【パート II：情報分野研究者のためのオンリーワン共有イノベーションプラットフォーム】

2. TSUBAKI：深い言語処理を特長とするオープンサーチエンジン基盤

黒橋　禎夫[*1]　　新里　圭司[*1]

＊1　京都大学情報学研究科

情報爆発時代の検索エンジン

「情報爆発」という言葉で形容されるように，World Wide Web（WWW）上には膨大な量の情報が発信されており，その種類はニュース記事，百科事典，種々のノウハウ，個人の発する口コミ情報など，多岐に渡っている．このような WWW 上の情報を効率良く利活用するためには，現状のようにページのランキングを行うサーチでは不十分であり，WWW 上の情報の集約・組織化が重要となる．具体的には，次のような技術を考えることができる[4]．

- ユーザの用途や趣向に合わせ検索結果のランキングを自動的に変更する技術
- あるトピックに関する関連概念を整理し，トピックの鳥瞰図的把握を提供する技術
- あるトピックに関する意見の分布を調べ，少数派，多数派などに分類する技術
- 検索結果に含まれる情報の信頼性や矛盾点を検出する技術

このような技術の実現には，その基盤となる検索エンジンが必要となる．現在いくつかの商用検索エンジンで，その検索結果を得るための API が提供されているが，これらを研究・開発の基盤として用いるには以下の問題がある．

(1) API 利用回数や取得可能な文書数に制限がある
(2) インデックスの更新が頻繁に行われ，再現性がない
(3) 検索結果のランキング尺度が公開されていない

そこで我々は，上記の問題点を解決したオープンサーチエンジン基盤 TSUBAKI[☆1] の構築・運用を行っている．TSUBAKI は，日本語 Web ページ約 1 億件を対象とした，研究用途に主眼をおいた検索エンジンであり，透明性・再現性のある検索結果をユーザへ提供する．また，API[☆2] も公開しており，1 日のアクセス数や，取得可能な検索結果数に制限を設けていない．

さらに，TSUBAKI には以下の特徴がある．

- Web 標準フォーマットによる大規模 Web ページの管理
- 深い言語処理を用いたインデキシング

Web 標準フォーマットとは，Web ページの解析結果の共有を目的に，我々が提案した XML 形式のフォーマットである．フォーマット化されたデータには，Web ページを対象とした研究を行う上で頻繁に利用されるデータ，たとえばアンカーテキストやページ内の日本語文，日本語文の言語解析結果などが含まれている．

また TSUBAKI では，ページのインデキシングに深い言語処理の結果を利用している．具体的には，単語だけでなく同義表現や係り受け関係（修飾関係）もインデックスに登録することで，前者で「ことば」の「ズレ」を吸収し，後者で「ことば」と「ことば」の結びつきを重視した検索を可能にしている．**図 -1** は，TSUBAKI をブラウザを通して用いた場合の画面である．画面は，「かぜ薬を飲む時の留意点」を検索した結果であり，「風邪薬を服用する」などの表現を含むページが検索結果として表示されていることが分かる．

本稿では，検索エンジン基盤 TSUBAKI のコンポーネントである，Web 標準フォーマット，言語解析，インデキシングについて述べる．

Web 標準フォーマット

● Web ページの解析結果の共有

自然言語処理コミュニティにおいて，Web ページからの知識獲得など WWW 上のテキストを対象にした研究が進められている．しかし，実際に Web ページを扱った研究を行おうとすると，研究に至るまでに直面する面

☆1　http://tsubaki.ixnlp.nii.ac.jp/index.cgi
☆2　http://tsubaki.ixnlp.nii.ac.jp/api.cgi

