修士論文 2000 年度 (平成 12 年度)

アノテーションによる
オンライン文書の単語理解支援システム

慶應義塾大学 大学院 政策・メディア研究科

東中 竜一郎

rh@sfc.keio.ac.jp

修士論文 2000 年度 (平成 12 年度)

論文題目

アノテーションによる
オンライン文書の単語理解支援システム

論文要旨

現在、さまざまなオンライン文書を誰もが簡単にネットワークを介して手に入れられるようになったが、それらの文書はさまざまな専門分野に関連し、語彙も多様なため、閲覧者にとっては必ずしも分かりやすい単語だけで構成されているとは限らない。
そこで、文書中の単語にあらかじめアノテーションと呼ばれる付加情報を与えておき、文書を閲覧する際に、中間サーバ（プロキシ）と呼ばれる仲介役のコンピュータを介して、その情報を利用することで、閲覧者の理解を促進する手法を考案した。具体的には中間サーバが閲覧中の文書を閲覧者が理解しやすいように、閲覧者の要求に応じて書き換える。
文書に含まれるすべての語にあらかじめ情報を与えておくのはコストが高いので、閲覧者にとってどの単語を理解困難であるかということを知るために、閲覧者がオンライン文書を閲覧している際に、理解困難な単語を共有のサーバに送るシステムを構築し、そのデータをもとに、アノテーションすべき個所を絞り込む。
さらに、閲覧者自身が単語の説明文を登録できる仕組みを開発し、蓄えられたデータを辞書として使うことで、以後閲覧者の理解の促進やオンライン文書間の関連付けに利用する。

キーワード

文書理解支援、アノテーション、
言い換え、オンライン辞書、自然言語処理

執筆者

慶應義塾大学 大学院 政策・メディア研究科

東中 竜一郎

Abstract of Master's Thesis Academic Year 2000

Title

A system supporting word/phrase understanding
of online documents using annotations

Summary

Nowadays, one can easily get many kinds of online documents via the Internet. These documents however vary in fields and terminology, and it is not necessarily true that these documents are composed only of understandable words or phrases.
Therefore, we have come up with a solution to assist word/phrase understanding by transforming online documents into more understandable ones. To accomplish this, we have developed a system, in which anyone can attach extra information called annotations to online documents, and view the documents through proxy servers, which convert or paraphrase the documents based on the annotations.
Since attaching extra information to all expressions that might be difficult seems costly, our system has a mechanism, by which ordinary viewers can tell the system which words/phrases are difficult. Based on the information, it is possible to pinpoint the items to annotate. Moreover, the system enables viewers to register definitions of the items, which will be used as dictionaries and will be useful in helping other viewers understand the document. The data collected can also be used to find similarities among online documents.

Key Words
document understanding support system, annotation, paraphrasing,
online dictionary, natural language processing

Author

Keio University Graduate School of Media and Governance

Ryuichiro Higashinaka

目次

第１章
はじめに

１．１　研究目的

Web上にさまざまな文書が存在するがそれらは多様な内容と形式を持つために文書によっては背景知識を持たない人にとって分かりにくかったり、意味が取れなかったりすることがよく起こる。たとえば専門用語であったり、一部でだけ通用するような単語を含む文書であると他の分野に携わっている人にとっては非常に難解なものになる。

これではWebの発達によって様々な文書にアクセスできる状況になったとはいえ、本当に文書が共有されているとは言えない。誰にとってもある程度理解可能な文書を提供することが可能になればWebの価値はより高まるのではないかと考える。

難解な語を含む文書に出会った場合、通例辞書を引いたり、人にその意味を尋ねたりして解決する。しかし、時間的制約などを考えると、分からない単語に出会う度にそのようなことを行うのは効率が悪い。そのため、オンライン文書における利用しやすい辞書を構築し、可能な限り辞書引き作業を自動化するとともに、その結果を閲覧者に分かりやすく提示する方法が求められる。

現在、Web上ではさまざまな文書が日夜生成されており、新たな単語、専門用語等は着実に数を増している。従来の辞書やオンラインで提供されている辞書も、すさまじいスピードで増えていくそれらの語彙に対応しきれていない感があり、辞書はそれらのダイナミックな語彙の変化に対応する必要がある。さらに、オンライン文書を閲覧する際に利用しやすい形ということを考えると、閲覧中の文書に統合される形が望ましいと考えられる。そうすれば閲覧者は辞書を引いているという感覚を持たずして、文書の内容をより深く理解することが可能になる。

