Web から動的に獲得した参考情報を利用する文章作成支援システム

中村 和正[†]　　吉永直樹[‡†]　　鳥澤 健太郎[†]
[†]北陸先端科学技術大学院大学　[‡]日本学術振興会
{kaz-naka, n-yoshi, torisawa}@jaist.ac.jp

1 はじめに

本研究では，ユーザが文書を作成する際に，作成する文章に関連する情報を Web から自動的に収集し，提示することにより，文書の作成支援を行うシステムを提案する．

近年コンピュータを用いて文書を作成する際には，作成しようとする文章の参考となる関連情報を Web から収集する機会が非常に多い．例えば Blog で，その日に自分が関与した事物（訪れた場所，買った商品，参加したイベントなど）に関する意見や情報の記述を行う際にも，記述の正確性・詳細性を高めるため 1) 事物の正式名称を確認したり，2) 事物に関する第三者の意見を参考にしたり，3) 訪れた場所のアクセスの方法，イベントの正確な日時，商品の仕様など事物の様々な側面（**属性**）の情報を確認したり付加したりするなど**情報要求**が頻繁に生じる．このような情報要求が生じると，ユーザは能動的に商用の全文検索エンジンを利用するなどして，書く作業を中断して参考情報を収集する必要がある．

本研究では，ユーザが文書を作成する際に頻繁に発生する情報要求に伴う「調べる」作業を極力抑えるために，1) ユーザの検索要求をこれまで知識獲得に使われていた**語彙統語パターン** [1] を利用して発見し，2) その情報要求タイプに応じた知識をWebからオンデマンドで獲得し，ユーザに提示することを目指す．ユーザの情報要求の発見および，Webから得られた知識の提示は，**AJAX**[*1] を用いることでシームレスに行い，極力書き手の手間を発生させないように工夫する．ユーザの検索要求は，実際には様々なタイプのものが存在すると思われるが，本研究では事物として具体物（例えばドイツやタイタニックなど）である固有名詞に焦点を当て，吉永と鳥澤による既存研究 [2] を用いて，具体物の様々な側面，すなわち属性とその属性値を獲得し，ユーザに提示する．

これにより，例えば「日本酒である太平山の原料米」に関して文章を作成する場合，ユーザがそれを知らなかったとしても本システムが「太平山の原料米は山田錦である」と情報を提示し，それを参考に Web を調べることなくスムーズに文章作成を続行することができるようになる．

2 研究の背景と特色

本研究では，ユーザ（クライアント）からの様々な情報要求にリアルタイムで応じるために，重い処理を行う情報抽出のモジュールはWeb文書を直に持っているサーバー側で行う．本節では，このようなクライアント・サーバー形式のWebアプリケーションを実現するために近年よく用いられているAJAX[*1]について簡単に説明する．

AJAXとは，JavaScriptのHTTP 通信のためのオブジェクトであるXMLHttpRequestを用いて，ブラウザのHTTPによる画面遷移とは非同期にサーバーとデータをやり取りし，動的にページ内容を変更する仕組みを指す．これにより，基本的に動的に情報更新されることのないWebページであるにも関わらず，シームレスな表示と画面遷移のない情報の更新が可能となる．AJAXの応用例として有名なものにGoogleマップ[*2]でのスクロール操作やシームレスな地図データの表示，Gmail[*3]のドラッグ操作や画面遷移のないメール情報の更新などがある．

[*1] AJAX（Asynchronous JavaScript+ XML）とはJavaScriptやXMLなどを用いたユーザーインタフェース構築技術の総称.