図 -1 「かぜ薬を飲む時の留意点」の検索結果

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat
Url="http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html"
OriginalEncoding="Shift_JIS" Time="2006-08-14 19:48:51">
<Header>
 <Title Offset="21" Length="39" Id="0">
  <RawString> 小泉総理プロフィール・信念 </RawString>
 </Title>
... 中略 ...
</Header>
<Text>
<S Id="1" Length="70" Offset="525">
 <RawString> 小泉総理の好きな格言のひとつに「無信不立（信無くば立たず）」があります．</RawString>
 <Annotation Scheme="KNP">
  <![CDATA[* 1D < 文頭 >< サ変 >< 人名 >< 助詞 >< 連体修飾 >< 体言 >< 係：ノ格 >< 区切 :0-4>
小泉 こいずみ 小泉 名詞 6 人名 5 * 0 * 0 NIL < 文頭 >< 漢字 >< かな漢字 >< 名詞相当語 >< 自立 >< タグ単位始 >< 文節始 >< 固有キー >
... 中略 ...
ます ます ます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 NIL < 表現文末 >< かな漢字 >< ひらがな >< 活用語 >< 付属 >< 非独立無意味接尾辞 >
．．． 特殊 1 句点 1 * 0 * 0 NIL < 文末 >< 英記号 >< 記号 >< 付属 >
EOS]]>
 </Annotation>
</S>
... 中略 ...
</Text>
</StandardFormat>
```
図 -2 標準フォーマット化された Web ページの例

倒な処理が多い．具体的には，大規模ページ集合のクロール，クロール結果からの日本語ページ抽出，ページからの文抽出がそれにあたる．文抽出を例に挙げれば，Web ページの文区切りは不明瞭な場合が多く，新聞記事などのテキストデータであれば句点を手がかりに文抽出が可能であるが，Web ページの場合は，HTML タグや顔文字，“(笑)”などの感情表現が文区切りとして利用されることも少なくない．このため，文区切りの検出は泥臭い処理になるが，その一方で最も基本となる処理であるため，ここでの性能は，その後の言語解析，アプリケーションの性能を大きく左右する．そのため，標準となる大規模な Web ページの集合を用意し，上述した研究利用に至るまでに必要な前処理を施し，それらを共有することは重要であり，言語資源としての Web ページの利便性の向上が期待できる．

このような考えのもと，TSUBAKI では，Web から 1 億件の日本語 Web ページを取得し，それらに対し，文抽出などの前処理を施したデータを公開している．データは，我々が提案する Web 標準フォーマットという XML 形式で，ページごとに保存されている．Web 標準フォーマットに変換されたページの例を**図 -2** に示す．Web 標準フォーマットでは，ページのタイトル，URL，リンク情報，日本語文とその解析結果などの情報を 1 つのファイルで集中的に管理しており，データベースなどのリソースを切り替えることなしに，利用したいデータにアクセスできるようになっている．フォーマット内のデータには，既存の XML 文書検索モジュールを利用することで，容易にアクセスすることが可能である．

● Web 標準フォーマットコレクションの構築

2007 年 5 月から 7 月にかけて情報通信研究機構知識処理グループにてクロールされた約 2 億 3 千万件の Web ページから，ページ内のメタ情報，助詞の含有率などを手がかりに 1 億件の日本語ページを抽出した．そして，これらを Web 標準フォーマットに変換し，大規模 Web 標準フォーマットコレクションを構築した．変換に用いた計算機環境は，Intel CPU Xeon 3.0GHz × 4，メモリ 4GB のスペックを持つ計算機 162 台であり，GXP2[1] を用いて並列に変換処理を行った．

上記の環境を用いた結果，日本語 Web ページ 1 億件の Web 標準フォーマット化に約 4 週間要した．この 1 億ページにはおよそ 60 億文含まれており，これらに対し，後述する言語解析が施されている．データのサイズはオリジナルの Web ページが 0.6TB，標準フォーマットは 5.2TB である．どちらも gzip で圧縮後のサイズである．

