすることで、出現しているキーワードに対応する WIX ファイル内の target タグに記述されている URL へのハイパーリンクが図 4 の記事には生成される。

図 2 Chrome Extension(WIX ツールバー)



図 3 WEB ページ例 1



図 4 WEB ページ例 2(アタッチ)

2.3 Closed WIX

現在, WIX システムではユーザが好きな WIX ファイルを登録して, アタッチするという仕様になっており, 私的目的が強い. しかし, WIX は企業や学校など狭い領域での利用も考えられる. このような狭い範囲での WIX の利用を Closed WIX と呼ぶ. Closed な領域では, 特定のキーワードに対して, その WEB ページ特有のリンクが貼られていることが多い. また, WEB ページ閲覧者から見ても, キーワードに対して, 特有なリンク先を必要とするケースが多いと想定される. そのため, アタッチする際に「wikipedia.wix」や「dictionary.wix」のような汎用性の高い WIX ファイルよりも, その Closed な WEB ページ特有の WIX ファイルを利用する方が有用性が高いと言える.

2.4 オーサ主導アタッチ

先述した通り, WIX システムでは, ユーザーが好きな WIX ファイルを登録してアタッチするというユーザ主導のアタッチが主流である. 別のアタッチの方式として, ユーザ主導ではなく, web ページの作成者主導で結合処理を行う研究も進められている. これはユーザのリクエストに対して web ページをレス

ポンスする際に WIX ファイルとの結合処理を自動的に行って, リンクが生成されたあとの web ページ, つまりアタッチ済みの web ページを返すという手法がある. このようなアタッチ手法を”オーサ主導アタッチ”と呼ぶ. オーサ主導アタッチは, Web Author が Web ページにアンカータグを記述しなくても, 自動的に常に最新のリンクの参照関係を持たせることができるという点で優れたアタッチ手法である. 特に Closed WIX において, WEB ページごとに特有な WIX ファイルをアタッチさせることができるので, ユーザ主導アタッチよりも効果的であると言える.

3. WIX Analyzer ツール

3.1 本研究の背景

2.4 節でも述べたオーサ主導アタッチでは, html 毎だけでなく, ディレクトリ毎にアタッチする WIX ファイルを指定できる. オーサ主導アタッチで使用する WIX ファイルはどのようなものが適しているか. 例として, 慶應義塾大学理工学部の WEB ページを扱う. 理工学部のトップページの下に「情報工学科」「機械工学科」「応用化学科」などの WEB ページがあり, 「情報工学科」からのリンクからは「遠山研」「斎藤 (博昭) 研」「斎藤 (英雄) 研」などがある (図 5). 情報工学科のページ群では, 遠山研というキーワードに対して, リンク先は遠山研のホームページ 1 つに一意的に決まるが, 斎藤研というキーワードでは, 斎藤 (博昭) と斎藤 (英雄) の 2 つが考えられるため, 一意的に決まらない. しかし, 斎藤 (博昭) 研のページでは, 斎藤研というキーワードのリンク先は, 斎藤 (博昭) 研のトップページへのリンクがほとんどであり, 一意的であると考えられる (図 6). このようにディレクトリによって, 同じキーワードでもリンク先の一意度は変化する. そこで, オーサ主導アタッチする際, 一意度が高いリンクをターゲットとして持ったエントリで構成された WIX ファイルを使えば, そのディレクトリでは不適切なリンクへ飛ぶ事がなくなると想定できる. この場合だと, 斎藤 (博昭) 研のホームページ内で斎藤 (英雄) 研へのハイパーリンクをアタッチするのは不適切だと考えられる. このようなアンカーテキストとリンク先に対する一意性が高いエントリの集合を WIX ファイル化すれば, オーサ主導アタッチにて用いる WIX ファイルの作成支援に非常に効果的であると考えられる. また, オーサ主導アタッチでは, WEB ページごとにアタッチする WIX ファイルが異なることが多いが想定できる. 手動での記述の場合, ユーザーの意図が最も反映された WIX ファイルが作成できるが, 手間と時間がかかる [3]. それぞれの WIX ファイルを手動で作成することは WIX ファイル作成者にとって負担になる. そこで, 本研究では, WEB ページ群でのアンカーテキストとリンク先の関数従属性 (一意度) を解析し, 可視化すると同時に, 一意度の閾値を越えたエントリ集合で構成された WIX ファイルを生成するツールを提案し, 実装した.