翻訳ソフトなどでは、よくマウス辞書が使用される。分からない単語の上にマウスポインタを置くとその単語の訳がポップアップして表示されるというものであるが、これはオンライン文書と辞書との統合の良い例である。しかし、従来のマウス辞書では語義が解決できておらず、複数語義のある単語について辞書引きを行った場合、その単語に関するすべての語義を表示してしまう。これは使用される辞書が非常に一般的なためと、文書のコンテキストからその単語がどの意味に使われているのかを決定する機構が存在しないためである。従って、語義の候補が複数個ある場合、閲覧者はその中から適切な候補を選ばなくてはならない。

単語の語義を決定できたとして、次に問題となるのは辞書引き結果の閲覧者への提示方法である。上に挙げた定義文のポップアップはその1つであるが、他に「挿入」と「グロッサリー（用語集）」、「原文の言い換え」が挙げられる。挿入は定義文を文中に埋め込む方法で、グロッサリーは文書に出現する難解な用語をその説明と共に別ウインドウに一覧として表示する方法、原文の言い換えは原文の表現を辞書の定義文に基づいて言い換えて提示する方法である。言い換えは意味的な情報を多く必要とし、一般に実現が困難である。たとえば、1つの表現に対して一般に多種多様な言い換えの可能性が存在することや、閲覧者によって言い換えの嗜好が異なる点などである。

本研究の最終的な目的は閲覧者の単語理解の促進であるが、そのためには以下の目的が達せられなくてはならない。

（1） オンライン文書における利用しやすい辞書の作成
（2） 辞書内容の効率的な閲覧者への提示

１．２　研究概要

まず、文書中の単語にあらかじめアノテーションと呼ばれる付加情報を、語義情報として与えておき、文書を閲覧する際にその付加情報を中間サーバ（プロキシ）と呼ばれる仲介役のコンピュータを介し利用することで、閲覧者の理解を促進する手法を考案した。具体的には中間サーバが閲覧中の文書を、閲覧者が理解しやすいように、閲覧者のデマンドに応じて書き換える。たとえば処理の流れは図１のようになる。

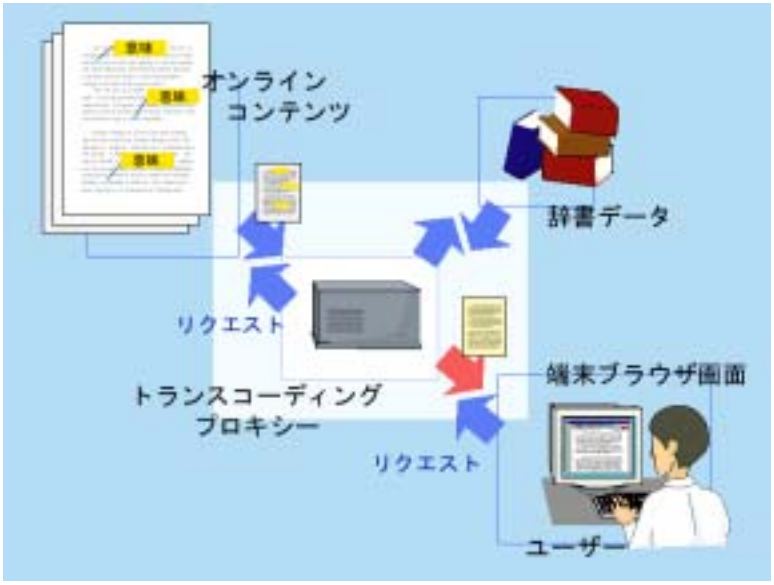

図 1　処理の流れ

しかし、文書に含まれるすべての語にあらかじめ情報を与えておくのはコストが高いと考えられるので、多くの閲覧者にとって理解困難となる語に的を絞って情報を用意するのが望ましく、また、語の情報を閲覧者に提示するだけでなく、その情報を元に用語辞書を作成するなど再利用しやすい形にすることによって、語彙情報を作成するコストを引き下げる必要がある。

第１に、閲覧者にとってどの単語を理解困難であるかということを知るために、閲覧者がオンライン文書を閲覧している際に、マウスのドラッグなどの簡単な操作で自分が理解できない、または理解困難な単語を共有のサーバに送るシステムを構築した。そのデータをもとに、文書中のどの単語が、どの程度理解困難であると一般的に考えられているのか知ることができ、アノテーションの個所を絞り込むことに利用できる。

