

WebIndexにおけるキーワード適合型サービス推薦システム

井上明莉咲[†] 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: [†]arisa@db.ics.keio.ac.jp, ^{††}toyama@ics.keio.ac.jp

あらまし Web Index(WIX)とは、Web ページを閲覧する人が Web ページ中の単語に関連する他の Web ページにアクセスすることを容易とするため、単語をハイパーリンクに変換するシステムである。本システムでは、単語リストと検索エンジン、SNS、EC サイトなど(これらの総称をサービスとする)の検索結果ページへのハイパーリンクを生成する仕組みも構築されている。本研究では、ユーザの満足度の向上を目的として、単語の分類ごとに適切なサービスの推薦方法を提案する。

キーワード Web 情報システム, Web Index, 推薦

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増加し、ユーザが検索エンジン、EC サイト、SNS など(総称してサービスと呼ぶ)を使用することで必要な情報を得るようになった。また、サービスだけでなく、ハイパーリンクを利用することで、異なる Web ページからさらなる情報を得ることも可能である。しかし、現在の一般的な Web 形態では、Web ページ作成者による特定のページへのリンクしか持つことができない構造となっており、そのリンクが必ずしも Web ページ閲覧者の欲する情報であるとは限らない。

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(以下、WIX と呼ぶ)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]。WIX システムでは単語(keyword)と URL(target)の組み合わせであるペア(WIX エントリ)の集合を XML 形式で記述した WIX ファイルと呼ばれるものを用い、閲覧中の Web ページに結合(アタッチ)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 へのハイパーリンクに変換する。現在の Web 形態では、Web ページ作成者によって特定のアンカーテキストから特定のページへのリンクが関連付けられるという構造が一般的である。WIX では、アンカーテキストとリンクを Web ページから独立した「キーワードとリンク先の集合」として扱い、任意のドキュメントに対してユーザ主導で「結合」することでドキュメント内のキーワードを対応する URL のハイパーリンクに自動的に変換する。その結果、Web ページ作成の時系列という壁を越え、古い Web ページからより新しい Web ページへのリンクも可能となる。keyword と target の組み合わせを用いる WIX ファイルアタッチの場合、target を増やすたびにタプル数が単語数分増える為、データが膨大になってしまう。それを受けて、Wordtank アタッチ機能を追加した。この機能は、Wikipedia の見出し語群と google, twitter, Amazon, Yahoo, YouTube 5 サービスの URLprefix(検索用 URL のクエリパラメータを除いたもの)を結合し、動的にリンクを生成する。本研究では、Wordtank ア

タッチのユーザ満足度向上を目的とし、単語ごとに適切なサービスを推薦するシステムを提案・実装する。

本論文の構成は以下のとおりである。まず、2 章で WIX システムの概要及び Wordtank アタッチの現状を説明し、次に 3, 4 章で本研究の設計と実装について述べる。最後に 5 章で評価について述べ、6 章でまとめを行う。

2. Wordtank アタッチ機能

2.1 システムアーキテクチャ

WIX システムではアタッチに必要な情報を WIX ライブラリ・WIXDB・Find インデックスの 3 つの異なる形態で管理する。これら 3 形態について説明する。

2.1.1 WIX ライブラリ

WIX ライブラリでは、全ての Wordtank 生成用ファイルの XML テキストをそのまま保存しており、ファイル単位での情報管理を行っている。アタッチの際には各エントリに対して辞書式マッチングを行うため、ファイルをエントリ単位に分解し WIXDB に格納する。

2.1.2 WIXDB

WIXDB では、ライブラリで管理している Wordtank 生成用ファイルをエントリ単位に分解し、wordtank テーブル(表 1)で管理する。

表 1 wordtank テーブル

eid	keyword
1	日本ハム
2	エンゼルス
3	大谷翔平
⋮	⋮

2.1.3 Find インデックス

Find インデックスでは、WIXDB の wordtank テーブルからエントリ情報をメモリ上に展開する。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。