[*2] Googleマップ　http://maps.google.co.jp

[*3] Gmail　http://mail.google.com

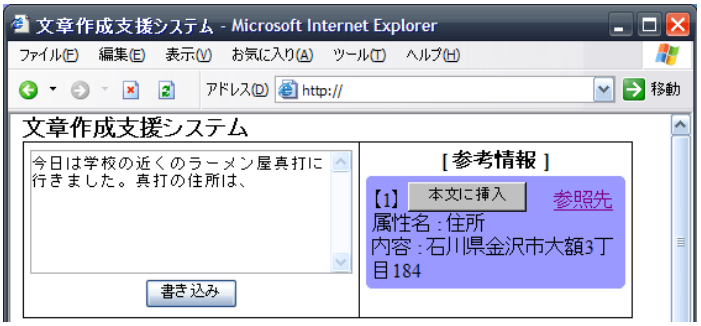

図 1：文章作成支援システム

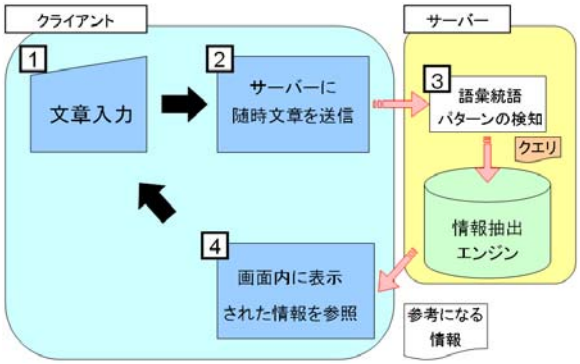

図 2：提案システムの処理の流れ

本研究では，AJAX で検索に関わる処理全てをバックグラウンドでサーバーに行わせることにより，本来情報要求が発生するたびにクライアントが行う必要があった操作手順を減らしている．また，AJAX は，シームレスに参考情報を表示する目的でも用い，本システムの操作性を高め，使い勝手を向上させる．

3 提案システム

本節では本研究で提案する文章作成支援システムについて説明する．

3.1 本システムの概要

本システムのクライアント部に含まれるエディタの画面を図 1 に示す．本システムはブラウザ上で動作する Web アプリケーションであり，クライアントにはブラウザ以外のソフトウェアは必要なく，クライアント部のエディタ画面には AJAX コードが埋め込まれているだけである．

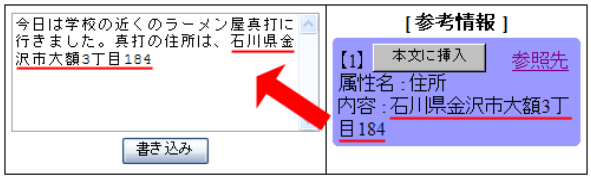

図 3：提示された情報を本文に挿入

本システムを利用する際の実際の手順は図 2 のようになる．本システムはユーザの入力を監視し，一定間隔で入力中の文章をサーバーに送り続ける．例えば「真打の住所は」と文章を入力（図 2-1）したタイミングで，本システムが自動的に AJAX を用いてバックグラウンドで入力中の文章をサーバーへ送信（図 2-2）したとする．文章を受け取ったサーバーは，そこから語彙統語パターンを検知（図 2-3）し，情報抽出エンジンへの問い合わせを行うためのクエリを作成して，情報抽出エンジンに入力し，得られた結果をエディタ画面内に参考情報として表示（図 2-4）する．ユーザは必要に応じて表示された情報を参考にしたり，本文に挿入（図 3）することができる．

本システムでは文章作成と情報抽出をクライアントとサーバーに分離させているため，ユーザはサーバーの処理の完了を待たずとも即座に文章の入力を続行することができる．文章入力を継続し，サーバーから参考となる情報が提示されてから改めてその部分を入力するといったことも可能である．

例えば Blog を書いているユーザが「今日は学校の近くのラーメン屋真打に行きました．真打の住所は，」という文章を入力しているとする．ここでユーザが真打の住所を知らなかったとしても，本

システムが「真打の住所」を情報要求として検知できれば, 真打の住所に関する情報を Web から発見し,「石川県金沢市大額 3 丁目 184」という情報を画面内に提示することで，ユーザの文章作成を支援できる（図 3）.