第２に、閲覧中文書に含まれる単語に、閲覧者自身がその単語の説明文を登録できる仕組みを開発した。この仕組みによりオンライン文書の作成者でなくとも文書中の単語についての情報を用意することができる。登録されたデータは閲覧中文書のURLと対でデータベースに蓄積される。蓄えられたデータは、あるURLに対する単語辞書と考えられ、以後閲覧者の理解を促進することに利用されたり、他の文書に関連付けられた単語辞書らとともに、再利用、再構成される。また、単独のコンテンツとしても利用可能である。さらに、それら辞書の類似性を計算することにより、広く散在するオンライン文書間の自動関連付けに使用でき、オンライン文書群から知識を発見する手がかりとなる。

上記のように作成された辞書を閲覧者が効率よく利用する手段として、前述したポップアップ、挿入、グロッサリー、原文の言い換えという４つの提示手法を、プロキシを利用して実現した。

以上に説明してきた流れにより、閲覧者の単語理解支援を実現する。

第２章
関連研究と本研究の特徴

本章では関連研究と本研究の特徴について述べる。関連研究は大きく分けて 2 つの部類に分けられ、ひとつはオンライン文書に直接意味的な情報を埋め込み、オンライン文書の機械的処理を実現するものであり、もうひとつはオンライン文書には直接情報を埋め込まず、メタ情報のみをオンライン上に蓄積し、同様の処理を実現しようとするものである。関連研究について述べた後、本研究との差異を列挙し、表にまとめ、解説する。

２．１　関連研究

２．１．１　トランスコーディング

デジタルコンテンツがあたりまえのものとして世の中に溢れ出したのは 20 世紀の情報技術の進歩からすると必然的であり、それら膨大なコンテンツを活用するための技術もさまざまなものが発明され、進歩を遂げていくことは間違いがない。これまでは、ともかくコンテンツを作成して流通させることが主目的であったのに対し、これからは、それらのコンテンツをいかに賢く利用するか、あるいは、いかに多様に、多目的に利用するか、ということが最も重要な課題になる。

デジタルコンテンツの高度利用の主なものに、パーソナライゼーションとアダプテーションがある。デジタル放送の映像や Web ページなどのデジタルコンテンツをユーザの好みに応じて変換することをパーソナライゼーションと呼び、それらのコンテンツをパソコンや PDA(Personal Digital Assistant)や携帯電話などのデバイスの特性に合わせて変換することをアダプテーションと呼ぶ。

デジタルコンテンツのパーソナライゼーションとアダプテーションを合わせたものをトランスコーディングと呼ぶ。現状では、インターネットへのアクセスはパソコン経由で行なわれることが多い。しかし、この様相は近年、急激に変わりつつある。パソコンに加えて、携帯電話や PDA、テレビ、カーナビなどを使ってインターネットにアクセスする機会がますます増加するだろう。このとき重要となるものがトランスコーディングである。

たとえば、パソコンで表示することを前提にして作成した Web ページを携帯電話などで表示す

る場合、画像の縮小やテキスト部分の圧縮といった操作を自動的に行う必要がある。トランスコーディングには、少ない伝送容量を使ってサーバからクライアントにコンテンツを配信できるという利点の他に、ユーザの嗜好に応じた理解しやすいコンテンツを生成できるといった利点がある。トランスコーディング技術を使えば、画面の表示機能やデータ伝送速度など、それぞれ違った仕様や制約をもつ多様な機器に対して、1 つのコンテンツ・ソースから情報やサービスを提供できるようになる[3]。コンテンツ・プロバイダやサービス・プロバイダは、それぞれの機器に対応したコンテンツを別個に用意しなくても済む。具体的な応用例としては、パソコン向け Web コンテンツのトランスコーディングによって、i モード向けのコンテンツを生成するといった利用法がある。現状のようにコンテンツ・プロバイダは、パソコン向けと i モード向けのコンテンツを作り分ける必要がなくなる。

２．１．２　セマンティック・トランスコーディング

このトランスコーディングをさらに進めて、テキストの要約などの内容に基づく処理の精度を高める工夫を盛り込んだのが、長尾らの提案するセマンティック・トランスコーディングである[5]。具体的には、コンテンツに含まれるテキスト文要素に言語的な付加情報（アノテーションと呼ぶ）を加えることによって、要約や翻訳の精度向上を図る。

