WEB 文書を対象とした KWIC システム

吉平健治† 武田善行‡ 関根聡†
†ランゲージクラフト研究所 ‡豊橋科学技術大学

１． はじめに

言葉を調べるには、言葉が実際にどのように使われているかを生のデータで調べることが重要である。現在、規模が大きく自由に入手可能な言語データとしては WEB があるが、このデータを収集、整理、検索することは規模の問題から容易ではない。WEB データを対象にした KWIC システムはすでに、[KWiCFinder homepage] [WebKWIC homepage] [WebCorp homepage]などあり、また WEB データの KWIC を利用したシステムには [山本他 2003]があるが、それらは WEB のかなり限定された一部を利用しているか、サーチエンジンが返す（内容的に偏った）少ないデータを利用しているものである。しかしながら、言語の研究では対象データの偏りのなさや規模が重要になることがあり、データの規模が大きいことによる困難を克服していかなければいけない。例えば大きな問題として以下のような問題がある。データの収集においては、クローラーがネットワークにかける負荷やネットワークの幅と収集速度の問題を解決しなければならない。集められたデータの規模は現在一般に購入できる IDE 規格のハードディスクの最大容量を超えており、簡単に整理できるものではない。また、検索をする際には、データは 32 ビットで表される４G バイトを超えており、この規模に対応した検索システムを使用しなければならない。

我々は、この大規模な文書を対象とした KWIC システムを作成した。約 350G バイトの WEB ページにある約 100 億文字以上の日本語データを対象にしている。収集は、この目的のための独自のインターネットアドレスを確保し、ネットワークに負荷をかけないように注意をしながら、2 ヶ月以上かけてデータを収集した。整理には cdb によるコンスタントデータベースシステムを利用し、容易かつ迅速にデータへのアクセスができるようにした。検索では独自に開発した 40 ビット長インデックス対応のサフィックスアレーによって、任意の文字列を含む文を瞬時に検索し抽出することができる。KWIC のインターフェースとしては一般的な機能である文字揃えやソートをサポートしている。

２． WEB データ

大規模な WEB データはすでに国立情報学研究所で、NTCIR の Web Retrieval Task を目的に 100G 相当の約 1100 万ページのデータが集められている。[Eguchi et. al 2002] しかし、このデータは著作権の問題からタスク参加者にしか入手できない。そこで、我々は独自に WEB データを収集した。対象とした WEB データは、内容の信頼性を重視し、Yahoo! Japan のカテゴリされているページを起点として収集した。クローラーツールには、UNIX 上で標準的な、GNU Wget を利用した。

WEB には様々なファイルがあるが、文書のみを集めることを目的としているため、対象としたファイルは html,htm または txt の拡張子がついたページだけとした。また robots.txt の作法にしたがい、ダウンロードを明示的に拒否しているサイトもしくはページからのファイルのダウンロードは行っていない。１サイトからは最大 3500 ページを収集した。収集したデータ規模の概要は表 1 のようになっている。

サイト数	２９万サイト
ページ数	２０６６万ページ
全ファイル規模	３５０G バイト
文章部分規模	２０．１G バイト
文章部分文字数	１１８億文字

表 1　データの規模

３．　サッフィックスアレー

我々が構築した KWIC システムでは、大規模な文書の高速な検索を実現するための索引付け法として、データ構造にサフィックスアレーを用いた。検索対象となるテキストにおいて、ある位置からテキスト末尾までの範囲によって定まる文字列をサフィックスと呼ぶ。サフィックスは文字列の先頭位置によって一意に定まる。この先頭位置(インデックスポイント)の配列を、対応するサフィックスの辞書順でソートしたものがサフィックスアレーである。サフィックスアレーの構成要素を

図 1 に示す。我々のシステムでは、インデックスポイントを 40 ビットとして実装したため、最大 1TB までのデータを索引付けすることができる。検索時には二分探索を行う。

(a)テキストとインデックスポイントの対応

テキスト	a	p	p	l	e
インデックスポイント	0	1	2	3	4

(b)ソートされたサフィックス

インデックスポイント	サフィックス
0	apple
4	e
3	le
2	ple
1	pple

(c)サフィックスアレー

サフィックスアレー	0	4	3	2	1

図 1 サフィックスアレーの構成要素

サフィックスアレー採用による最も大きな長所は、任意文字列の検索が可能になるという点である。メモリ効率のより良い索引付け法として転置ファイル法なども考えられるが、多言語情報への拡張性や、既存の文法規則にそぐわないような例外的な表現への対応可能性を考えた場合に不利である。特に、新聞記事や百科事典のような表現の統制が強いデータに比べて, WEB データは例外的な表現やノイズを多く含むことが知られている。また、研究・分析用ツールとしての見地から、例外的な表現やノイズの分析が出来ることは利点といえる。

４．　KWIC システム

KWIC は"KeyWord In Context"の略号で、検索キーワードを中心に、その前後の文脈を同時に表示する索引手法である。今回開発した KWIC システムは、表１にある WEB データに対して３．で述べたサッフィクスアレー構造を採用し、Web サービスとして標準的なブラウザから利用できる仕組みを提供している。