構築したWeb標準フォーマットコレクションは，TSUBAKIが提供するAPIを利用することで取得可能である．また，このデータは，同じく特定科研情報爆発において運用されている共有計算機環境InTrigger☆3にも配置してあり，InTriggerユーザであれば，APIを介さずに誰でも利用することが可能である．

深い言語処理に基づくインデキシング

TSUBAKIが検索対象としているのは，前節で述べた日本語Webページ1億件である．これらのインデキシングには，Web標準フォーマットに埋め込まれている言語解析結果を利用している．本章では，文に対して適用される言語解析，および解析結果から作成されるインデックスについて述べる．

●言語解析

言語解析としては，形態素解析，構文解析に加え，文内の語・句と同義関係にある表現の対応付けを行う．まず，Webページから抽出された日本語文に対して形態素解析，構文解析を行う．形態素解析とは，文を単語列に分割する処理であり，構文解析とは，単語間の係り受け関係を同定する処理である．形態素解析の際，「こども」「子ども」「子供」のような表記の揺れの解消も同時に行われる．構文解析後，文内の単語または句と，同義関係にある表現（正確には，同義関係にある表現のグループID）の対応付けを行う．これら単語や句の間の同義関係は，国語辞典，Webテキストから自動獲得したものを利用する[3]．

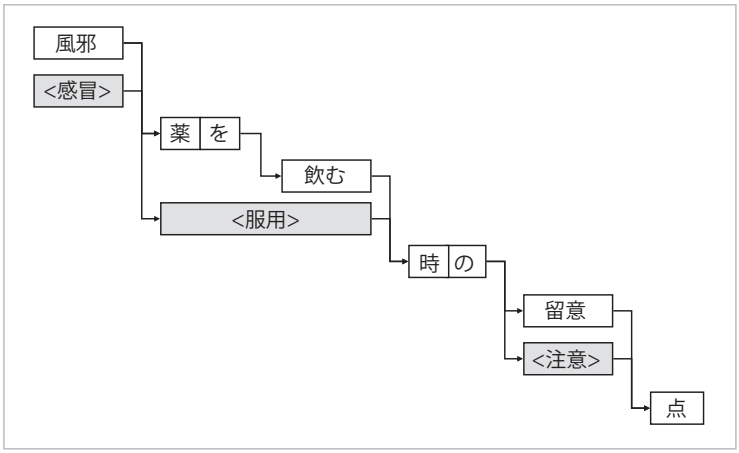
図-3は，「かぜ薬を飲む時の留意点」を言語解析した結果である．近年の言語処理技術の発展に伴い，ここまでの処理（形態素解析，構文解析，同義関係の獲得および同定）は，Webページ中の文のような崩れたものであっても，実用レベルの精度で実行可能である．

●インデキシング

TSUBAKIでは転置インデックス方式を採用しており，各ページの索引となる表現は，対応するWeb標準フォーマット化されたデータから抽出される．既存の商用検索エンジンの多くは，単語だけに注目してインデキシングを行っているが☆4，TSUBAKIではページを言語的に深く解析することで得られる同義表現や係り受け関係についても注目しており，この点がTSUBAKIの特長である．たとえば，図-3に示した「かぜ薬を飲む時の留意点」の解析結果からは以下の表現が抽出される．

単語：風邪，薬，を，飲む，時，の，留意，点
係り受け：風邪→薬，薬→飲む，飲む→時，時→留意，留意→点
同義表現：＜感冒＞，＜服用＞，＜注意＞
係り受け（同義表現を考慮）：＜感冒＞→＜服用＞，風邪→＜服用＞，＜感冒＞→薬，＜服用＞→時，時→＜注意＞，＜注意＞→点

各インデックスに登録される情報，サイズを**表-1**に示す．TSUBAKIでは，フレーズ検索や近接検索など，語の出現位置を考慮した検索をサポートするため，索引表現の出現頻度に加え，出現文，出現位置をインデックスに登録している．



四角内の表現は単語を表しており，矢印は係り受け関係を表す．また，<>で囲まれた表現は，対応する語・句と，同義関係にある表現のグループIDを表す．

図-3　言語解析結果の例

	単語	係り受け	同義表現	係り受け（同義表現を考慮）
索引表現自身	○	○	○	○
文書頻度	○	○	○	○
出現文書情報	○	○	○	○
出現文情報	○	×	○	○
出現位置情報	○	×	○	○
サイズ [TB]	1.17	0.89	1.84	4.81*

* 同義表現を考慮した係り受けインデックスについては，データサイズを小さくするため，1億ページ中で文書頻度が10以上のみ．

表-1　TSUBAKIで用いるインデックスデータ

☆3 https://www.logos.ic.i.u-tokyo.ac.jp/intrigger/registration/

☆4 最近では，言語解析の結果を利用した検索エンジンPowerset（http://www.powerset.com/）も登場しているが，ベータ版の公開にとどまっており，本稿執筆時点では，実際にWebページを対象にした検索はできない．

パラメータ	型／値	説明
query	*string*	検索クエリ(utf8)をURLエンコードした文字列．検索結果を得る場合は必須．
start	*integer*	取得したい検索結果の先頭位置．
results	*integer*	取得したい検索結果の数．
logical_operator	AND/OR	検索時の論理条件．デフォルトはAND．
only_hitcount	0/1	ヒット件数だけを得たい場合は1，検索結果を得たい場合0．デフォルトは0．
id	*string*	個別の文書を取得する際の文書ID．オリジナルのWeb文書，または標準フォーマット形式の文書を得る際は必須．
format	html/xml	オリジナルのWeb文書，または標準フォーマット形式のWeb文書のどちらを取得するかを指定．idを指定した際は必須．

表-2　APIで指定可能なリクエストパラメータの一例