たとえば、付加情報を使ってコンテンツに含まれるテキスト文の曖昧さを軽減すると、正確な要約や翻訳が期待できる。コンテンツにアノテーションを付ける手間が増すが、重要な情報はアノテーションをつけて正しく情報を伝え、共有すべきという考えに基づいている。このアノテーションはコンテンツの内容理解を促進するものと位置付けられ、現在、長尾らは多くの人々が文書の内容に関する補足的情報を付加できるような枠組み作りや、その情報を加味して文書を読者に適した形に加工する仕組み作りに取り組んでいる。

セマンティック・トランスコーディングは、基本的にテキストコンテンツの処理を中心としたものであるが、その手法は映像や画像などの非テキスト・コンテンツの加工にも応用され、マルチメディア・データを含むコンテンツに適用できる。

セマンティック・トランスコーディングは、ユーザが指定した Web 上の新聞記事などのコンテンツを任意の圧縮率で要約して表示したり、テレビ番組などの映像データからユーザの好みに応じた話題だけを抜き出して、ダイジェスト映像を作成するといったことを可能にする。さらに、要約したコンテンツを翻訳したり、テキストを音声化して聴くこともできる。図 2 はセマンティック・トランスコーディングシステムの構成を表している。

コンテンツサーバにおかれたテキスト、画像、音声、映像などのコンテンツはトランスコーディ

ングプロキシサーバによって、ユーザの使用するデバイス（パソコン、携帯電話、カーナビなど）や、ユーザの要求（概要をつかみたい、母国語で読みたい、声で聞きたい、など）に合わせて加工される。

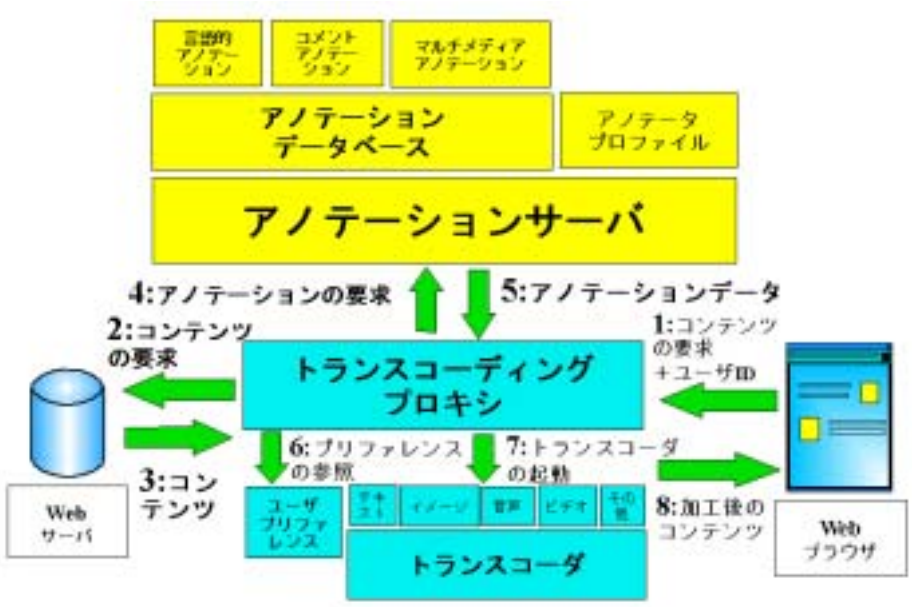

図 2　セマンティック・トランスコーディングシステムの構成

このとき、アノテーションと呼ばれる付加情報を用いて、より精度の高い要約・翻訳を行う。アノテーションはアノテーションサーバに蓄えられている。図 3 で示されるように、アノテーションは、現在の Web に上位構造を作る基盤になる。現在の Web コンテンツが最下層で、アノテーションはコンテンツに情報を付け加えるメタ（上位）コンテンツ、さらにメタコンテンツに対するメタコンテンツのように階層をなしている。