2.1.4 Wordtank アタッチ

WIX ファイルの集合とサービスの集合をブックマークとして記録しておき、WIX ツールバー (図 1) にボタンを用意する。



図 1 WIX ツールバー

ユーザが Web ページを閲覧中に、このボタンをクリックすると、サーバーサイドにおいて閲覧 Web ページと Find インデックスとの辞書式マッチングが行われ、リンクを生成する際に必要となる情報を取り出せる形に書き直された HTML 文書がレスポンスとして返され、対象文字列の色が変わる。この処理を Wordtank アタッチと呼ぶ。対象文字列をクリックすると、文字列とサービスの id 情報がサーバサイドに渡される。サーバサイドでサービスの URL と文字列を結合しリンク先ページを表示する。

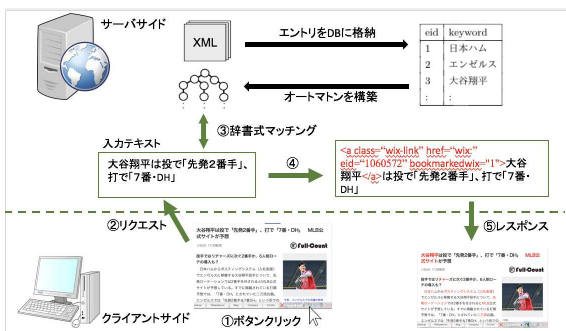


図 2 Wordtank アタッチの流れ

なお図 1 左側の Wikipediaja, Wikipediaen, Blog, Company, EJ Dict とラベルされているボタンは WIX ファイルを用いたハイパーリンク生成処理 (WIX ファイルアタッチと呼ぶ)、右側の Google, Twitter, Amazon, Yahoo!, YouTube のロゴがついたボタンは Wordtank アタッチを行うためのものである。

WIX システムのクライアントサイドは、Firefox add-on や Google Chrome Extension によって実装されている。図 3 は Google Chrome Extension の一例である。

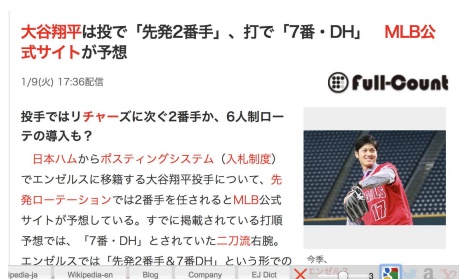


図 3 Wordtank アタッチ例

2.2 Wordtank アタッチ機能の現状

このように単語群と prefix を組み合わせ、動的にハイパーリンクを生成することで、エントリ数を増やさずリンク先ページのバリエーションを豊かにすることが可能になった。しかし現システムは、単一の単語リストを用いてアタッチが行われているため、不必要なリンクが生成されてしまう。例えば、図 3 の「大谷翔平」に Instagram のリンクをアタッチすると、同ページ内の Instagram で検索するほとんどない「日本ハム」のリンクも生成されてしまう。また、サービスを増やすことが難しい設計になっており、図 1 の様にツールバーにサービス毎のアタッチボタンを用意するため、対応可能なサービス個数がツールバーサイズに制限されたり、サービスが増えるに従ってユーザが目的のボタンを選択するコストがかさむことが懸念される。

3. 提案設計

2.2 で挙げた点を解決するにあたり、以下の様に仕様を変更する (図 4)。サービス毎にアタッチせず、ボタン一つでドキュメント内のテキスト全てに Find インデックス処理を行う。その後にユーザが調べたいと思う文字列にマウスオーバーすると適したサービスが設定個数分表示され、候補の中から求めるサービスを選択する。この方法であれば、単語毎に適切なサービスのみアタッチしているため、無駄なリンクは生成されず、かつユーザコストが一定に保たれるため、登録サービス増加に伴う問題点も解消することができる。