3.2 ツール概要

WIX システムにおいて, WEB ページ群内のアンカーテキストとリンク先の解析を行うことで WIX ファイルの作成を支援するという観点から, 本ツールを WIX Analyzer ツールと呼ぶ.



図5 ディレクトリとリンクの一意度の関係

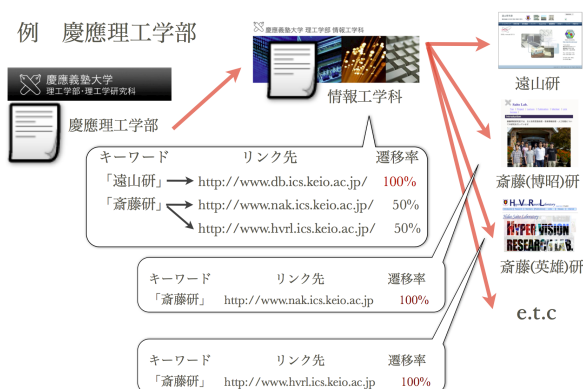


図6 ディレクトリとリンクの一意度の関係 (cont.)

本ツールは Web アプリケーションとなっており、ツール専用ページを持つ。まず、本ツール使用者は解析したい Web サイトのルートとなる URL を指定する。指定された URL をクロールし、アンカーテキストとリンク先を抽出する。抽出された URL に対して同様の操作を行い、最初に指定した URL 配下のディレクトリや html ファイルを収集する。このとき、Closed WIX において、外部リンクは有用性を持たないことが多いと考えられる。そのため、最初に指定した URL とルートドメインが一致しない URL については排除を行う。

次に、解析として、収集された html ファイル毎にアンカーテキストとリンク先のペアの出現数をカウントし、指定した URL 配下での総出現数を算出する。アンカーテキストとなるキーワードの中で、同じキーワードでもリンク先が異なる場合は、キーワードごとでのリンク分岐を表記し、分岐率（一意度）を表示する。

ここでの一意度とは、キーワードに対するリンク先の関数従属性を表す。一意度の具体的な説明は、3.4 小節で行う。分岐率とキーワードの出現回数の 2 つ座標軸について分布図を表示して可視化する。また、指定した分岐率の閾値を越えたエントリ情報だけで構成された WIX ファイルを生成することを可能にした。

3.3 ディレクトリ指定

本ツールでは、指定した URL 配下のディレクトリと html ファイルを収集し、そのディレクトリ構造をマウスオーバーすることでプルダウンで表示できる。特定のディレクトリ

をクリックしたとき、そのディレクトリ配下での解析結果を動的に表示することができる。また、解析結果として、参照 URL 数と解析実行時間も表示する。図 7 は慶應義塾大学理工学部のホームページ (<http://www.st.keio.ac.jp/>) を解析し、ディレクトリ構造を表示した一例である。図 7 では「faculty」の部分をクリックしたことで、対象ディレクトリ範囲が「<http://www.st.keio.ac.jp/departments/faculty>」へと変更され、この URL 配下にある html ファイル群での解析結果を動的に出力することができる。



図7 ディレクトリ指定

3.4 アンカーテキストとリンク先の一意度の表示

解析結果の表示方法として、「全リンク表示」「分岐リンク表示」「高頻度リンク表示」の 3 種類の表示方法を用いる。どの表示方法でもアンカーテキストのキーワードごとに表示するという点は共通である。「全リンク表示」では、解析結果のアンカーテキストとリンク先のペアの情報をすべて表示する。「分岐リンク表示」では、同じキーワードに対してリンク先の 2 種類以上に分岐するリンクのペアのみを表示する。「高頻度リンク表示」は、一意度が 80% を越えたリンクのペアのみを表示する。図 8 は「分岐リンク表示」の例である。図 8 において、「計測工学科」というキーワードに対して、95.8% が「<http://www.st.keio.ac.jp/ygprize/measurement.html>」へのリンクとなっていることがわかり、分岐率（一意度）は 95.8% ということになる。表示における工夫として、同一キーワードでの URL 表示は分岐率が高い順に表示し、一意度が 80% を越えたエントリは赤字で、50% を越えたエントリは青字で表記する仕様となっている。