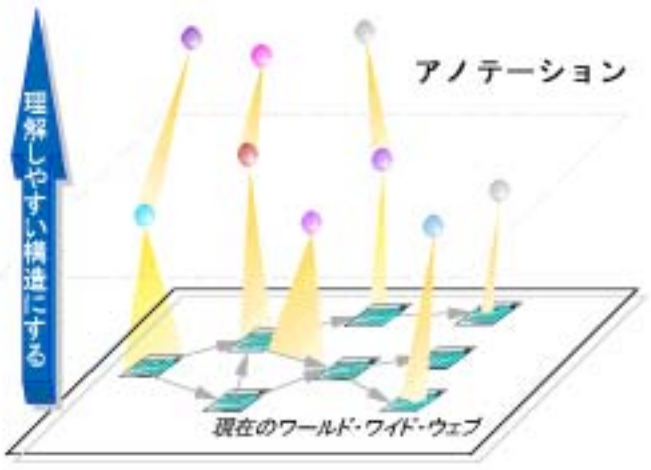

図 3　アノテーションによる WWW の拡張

セマンティック・トランスコーディングの手法を使って、具体的には HTML 文書などの Web コンテンツが抱える、以下の 3 つの課題を解消できるだろう。

（1）HTML (HyperText Markup Language)ではレイアウトなどの文書の表現については規定している。しかし、文書の意味などといった内容に関してはほとんど何も規定していない。
（2）HTML などで記述したハイパーテキストは、各文書間のネットワーク構造を記述できる。ただしリンク情報が常に正しいとは限らず、その修正ができるのは基の文書の著者だけである。
（3）Web 文書の著者は一般にその読者のことを考慮して著作してはいない。なおかつ著者と読者の間に立って吟味・調整する役割の人間も通常はいない。

Web は、新しいスタイルの文書のあり方を示したという点において革新的だったと言えるだろう。Web コンテンツの自由度の高さも疑いようがない。しかし、現状では Web コンテンツを読者が読みやすいような体裁に機械的に変換することは非常に困難である。

図 3 にあるように、従来の Web コンテンツは一枚の平面上に存在する要素郡として捉えることができる。セマンティック・トランスコーディングでは、Web コンテンツを平面から立体に拡張する手法を提案する。コンテンツの各要素に意味や文書構造を示すアノテーションを付加する。このことによって Web コンテンツに、コンテンツの各要素の意味や文書構造を記述した上位構造を築く。代表的なアノテーションの例としては、リンク元の文書に埋め込まれていないハイパーリンクである外部リンクや、コンテンツに対するコメントなどが挙げられる。アノテーションを作成して公開することが容易になれば、Web コンテンツの表現力は大幅に高まり、その利用価値が飛躍的に向上する。

２．１．３ Global Document Annotation

GDA(Global Document Annotation,http://www.etl.go.jp/etl/nl/gda/)は、テキストの文章の意味構造に関するアノテーションである。それは、語間の係り受け、代名詞の指示対象、多義語の意味など、かなり細かい情報を含む。このタイプのアノテーションは、ドキュメントの内容理解に大きく貢献し、文書変換以外にも、たとえば、内容検索や知識発見などに利用される。

GDA は、具体的には XML（eXtensible Markup Language）[1] [4] 形式のタグファイルである。前項のセマンティック・トランスコーディングにおいても GDA が利用されている。GDA は多言語間に共通な意味的・語用論的タグをドキュメントに付与することにより、その機械的な内容理解を可能にし、ドキュメントの検索・要約・翻訳を実用的なレベルで実現するとともに、ドキュメントの作成・公開(共有化)・再利用を考慮した統合的なプラットフォームを構築して、世界的に普及させようという、壮大なプロジェクトである。

一般に、GDA ドキュメントはネットワーク構造を成しており、そのリンクには、タグの入れ子構造よって定義される関係と参照関係の 2 種類がある。また、GDA のタグ集合は 10 項目以上からなるが、さしあたり、そのうちで自動タグ付け作業が比較的大変だと思われる、統語構造、文法・意味関係、語義、照応、修辞関係という 5 項目だけを扱っている。

文法機能(主語、目的語、間接目的語)、主題役割(動作主、被動作者、受益者など)、および修辞関係(理由、結果など)は関係属性によって表示する。関係属性は rel="*"という形で表される。主語、目的語、および間接目的語の主題役割の判断は難しいことが多いので、文法機能(sbj、obj、iob)を用いる。

このようなタグ付けは多くの労力を要すると思われるが、タグ付けツールにいくつかの自然言語処理モジュール(統語・意味解析、照応解析など)を統合することによって、極力人間の負担を減らせるように工夫している。人間がインタラクティブに解析した部分は、事例として次の機会に再利用されるので、それによって解析の精度が少しずつ上がっていくことになる。解析の精度が上がれば、それだけ人間の負担が減ると思われるので、将来的にはタグ付けのコストは十分に少なくなると思われる。