以上のように文章入力時において情報要求が発生したとしても，検索にかかる手間をバックグラウンドでサーバーに任せることによって，従来のように改めて Web ブラウザを起動し, Web 上の検索エンジンにアクセスし，必要な情報を探すといった「調べる」作業のために文章入力を中断させられることがなくなり，ユーザはスムーズな文章入力を続行することができる.

3.2 語彙統語パターンに基づく検索要求の発見

文書入力中のユーザの情報要求を正確に推定することは，入力途中の未完成の文からユーザの意図を推測しなければいけないため，それ自体非常に難しい問題である．本研究では，情報要求のタイプを，対象とその属性に関する情報に限定し，従来自然文からの知識獲得に用いられてきた語彙統語パターン[1]を用いてユーザの情報要求の予測を行う.

語彙統語パターンは，知識獲得を行う際に獲得対象の知識（上位・下位関係，属性・属性値関係等）を含む語彙，統語的な文脈を表現したパターンである. 例えば上位･下位関係の獲得であれば，“A と呼ばれる B”といったパターン [3,4] が属性・属性値の獲得であれば，“A の B は C である”のようなパターン [5,6] が用いられる．これらの語彙統語パターンは，高い割合で目標とする知識を表現するため，知識獲得に用いられているのであるが，逆を返すと，このような統語パターンにマッチする表現をユーザが用いたときには，ユーザは高い確率で対応する知識に関する情報を記述することが予想できる．例えば“王貞治と呼ばれる”という入力が行われれば，ユーザは次に「野球選手」「監督」「ホームランバッター」といった「王貞治」の上位語を書くことを意図していると推測できるし, “ホテルオークラの宿泊料金は” という入力が与えられれば，ユーザは「ホテルオークラの宿泊料金」を記述しようとしているということが予測できる.

そこで本研究では，ユーザの情報要求の中でも特に頻度が高いと思われる，具体物の属性・属性値情報に焦点を当て，属性記述に関する語彙統語パターン「A の B」を利用して，ユーザの情報要求を予測することを試みる．具体的には作成中の文章の末尾を形態素解析し，その結果に「A（名詞句）の B（名詞句)」という語彙統語パターンを検知した場合に先行する名詞句 A を具体物名，A から“の”を介して係り受けの関係にある名詞句 B を A の属性語として，吉永と鳥澤の属性・属性値抽出システム [2] へのクエリを生成する.

3.3 Web からの属性・属性値の抽出

本研究では，ユーザの入力文から語彙統語パターンを用いて検知された具体物とその属性に関して，情報要求を表現するクエリを生成し，吉永と鳥澤 [6] により提案された属性・属性値抽システムの入力として与えることで具体物の属性・属性値を獲得する.

仮にユーザの検索要求を検索クエリとして適切に表現できたとしても，その検索クエリに情報抽出システムが高速に回答することが出来なければ, 我々のシステムがリアルタイムの動作を意図している以上，実用的システムとして運用することは難しい．我々の採用した吉永と鳥澤による属性・属性値抽出システムは，具体物の属性語を利用することで属性･属性値を含む可能性の高い Web ページを効率よく収集し, 得られた Web ページ中で入力の属性語を含む範囲を特定した後，入力の属性語の出現している文脈から属性・属性値の記述パターンを導出することで，属性・属性値を非常に高速に獲得することが可能である.

彼らが想定している入力は，具体物とその上位語であるが，上位語は実際には具体物の属性語を得ることに利用されており, (我々が検索クエリとして与える) 具体物とその属性語を入力としても, 彼らのシステムを運用することは可能である.