図 4 実装イメージ図

3.1 Type の導入

keyword ごとに異なるサービスの組み合わせを表示するには、wordtank テーブルに格納されている全ての単語一つ一つに推薦サービスを割り当てる必要がある。しかし、何十万とある keyword に割り当てを行うとなると、かなりの計算量と時間を要する。そこで次元を下げるために、似通った特性を持つ単語群を type と呼ぶものでグルーピングする。本研究ではサービス推薦方法に注力するため、グルーピングルールは DBpedia のオントロジ [5] を日本語版にマッピングしたもの [6] を借りた.type 付けした単語情報を 2.1 の WIX システムアーキテクチャで管理可能にするため図 5 の記述を持つファイルを用意し、表 2 のテーブルで管理する。

```

<keyword>ソクラテス</keyword>
<type>Philosopher</type>
</entry>
<entry>
<keyword>LimeWire</keyword>
<type>Software</type>
</entry>
<entry>
<keyword>うる星やつら</keyword>
<type>Anime</type>
</entry>
<entry>
<keyword>うる星やつら __高橋留美子__1</keyword>
<type>Manga</type>
</entry>
<entry>
<keyword>士郎正宗</keyword>
<type>ComicsCreator</type>
</entry>

```

図5 type付き wordtank ファイル

表2 typed_wordtank テーブル

eid	keyword	type
1	大谷翔平	BaseballPlayer
2	日本ハム	Company
:	:	:

3.2 システムの挙動

まずは、アタッチまでの流れ(図6)を述べる。ウェブブラウジング中のユーザがボタンをクリックするとドキュメント内のテキストとHTML リクエストがサーバ側に渡される。サーバ側ではテキストを入力として、の前処理としてタグ抽出処理を行なった後、Find インデックスにおいて辞書式マッチングを行う。この時、全エントリ情報のうち文書中に存在するキーワードを持つエントリを Web 文書中の出現位置とセットにして配列に格納する。なお Find インデックスは WIXDB のエントリ情報から Aho-Corasick 法による Find インデックスを構築している。マッチング結果は対象エントリの typed_wordtank テーブルの eid, keyword, type 名を type 情報を管理する type.info テーブルを用いて id に変換した tid と入力 HTML 文書の出現位置における開始位置、終了位置が返される。最後に Find インデックスの前処理で抽出したタグを用いて入力 HTML 文書を新たな HTML 文書に書き換える。この書き換える処理を typed_wordtank アタッチと呼ぶ。

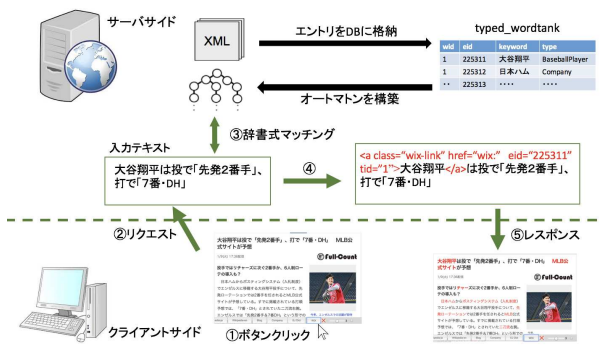


図6 システムの挙動 (アタッチ)

次に、アタッチが後からリダイレクト処理までの流れ(図7)を説明する。アタッチが行われて、赤文字になった文字列の中から興味のある単語にマウスオーバーするとタグに含まれている tid をサーバ側に渡す。サーバ側では、サービス推薦システムで

用いる情報を管理している type_service テーブルと type_info テーブルに対して tid で SELECT 処理を行う。type_service テーブルでは tid=1 をペアに持つ sid を抽出し、サービス情報を管理する service_info テーブルについて sid で検索をかけ、prefix (サービスの URL) やアイコン画像の URL などの必要なサービス情報を抽出する。type_info テーブルでの SELECT 処理では、推薦システムで用いる情報を選択する。これらの検索結果をクライアント側に送る。クライアント側では渡された情報を元に、表示するサービスの個数やどのサービスを表示するかの決定を行い、アイコンを表示する。ユーザが表示されたいずれかのアイコンをクリックすると、サーバ側にアイコンのサービスが持つ prefix とマウスオーバーされたキーワードが送られ、リダイレクト処理が行われ、期待する検索結果ページに飛ぶ。