２．１．４　セマンティック・ウェブ

セマンティック・ウェブ（http://www.semanticweb.org/）とは Web 文書を XML で記述して、「<object isa="Personal Computer">ThinkPad</object>は<price unit="yen">180000</price>円です。」という記述から、200000 円以下のパソコンには何がある？などの質問に答えられる仕組みを作ろうというものである。

Web 文書が人間が読むためのコンテンツであるという考えから、機械（エージェント）が世界のことを知るためのデータであるという考えに移行しようという趣旨であるが、これまでの HTML 文書の再利用までは考えられていないようである。しかし、この考えはすでに B2B(Business-to-Business, 企業間電子取引)の枠組みの中ではすでに取り入れられているものであり、特に目新しいものではない。ただし、セマンティック・ウェブは次に述べる SHOE や TopicMaps のアイディアを取り込んで、人間が自由に機械可読な知識をネットワーク上に投入していけるようなインフラに発展していく可能性がある。

２．１．５　ＳＨＯＥ（Simple HTML Ontology Extensions）

ＳＨＯＥ(http://www.cs.umd.edu/projects/plus/SHOE/)[2]とは HTML 文書の中に、内容に関わる概念や関係の形式的な記述（オントロジー）を追加するというものである。これによって、従

来の Web ページを知識源とした意味的な推論が行えるようになる。ただし、オントロジーを追加していく作業は高度なスキルを要求するため、一般のユーザが気軽に参加できるものではない。

以下はオントロジーを追加した Web ページの例である。
ページの最後にこのページの属するオントロジーのベースとなる URL の指定があり、その後、このページ内の単語をその中のカテゴリに入れていく指定がなされている。

```
<HEAD>
<META HTTP-EQUIV="SHOE" CONTENT="VERSION=1.0">
<TITLE> My Page </TITLE>
</HEAD>
<BODY>
<P> Hi, this is my web page.
    I am a graduate student and a research assistant.
<P> Also, I'm 52 years old.
<P> My name is George Stephanopolous.
<P> Here is a pointer to my <A
  HREF="http://www.cs.umd.edu/smith"> graduate advisor.</A>
<INSTANCE KEY="http://www.cs.umd.edu/users/george/">
        <USE-ONTOLOGY
                ID="cs-dept-ontology"
                URL="http://www.cs.umd.edu/projects/plus/SHOE/onts/cs.html"
                VERSION="1.0"
                PREFIX="cs">
        <CATEGORY NAME="cs.GraduateStudent">
        <CATEGORY NAME="cs.ResearchAssistant">
        <RELATION NAME="cs.name">
                <ARG POS=TO VALUE="George Stephanopolous">
        </RELATION>
        <RELATION NAME="cs.age">
                <ARG POS=TO VALUE="52">
        </RELATION>
        <RELATION NAME="cs.advisor">
                <ARG POS=TO VALUE="http://www.cs.umd.edu/users/smith">
        </RELATION>
</INSTANCE>
</BODY>
</HTML>
```

２．１．６ Topic Maps

Topic Maps(http://www.topicmaps.org/)も SHOE と同様に、Web ページに情報を付け加えていくための枠組みのようである。追加する情報には、トピックと呼ばれるキーワードの集合と、トピック間の関係などである。一つのページに複数のトピックが含まれている場合は、トピック間のつながりを定義して、あるトピックから別のトピックに移行する経路を記述することができる。

Topic Maps は、特定のトピックに関する情報をかき集めて、ユーザに分かりやすく提示することができる。ただし、やはりトピックやその関係の定義は、オントロジーと同様に高度な知識を必要とすると思われるため、一般のユーザが自由に参加できるようなものにはならないだろう。

２．１．７　Third Voice

Third Voice(http://www.thirdvoice.com/)はＷｅｂページそのものやＷｅｂページ内の単語についてメタ的にリンクを設定できる。この機構によってあるＷｅｂページから関連性のある他のページへのポインタが指し示され、ユーザは効率的に必要な情報群にアクセスすることが可能となる。BrowseUp(http://www.browseup.com/)も同様のシステムである。