基本的な動作は以下の通りである。まず、ユーザーがキーワードを入力し、その後、システムは検索された行数とサンプルを表示する。その結果を見てユーザーは表示方法などを指定し（図２）、その検索結果をファイルとしてダウンロードし、ユーザーは出力結果を獲得する。検索キーワードによっては、出力結果が予想以上に大きくなる可能性があるので、検索結果の数に上限を付けることも可能である。言語研究に対応できるようにするため、KWIC 検索結果の表示には一般的なソートや文字揃えの機能が実装されている。

５．　著作権についての考え方

言うまでもなく、WEB ページにある文章は著作者の著作権のついた著作物でる。しかし、本システムは Google などのロボット型検索エンジンを引き合いに出すまでもなく、その内容を提供しているわけではなく、検索の技術を提供している。検索された対象物は誰でも自由に無料で入手できるものである。また、データを集める際には、robots.txt の定義に従い、ダウンロードされたくないと宣言されているページは収集していないし、削除依頼のあった場合にはそのページをデータから削除する。検索結果の商用利用も禁止した上で、一般利用に向けて公開したいと考えている。

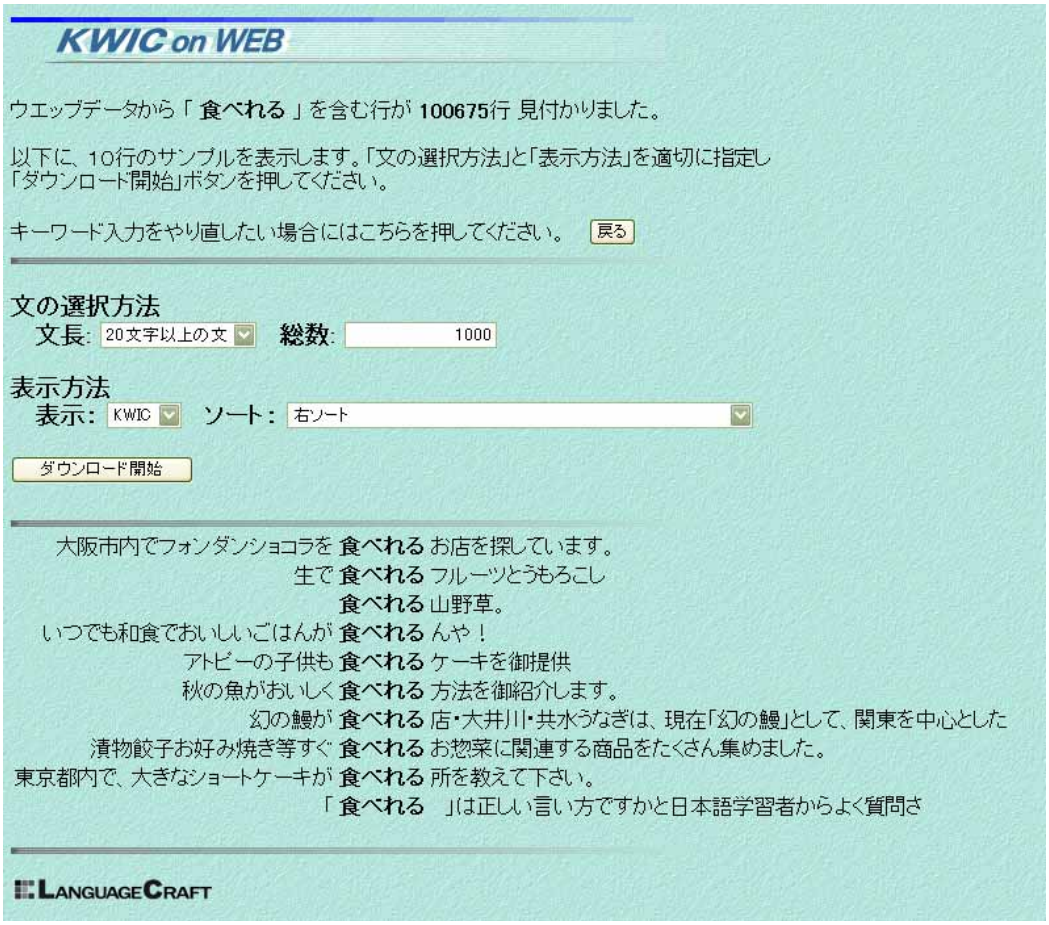

図 2　KWIC 画面のサンプル

6.　まとめ

我々は、大規模な WEB データを対象にした KWIC システムを作成した。言語データの規模は 2000 万ページ、100 億文字以上あり、その中から任意の文字列に対する KWIC データを抽出できる。http://languagecraft.jp/kw にて KWIC システムのデモンストレーションを体験することができる。将来的に、本システムを研究教育利用に限りある程度の量までは無料、それ以上は有料でサービスを提供することを検討している。

参考文献

山本真人、田中久美子、中川裕志　「検索エンジンに基づく多言語用例指南ツール：KIWI」言語処理学会第９回年次大会

Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando and Kazuko Kuriyama: “Overview of Web Retrieval Task at the Third NTCIR Workshop”, NTCIR Workshop 3 Meeting OVERVIEW, National Institute of Information

William H. Fletcher "Making the Web More Useful as a Source for Linguistic Corpora", 2002 North American Symposium on Corpus Linguistics (KWiCFinder)

KWiCFinder homepage: http://www.kwicfinder.com/KWiCFinder.html

WebKWIC homepage: http://prairie.lang.nagoya-u.ac.jp/program/webkwic.html

WebCorp homepage: http://www.webcorp.org.uk/

Web Concordancer homepage: http://www.edict.com.hk/concordance/