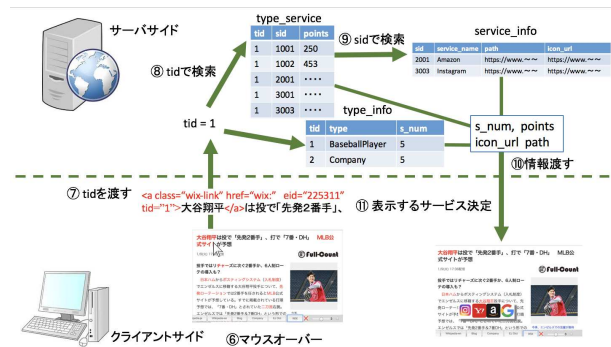


図7 システムの挙動 (アイコン表示)

3.3 サービス推薦

推薦システムは、大きく完全個人化、非個人化と分類できる[7]。どちらかのシステムなのかにより推薦方法が異なると考え、場合分けをする。なお、個人化の度合いに関わらず推薦システムは output-input-process model(O-I-P モデル)[8]に習い設計し、データの入力、嗜好の予測、そして推薦の提示の三つの段階で推薦を行う。

3.3.1 完全個人化の推薦システム

完全個人化は、利用者個人の過去の利用履歴に応じて異なる推薦をする場合である。そのユーザ専用のカスタマイズされるため、満足度が高い候補を提示すること可能と予測される。O-I-P モデルにおけるデータ入力段階での嗜好データ収集のアプローチには暗黙的と明示的の二種類があり、暗黙的な獲得とは、利用者の行動をから、利用者の嗜好や関心を推察することで嗜好データを得る方法で、明示的な獲得とは、利用者に好き嫌いや、関心のあるなしを質問し、利用者に回答してもらう方法である。本研究の場合、明示的なアプローチとして、各ユーザに初期設定で、type ごとに表示したいサービスを選択してもらう方法が考えられる。暗黙的なアプローチとしては、初期設定では全てのサービスを表示し、ユーザ毎にクリック数を取り、カウント数によって嗜好を予測する方法がある。

3.3.2 非個人化の推薦システム

非個人化は、全ての利用者について全く同じ推薦をする場合である。完全個人化の場合と同様、データ入力段階で暗黙的・明示的2つのアプローチをとる。

本研究の場合、明示的なアプローチとして表示するサービスの個数を決めるのに必要なデータや type ごとに頻繁に利用するサービス何かといったデータをアンケートで集めることが効果的だ。暗黙的なアプローチとしては、どの type にどのサービスをユーザが選んだかの履歴を取ることである。この履歴から type ごとの頻繁に利用されているサービス、最低限表示すべきサービスの個数を予測することが可能になる。

4. 実装

4.1 データ

4.1.1 type 付き wordtank

本研究では 3.1 で述べた通り、初期入力として DBpedia 日本語版を借りた。また、WIX システムでユーザは固有名詞に興味を持つ傾向があることを踏まえ同ジャンルの単語群を適宜 type 付けしテーブルに手動で追加した。例えば、レストラン名リストを restaurant type として扱うなどだ。

4.1.2 サービス

id を見ただけで EC サイト・検索エンジンなどの種類が特定できる様に、id の最高位を検索エンジンは 1, EC サイトは 2, SNS は 3, 専門性があるサイトは 4, 辞書は 5 とした。これら 5 種類のサービスを、検索エンジンや EC サイトが紹介されているページ [10] [11] を参考に、100 個収集した。表 8 は収集したサービスの一部である。

1001 yahoo(web)	2001 Amazon
1002 Google(web)	2002 価格.com
1003 Bing(web)	2003 メルカリ
1012 Yahoo(画像)	3001 Twitter
1013 Google(画像)	3003 Instagram
1014 Bing(images)	4003 Yelp
1015 Bing(Videos)	4039 テレビ王国
1018 Google(地図)	5002 goo 辞書