２．１．８　ユーザ参加によるオンライン辞書作成

現在、インターネット上では様々な手作りデータベースが増えており、そのなかに、辞書データベースが挙げられるだろう。掲示板のような閲覧者の書き込みによる辞書データベースの作成もいくつかあり、それらはたとえば「みんなで日本語辞書を作ろう！」（http://gakat.pos.to/jisho-tsuku/）などの運動である。これは単語とその説明文をフォーム等に入力し送信すると、その送られてきたデータを管理者が選別し、辞書に登録するというものである。

またＩＭＤＢ(Internet Movie DataBase,http://www.imdb.com/)やＣＤＤＢ（CD DataBase, http://www.cddb.com/）などのデータベースでは、データは主催者側で用意されるが、利用者数が多いため、少しの間違いでもあれば即座にユーザのレポートをもとにデータが修正される。ある意味ではユーザと主催者側が協力しあって、より利用価値の高い辞書の作成を目指していると言える。

２．２　本研究の特徴

まず、本研究は以上に述べた関連研究に、以下の点で類似している。

（１）セマンティック・トランスコーディングにおける、オンライン文書に付加情報を付け、プロキシサーバを介して閲覧者がその情報を利用する点。
（２）セマンティック・ウェブのように意味を表すタグ情報を用いて、文書の内容に関する機械的処理を実現する点。
（３）ＳＨＯＥのようにオンライン文書中に、語彙に関する付加情報を付け、それらを元にオントロジーを生成する手段を示している点。

（４）Topic Maps のように、オンライン文書同士を関連付ける手法を提示する点。
（５）Third Voice や BrowseUp に見られるようなＷｅｂコンテンツの部分に対して付加情報を関連付けることができる点。
（６）参加型のオンライン辞書作成等の動きにおける、一般の閲覧者が辞書作成に参加したり、ある特定の辞書に対しデータの誤りや指摘をするといったフィードバックをしている点。

そして、本研究には以上と比較して次の特徴が挙げられる。

（１）オントロジー生成といった一部の人にしかできないような高度な処理ではなく、誰にでも非常に簡単な操作で参加できる構造化された単語情報の構築が可能である点。
（２）閲覧者のフィードバックにより、単語情報の付加コストを引き下げることができる点。
（３）独自のウインドウに情報を表示せず、文書ブラウザと辞書を統合することで、オンライン文書の閲覧者は単語情報を自然に引き出すことができる点。
（４）集められた単語情報を元に、オンライン文書間の類似性を自動的に求めることが可能な点。
（５）関連システムと本研究におけるシステムを情報の作成者、情報作成のコスト、情報の詳細度、情報の再利用性の観点から表にすると以下のようになる。

システム名	情報の作成者	作成コスト	詳細度	再利用性
セマンティック・トランスコーディング	誰でも（コンテンツ作成者、第三者）	中	中	高
セマンティック・ウェブ	コンテンツ作成者	高	高	高
ＳＨＯＥ	コンテンツ作成者	高	高	高
Topic Maps	コンテンツ作成者	高	高	高
Third Voice	誰でも	低	低	低
BrowseUp	誰でも	低	低	低
ユーザ参加型辞書	誰でも	低	中	低
本システム	誰でも	低	中	高

第３章
語彙アノテーションと辞書の作成

本章では語彙アノテーションと辞書の作成について述べる。まずオンライン文書に含まれる任意の単語に語彙情報をアノテートする手順について述べ、次にそれら語彙情報のデータを利用し、辞書を作成する手順について述べる。

３．１　語彙アノテーション

オンライン文書に語彙情報をアノテートする場合、次のような問題点が考えられる。

（１）誰がどの単語を理解困難とするかが示されないため、アノテートする側にとってはどの語についてアノテーションをつけてよいかの指標がない。
（２）（１）の対処法として、なるべく多くの語にアノテーションを付与するやり方、またはアノテータが理解困難であると解釈した語に関してアノテートするやり方が考えられるが、前者の場合、大量にあるオンライン文書のことを考えると物理的に難しく、後者の場合、閲覧者の要望とのミスマッチが生じる可能性がある。
（３）多義語の場合、複数の語義をどのように区別してアノテートすべきか、一般には決められないことが多い。
（４）ある文書に関して定義した語義を他の文書において再利用するときに、異なる語に関するものを含めて、適切な候補を選び出すやり方は自明ではない。