検索スペックと利用事例

TSUBAKIでは，さまざまな検索条件をサポートしており，たとえば，通常の商用検索エンジンにも実装されているフレーズ検索に加え，クエリ中の単語がN単語以内に現れているかどうかを条件にする近接検索や，クエリに含まれる係り受け関係の有無を条件にした検索などが可能である．検索条件に一致するページは，クエリとの関連度に従ってソートされユーザへと提示される．検索クエリと文書の関連度は，OKAPI BM25[2]を基に求めている．

図-1は，「かぜ薬を飲む時の留意点」をTSUBAKIで検索した画面である．「かぜ」と「風邪」，「薬を飲む」と「服用」などの同義表現，「薬」と「飲む」の間の係り受け関係を用いて検索することで，適切なページを上位に提示できている．仮にGoogleなどの商用検索エンジンに同じクエリを与えた場合，自然文によるクエリを適切に扱えないため，望ましい検索結果は得られない．

検索は，通常のブラウザ検索に加え，APIを用いて行うことも可能であり，通常検索と同様にさまざまな条件を指定して検索することが可能である．**表-2**にAPIで利用可能なパラメータを示す．APIを用いた検索は，表のパラメータを用い，REST形式でサーバへアクセスすることで実現される．その実行速度は，1クエリにつき1000件分の検索結果を得るのに20秒程度である．

TSUBAKI APIはさまざまな場面で利用可能であるが，現在までに以下の目的で用いられている[4]．

(1) 知識獲得のための，大規模構文解析済みデータの取得
(2) 類義語・関連語獲得における，ヒット件数に基づく語と語の共起の強さの計算
(3) 質問応答システムにおける，解答を含むWebページの取得
(4) 検索結果クラスタリングシステムにおける，クラスタリング対象となるページの取得

今後の展開

本稿では，開発・運用を進めているオープンサーチエンジン基盤TSUBAKIについて述べた．TSUBAKIでは，日本語Webページ1億件を対象とした検索が可能であり，APIを介して誰でも自由に検索結果を取得できる．その特徴としては，(1) Web標準フォーマットによるWebページの管理および共有，(2) 深い言語処理を用いた柔軟な検索が挙げられる．

今後の課題は，より多くのユーザがストレスなく利用できるように，計算機環境，ソフトウェアの整備を進め，検索速度の向上，検索機能の強化をはかる予定である．さらに，ユーザが開発した検索モジュールをTSUBAKIの計算機環境にアップロードすることで，共通のデータセットを用いて簡単に検索指標を評価できるプラットフォームを構築し，公開する予定である．現在はそのために，評価データおよびソフトウェアの整備を行っているところである．

参考文献
1) Kaneda, K., Taura, K. and Yonezawa, A. : Virtual Private Grid : A Command Shell for Utilizing Hundreds of Machines Efficiently, In 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2002) (2002).
2) Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A. and Lau, M. : Okapi at TREC, Text REtrieval Conference, pp.21-30 (1992).
3) Shibata, T., Odani, M., Harashima, J., Oonishi, T. and Kurohashi, S. : SYNGRAPH : A Flexible Matching Method based on Synonymous Expression Extraction from an Ordinary Dictionary and a Web Corpus, Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008) (2008).
4) 鳥澤健太郎，中川裕志，黒橋禎夫，乾健太郎，吉岡真治，藤井　敦，喜連川優：キーワードサーチを超える情報爆発サーチ―自然言語処理で価値ある未知をマイニング―，情報処理，Vol.49, No.8, pp.890-896 (Aug. 2008).

（平成20年5月2日受付）

黒橋　禎夫（正会員）
パートI「キーワードサーチを超える情報爆発サーチ」を参照

新里　圭司：shinzato@nlp.kuee.kyoto-u.ac.jp
昭和54年生．平成18年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．同年より京都大学大学院情報学研究科特任助教．自然言語処理の研究に従事．