図 8 100 個のサービス (一部)

4.2 推薦システム実装

3.3 であげた推薦手法のうち、非個人化手法を実装した。表示個数と推薦サービスを決めるのに必要な嗜好データをアンケートで収集し、初期値を設定する。アンケートのみであるとサンプル数が少ないため、クリック数を用いて初期値のアジャストを行う。

4.2.1 初期値設定 (表示個数)

表示個数を決定するために実施したアンケートは、ユーザが不自由に感じるアイコン数を知る事を目的としたもので、1 20

個のアイコンの並び (30 通り用意) から特定のアイコンを見つける体験をしてもらい、1. 探すのに時間がかかった (反射的に見つけられなかった)、2. 探すのが面倒だと感じた、3. 画面が騒がしいと感じた、4. 簡単に見つけられたの 4 つの項目から当てはまるものを選択してもらおう (図 9)。結果、アイコン 6 個以上で誰かしらが選択肢 4 以外を選んでいることがわかった。この結果より、初期設定では全ての type で、推薦するサービスの個数を 5 に設定した。

Yahoo!のロゴを探してください。以下の選択肢から当てはまるものを選択してください(複数可)。

☐ 反射的に見つけることができなかった(探すのに時間がかかった)

☐ 探すのが面倒だと感じた

☐ PC画面に出てきた時にごちゃごちゃした・うるさい印象を受けた

☐ 探しやすかった

図 9 表示個数を決めるアンケート例

4.2.2 初期値設定 (推薦するサービス)

type ごとにどのようなサービスを推薦すべきか判断するために必要なデータを、収集することを目的にアンケートを作成した。本アンケートは、22 個の type に属する単語について 100 個のサービスの中から、その単語を検索する際に利用する頻度が高いサービスをランキング形式で最大 5 つ回答してもらおう (図 10)。アンケートに用いられた 22 個の type は、図 11 の通りである。結果を用いて O-I-P モデルの 2 段階目、嗜好の予測を行う。各 type でのサービスの利用度を points という値を用いて表現し、points はユーザ毎に各問で 1 位に選ばれれば 5 点、2 位であれば 4 点、3 位であれば 3 点、4 位であれば 2 点、5 位であれば 1 点を points に加算する。初期設定では、points が高い順に上位 5 つのサービスを推薦候補とした。

次の単語を/で検索するときに使うサービスは？王貞治、大谷翔平 (BaseballPlayer) *

yahoo検索(画像)

yahoo知恵袋

ヤフオク!

google検索(web)

google検索(画像)

yahoo検索(web)

yahooショッピング

Bing(images)

NAV

1位

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

2位

☐

☐

☐

☐

☐

☐

☐

☐

☐

3位

☐

☐

☐

☐

☐

☐

☐

☐

☐

4位

☐

☐

☐

☐

☐

☐

☐

☐

☐

5位

☐

☐

☐

☐

☐

☐

☐

☐

☐

図 10 BaseballPlayer type についてのアンケート

- Artist

• ComicsCreator

• Company

• Actor

• Disease

• Film
- Station

• Book

• HotSpring

• Model

• Politician

• FigureSkater
- University

• Manga

• Artwork

• ShoppingMall

• BaseballPlayer

• Restaurant
- Comedian

• VideoGame

• City

• TelevisionShow

図 11 アンケートに用いた 22 個の type

4.2.3 アジャストメント

アンケートのみだと、サンプルが少ないため、クリックカウントという形で全ユーザの嗜好データを収集し、初期値を調整する。推薦する5個のサービスの他に5つサービスを表示し、1クリック1カウントとして click_count という値に加算する。type T におけるクリックカウント総数がサービスの個数と同じになると、表示するサービスの個数を調整し、click_count の値を points に加算する。また、クリックカウント総数がサービスの個数になる前に、新しい type あるいはサービスが追加された場合は、全ての click_count の値を初期化する。