3.5 分布図の可視化

3.2 小節でも述べたようにアンカーテキストとリンク先のエントリ情報について、横軸を一意度 [%]、縦軸をキーワードの出現回数 [回] として、分布図を表示する。例として図 9 に前述の慶應義塾大学理工学部の WEB ページでの分布図を示す。分布図のプロット位置によって、WIX ファイルを生成することに価値があるかが一見してわかる。プロット位置が右側に多いと

リンク抽出と出現頻度			
全リンク抽出	全リンクと出現頻度の集計	出現頻度リンク抽出	グラフ
keyword : 計測工学科			
target : http://www.st.keio.ac.jp/ygrize/measurement.html	count : 23 / 24	percentage : 95.8%	
target : http://www.st.keio.ac.jp/ob_relay/ob_atmeas.html	count : 1 / 24	percentage : 4.2%	
keyword : 電子工学科			
target : http://www.st.keio.ac.jp/departments/faculty/elec.html	count : 31 / 69	percentage : 44.9%	
target : http://www.st.keio.ac.jp/ygrize/elec.html	count : 23 / 69	percentage : 33.3%	
target : http://www.st.keio.ac.jp/campus/laboratory/02.html	count : 13 / 69	percentage : 18.8%	
target : http://www.st.keio.ac.jp/learning/learning_elec.html	count : 1 / 69	percentage : 1.4%	
target : http://www.st.keio.ac.jp/ob_relay/copy_of_ob_elec.html	count : 1 / 69	percentage : 1.4%	
keyword : 電気工学科			
target : http://www.st.keio.ac.jp/ygrize/electrical.html	count : 23 / 24	percentage : 95.8%	
target : http://www.st.keio.ac.jp/ob_relay/ob_electrical.html	count : 1 / 24	percentage : 4.2%	
keyword : 2年次からの所属学科を通知「学科分け説明会」が開催されました			
target : http://www.st.keio.ac.jp/news/2007/203_001.html	count : 5 / 8	percentage : 62.5%	
target : http://www.st.keio.ac.jp/news/2008/1218_001.html	count : 3 / 8	percentage : 37.5%	

図 8 アンカーテキストとリンク先の一意度の表示

一意度が高い有用なリンクであり、左側に多いとほとんど使われないリンクであり有用性が低いと理解できる。

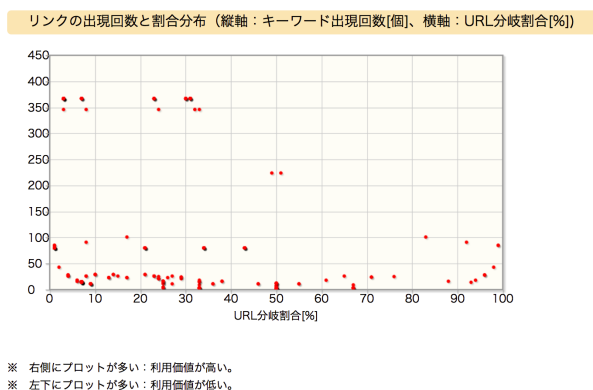


図 9 分布図

3.6 WIX ファイル生成

表示されているエントリ情報に対して、指定した一意度の閾値を越えたエントリだけを抽出し、WIX ファイルの形式で表示する。一意度の閾値はデフォルトで 80%とし、10%区切りで変更することが可能である。また、アンカーテキストが単語ではなく文章型のキーワードに対しては WIX ファイル化の意義がないと考えられるため、記号を含むキーワードを取り除くフィルタリングをすることが可能であり、一意度が 100%で分岐がないリンクも WIX ファイル化するかをユーザ指定で決定することができる。全リンク数に対するエントリ数も出力することでフィルタリング効果を一見できる仕様にした。

WIXファイル作成			
リンク分岐割合	完全一意リンク	キーワードフィルタ	該当エントリ数
80%	☐ 含む ☒ 含まない	☐ 行う ☒ 行わない	10 / 435
<pre><?xml version="1.0" encoding="utf-8"?> <wix> <header> <body> <entry> <keyyear>卒業生の方</keyyear> <target>http://www.st.keio.ac.jp/news/news_2003_cor.html</target> </entry> <entry> <keyyear>卒業生の方</keyyear> <target>http://www.st.keio.ac.jp/news/news_2003_cor.html</target> </entry> <entry> <keyyear>受験生の方</keyyear> <target>http://www.st.keio.ac.jp/news/news_2003_cor.html</target> </entry> <entry> <keyyear>在校生の方</keyyear> <target>http://www.st.keio.ac.jp/news/news_2003_cor.html</target> </entry> <entry> <keyyear>教職工学科</keyyear> <target>http://www.st.keio.ac.jp/ygrize/mathematics.html</target> </entry> <entry> <keyyear>計測工学科</keyyear> <target>http://www.st.keio.ac.jp/ygrize/measurement.html</target> </entry> </body> </header> </wix></pre>			
ファイル名	main.wix	WIX	