彼らは論文で 50 個の具体物とその上位語のペアを入力として，全文検索エンジンを用いて 10

プレイステーション３の価格	→ 62,790円 (HDD 20GB)、オープンプライス (HDD 60GB)
任天堂の本社	→ 601-8501 京都府京都市南区上鳥羽鉾立町11-1
東京大学の住所	→ 東京都文京区本郷7-3-1
アメリカのGDP	→ 10兆7280億ドル(2004年:名目)
名探偵コナンの作者	→ 青山剛昌
タイタニックの監督	→ ジェームズ・キャメロン
Googleの時価総額	→ 150,841 百万
エキスポランドの住所	→ 大阪府吹田市千里万博公園1-1
ホテルオークラの住所	→ 〒105-0001 東京都港区虎ノ門2-10-4
革命の作曲	→ ショパン
ドイツの首都	→ ベルリン 北緯 52度30分 東経 13度22分
トルコの公用語	→ トルコ、キプロス、北キプロス、ブルガリア
ドラゴンボールの作者	→ 鳥山明
スーパーファミコンの発売日	→ 1990年11月21日
愛媛県の県庁所在地	→ 伊予市
木村拓哉の出身地	→ 東京都
EIZOの本社	→ 924-8566 石川県白山市下柏野町153

図 4：属性・属性 値抽出の一例

件の Web ページを知識源として収集した場合，74%程度の具体物に対して正しい属性・属性値を獲得することに成功したと報告しており，また 10 件程度の Web ページを知識源として属性知識の獲得を試みた場合, 経験的にかかる時間は 10 秒程度であり，我々の期待する高速性を十分達成できていると結論づけられる.

図 4 に，実際に抽出できた一例を示す．適切な具体物名と属性名が与えられれば，時事的な情報である商品の価格や株価，ホテルなどの住所，映画の監督名や芸能人の出身地など，多岐に渡る属性値が抽出できることが確認できた.

定量的な評価はまだ厳密に行っていないものの, 具体物とその適切な属性を 「A の B」という形で著者が選んで入力した限りでは，50%ほどの確率で適切な情報が 1 番目に提示され，2～3 番目までの候補の提示を要求すると 60～70%ほどの確率で適切な情報が提示された．また，吉永と鳥澤の論文にあるように，今後の精度向上も見込めることから，十分に文書作成の支援ができる情報抽出システムになるものと期待している.

4 まとめ

本論文では，入力中の文章に関連した情報を AJAX と既存の情報抽出エンジンを用いて提示することで，ユーザへの文章作成を支援する手法の 1 つとして，本システムを提案した．従来の知識獲得に用いられてきた語彙統語パターンを手がかりに，システムはユーザの情報要求を検知する．そのため本研究の目的である，文章作成の中断となるような「調べる」作業を発生させることなく，スムーズに文章を作成することが可能となった.

今後は，入力中の文章から検知する語彙統語パターンを増やすことで，ユーザの様々なタイプの情報要求に対応し，情報要求のタイプに応じて適切に情報抽出エンジンを使い分け，多岐に渡る情報の提示を行いたいと考えている.

参考文献

[1] Marti A. Hearst, Automatic acquisition of hyponyms from large text corpora. In Proc. of COLING, pp.539 – 545, 1992.

[2] 吉永直樹, 鳥澤健太郎, Web からの具体物の属性・属性値情報の自動獲得, 言語処理学会第 13 回年次大会発表論文集, 2007.

[3] 安藤まや, 関根聡, 石崎俊. 定型表現を利用した新聞記事からの下位概念単語の自動抽出.情報処理学会研究報告, 2003-NL-157, pp.77-82, 2003.

[4] 今角恭祐, 並列名詞句と同格表現に着目した上位下位関係の自動獲得, 九州工業大学修士論文. 2001.

[5] 高橋哲朗, 乾健太郎, 松本裕治. 言語パタンと統計的共起尺度による属性関係抽出. 言語処理学会第11 回年次大会発表論文集, pp.432-435. 2005.

[6] 徳永耕介, 風間淳一, 鳥澤健太郎. 属性語の Web 文書からの自動発見と人手評価のための基準. 自然言語処理, 13(4), 2006.