5. 評価

5.1 適合度

初期値設定で選ばれたサービス群と、Wordtank アタッチに対応していたサービス群がどれほど同じサービスを含んでいるか考察する。既存の Wordtank アタッチのサービス (Google(web), Yahoo(web), Amazon, Twitter, YouTube) が 22type のうちいくつの type の推薦候補に含まれているか数える。google(web) が 22type 中 22type, yahoo(web) が 22type 中 22type, amazon が 22type 中 2type, Twitter が 22type 中 2type, Youtube が 22type 中 5type のサービス推薦候補に含まれていた。このことより、Wordtank アタッチで用いられていた amazon, twitter の利用頻度は低い事がわかった。また、推薦システムで新たに登場したサービスは、18 個あり、より専門性があるサービスをユーザに提供できたと言える。

5.2 満足度

ユーザ満足度とは抽象的な尺度でかつユーザの主観によるものである。そこで、実際にサービス群を見てもらい、アンケート (図 12) に回答してもらった。満足度を測るための質問として、1. もっとも利用したいサービスがあったか、2. 1 番ではないが利用したいサービスがあったか (質問 1 を選択した人以外)、3. 利用したサービスが全てあったか、4. 不必要なサービスがあったかの 4 つを用意した。22type において、既存の Wordtank アタッチに対応しているサービス群と初期値設定で選ばれたサービス群に対してアンケートを実施した結果、以下ようになった (表 3)。スペースの問題で、ここでは 14 人の各問いにおける投票数の合計 (22type 分) を用いる。質問 1, 2 の合計が 308 であることより、全回答者にとって、いずれの type でも、必ず 1 つは利用したいサービスがあったことがわかる。また、質問 3, 4 の結果を既存システムに対応しているサービス群と比較すると、推薦システムを用いることでコンテンツが充実したと言える。

千と千尋の神隠しなど映画タイトル(Film)を調べる場合、
{google検索(web)、yahoo検索(web)、YouTube、yahoo検索(画像)、Hulu)の中には、*

- ☐ もっとも利用したいサービスがあった
- ☐ 1 番ではないが利用したいサービスがあった
- ☐ 利用したいサービスが全てあった
- ☐ 不必要なサービスがあった

図 12 評価用アンケート

	質問 1	質問 2	質問 3	質問 4
Wordtank アタッチに対応しているサービス群	265	43	34	195
推薦サービス群	292	16	55	118

表 3 満足度アンケート結果

6. 終わりに

本論文では、Wordtank アタッチの懸念点である、不必要なリンク生成、サービス増加に伴う限界や利用コストの増加を解消するために、type 付きアタッチ機能及びサービス推薦システムを提案、実装した。

今後の課題としては、主に単語の充実・単語に type を割り当てる方法 (単語と type の関連性決定) を考えることがあげられる。本研究では時間の関係上、日本語版 DBpedia の単語とクラス情報を借りた type 付けを DBpedia のオントロジ依存せず、知識ベースなどを用い独自の type ラベリングルールを考案できれば、typed_wordtank が保持する単語をソース元関係なく集めることが可能になるのではと予想している。また、ラベリングの自由度が上がったところに伴い一単語が複数の type を持つ場合のサービス推薦方法も必要になる。

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, 戴達也, 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] 金岡慧, 遠山元道. 自動更新型 WIX ファイル生成システムおよび Deep Web に対するアタッチ機構の構築. DEIM2014. 2014.
- [5] <http://mappings.dbpedia.org/server/ontology/classes/>
- [6] http://mappings.dbpedia.org/index.php/Mapping_ja
- [7] J. Ben Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. Data Mining and Knowledge Discovery, Vol. 5, pp. 115153, 2001.
- [8] J.A.Konstan and J. Riedl. Recommender systems: Collaborating in commerce and communities. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, Tutorial, 2003.
- [9] 大谷翔平についてのニュース: <https://headlines.yahoo.co.jp/hl?a=2018010>

9-00101548-fullcount-base

- [10] 検索エンジンまとめ : <http://www.coolinx.net/jp/search/kaigai.html>
- [11] EC サイトまとめ : <https://www.ebisumart.com/blog/casestudy/#1>