図 10 WIX ファイル生成

4. 評価

提案手法の有用性を評価するために、本ツール使用時と未使用時での WIX ファイル作成時間とタイピング数の比較、生成 WIX ファイルの精度測定、パフォーマンス評価の 3 項目について実験を行い、評価を行った。

4.1 WIX ファイル作成時間とタイピング数

図 11 は、筆者が本ツールを使用せずに、事前にエントリ数の異なる 10 種類の WIX ファイルを作成したときの作成時間とタイピング数を示したグラフである。このときのタイピング数は、クリック数とキーボードタイプ数の合計値である。グラフからエントリ数が増加するにつれて作成時間とタイピング数も線形的に増加していることがわかる。一方、図 12 は、本ツールを用いて、WIX ファイルを作成したときのグラフだが、エントリ数が増加しても作成時間とタイピング数は変化せず、タイピング数は約 20 回、作成時間は約 6 回とほぼ一定であった。この結果から、手書きの WIX ファイルの方が精度が高いとはいえず、大幅に WIX ファイル作成者の負担を軽減することができると考えられる。

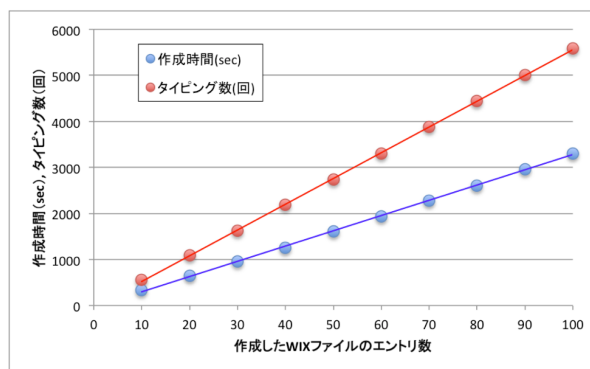


図 11 WIX Analyzer 未使用時

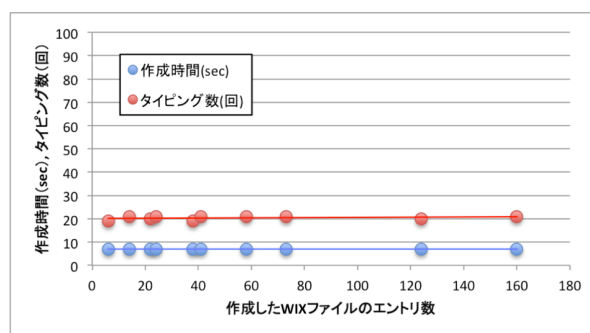


図 12 WIX Analyzer 使用時

4.2 生成 WIX ファイルの精度

生成した WIX ファイルのエントリが有用性があるかどうかを本ツールで作成した異なる 10 個の WIX ファイルにて、期待通りのエントリが抽出できているかをユーザ判断でカウントした。全エントリ数に対し、有用性があるエントリの数の割合をパーセンテージ化した時の結果が表 1 であった。完全一意とは、

アンカーテキストに対するリンク先の割合が 100%であるエントリである。完全一意のエントリは、文章的なものや汎用性の低いリンクがほとんどであったため、WIX ファイルに含んだ場合は不適切なエントリとして判断できるものが多かった。完全一意を含まないエントリは約 90%のエントリが適切と判断できた。しかし、これらの WIX ファイルのエントリ数は少なく、少ないものでは 10 エントリという WIX ファイルも存在したため、完全一意であり、且つ利用価値のあるリンクを抽出する必要があると考えられた。

表 1 WIX ファイルの精度					
割合	0 - 59%	60 - 69%	70 - 79%	80 - 89%	90 - 100%
完全一意含む	0	2	5	3	0
完全一意含まない	0	0	0	3	7

4.3 パフォーマンス測定

本システムの有用性を評価するために、本ツールを利用したときの抽出エントリ数と参照 URL 数に対する解析実行時間のパフォーマンス評価を行った。対象としたのは、特定の企業や小学校や大学等などの狭い領域での WEB ページ 13 件である(表 2)。これらの WEB ページを解析したときのルートディレクトリ配下にある全エントリ数と URL 数と解析実行時間の関係性を調べた。

図 13、図 14 にて図示されるように、エントリ数や参照 URL 数が増加するにつれて解析実行時間も線形的に増加していることがわかる。また、いくつか近似直線から離れているプロットも見受けられる。これは、エントリ数が極端に多いが参照 URL 数が少ない場合や、ディレクトリ構造の層が深い場合に、近似直線から逸脱したプロットが発生してしまうと考えられる。

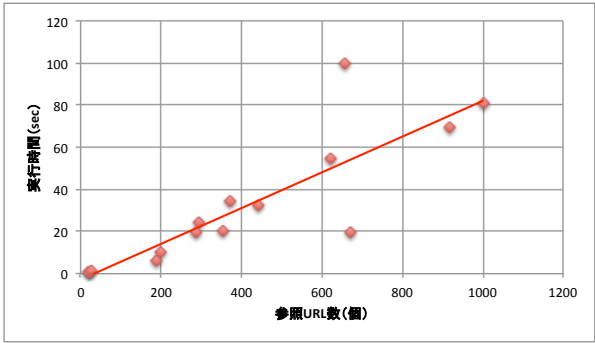


図 13 参照 URL 数と実行時間の関係

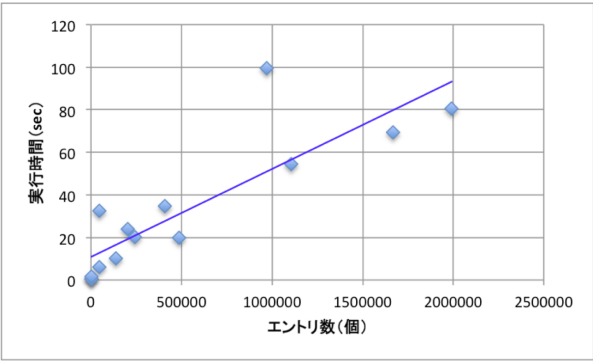


図 14 抽出エントリ数と実行時間の関係

表 2 解析に利用した WEB ページ

番号	WEB ページ	URL
1	遠山研究室	http://www.db.ics.keio.ac.jp/
2	慶應理工学部	http://www.st.keio.ac.jp/
3	早稲田大学	http://www.sci.waseda.ac.jp/
4	青山大学	http://www.aoyamagakuin.jp/
5	矢上小学校	http://www.edu.city.yokohama.jp/sch/es/yagami
6	横浜小学校	http://yokosyo.sakura.ne.jp/
7	ソニー	http://www.sony.co.jp/
8	東芝	http://www.toshiba.co.jp/
9	テレビ朝日	http://news.tv-asahi.co.jp/
10	駿台	http://www.sundai.ac.jp/
11	任天堂	http://www.nintendo.co.jp/
12	横浜ベイスターズ	http://www.baystars.co.jp/
13	千葉ロッテマリーンズ	http://www.marines.co.jp/

5. おわりに

5.1 結 論

本論文では、指定した URL 配下にある Web ページ群でのアンカーテキストとリンク先を解析し、その一意度を調べることによって、散布図で可視化するとともに、一意度の閾値を越えたエントリのみで構成された WIX ファイルを生成するツールを提案し、実装した。本ツールを利用することで、リンク解析による WIX ファイル化を行う上で、有用性の高い WEB ページかを可視化することができるので今後の WIX ファイル作成支援に役立つと言える。しかし、まだ解析時間や解析方法や一意度の測定精度、また、WIX ファイル化時のフィルタリング精度について改善の余地があると考えられる。そのため、更なる利便性の追求とともに精度の向上を目指していくことが今後の課題と言える。

参考文献

[1] 長谷川 駿, 遠山 元道. Web Index における関連コンテンツ推薦システム. データ工学ワークショップ, DEIM2013. 2013.
[2] 林 昌弘, 青山 峻, 朱 成敏, 遠山 元道. Keio WIX 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.

- [3] 市東 隼, 分部 亮太, 朱 成敏, 遠山元道. Keio WIX システム (3) コンテンツ作成. データ工学ワークショップ, DEIM2011. 2011.