これら上記問題点を解決するために、以下の手法、システムを考案した。

（１）オンライン文書内の分かりにくい語やフレーズを自由に登録し、共有できるシステムの構築
（２）（１）で登録された、分かりにくい部分に誰でも簡単に説明文を登録でき、その内容から共有可能な辞書を作成するシステム
（３）（２）で作成された辞書を用いて、文書を分かりやすく書き換えて表示するシステム

３．２　オンライン文書内の理解困難な語の共有システム

オンライン文書において、閲覧者が理解困難だと感じている語がどれであるかを知ることは、そ

のページの作成者にとっても、アノテートする人にとっても非常に重要なことである。前者には、よいフィードバックとなり、後者にはアノテートすべき個所がはっきりするため、アノテーションの手間を必要最小限にできる。また閲覧者に対し、彼らが分からないと登録した語がどのくらい、またどのようにアノテートされているかを示すページを用意することで、閲覧者はフィードバックを非同期に受けることができる。
本研究による理解困難な語の共有システムは以下の要素から構成される。
閲覧者、オンライン文書を閲覧するためのブラウザ、共有データを蓄積するための辞書サーバ、理解困難な語の登録用ユーザインタフェース、外部サーバとの通信用モジュール、閲覧中の文書の表示を変更する文書変換サーバである。

閲覧者が理解困難な語を登録する場合の手順は以下のようになる。

（1）閲覧者はオンライン文書を閲覧する際に、ブラウザ内にブラウザ上で選択した語を辞書サーバに送るためのユーザインタフェースと通信モジュールを読み込む。
たとえば、プロキシを介して閲覧中の文書に辞書サーバや文書変換サーバとの通信をするモジュールを埋め込んだり、理解困難な語登録用のメニューを追加する。(図 4)

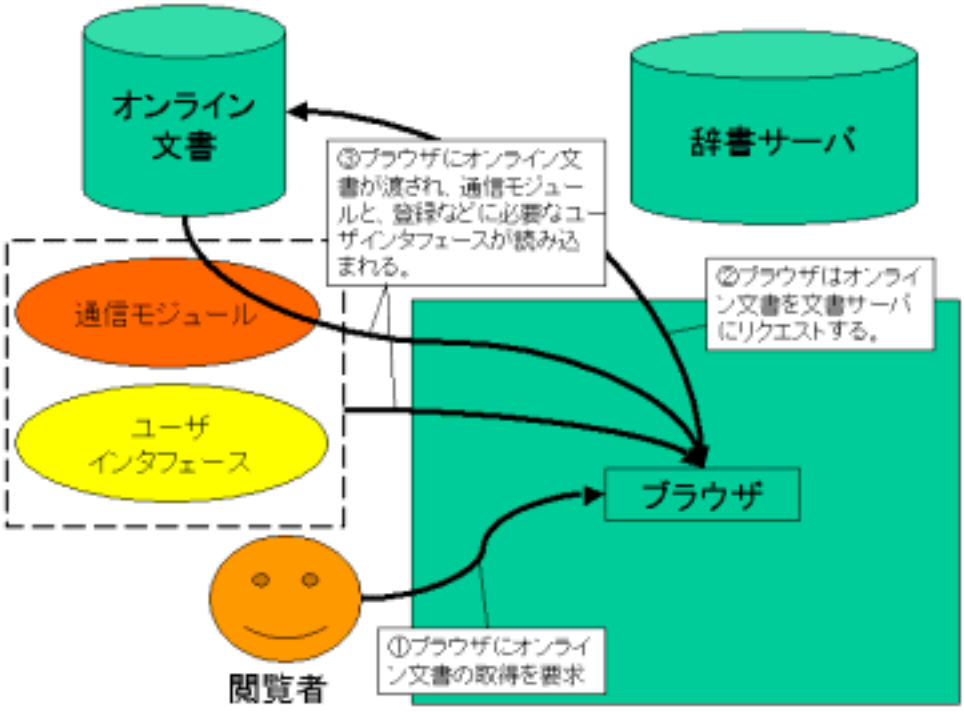

図 4　最初に文書を読み込むときの処理の流れ

（２）閲覧者は埋め込まれたユーザインタフェースと通信モジュールを利用し、オンライン文書中で自分が理解困難であるという語をマウスのドラッグ操作等で選択し、辞書サーバにそのデータを送信する。
辞書サーバは送信元のURLと送られてきた単語等のデータを組で保存する。もし、同じURLで同じ単語が複数回登録された場合、その回数を覚えておく。（図 5）

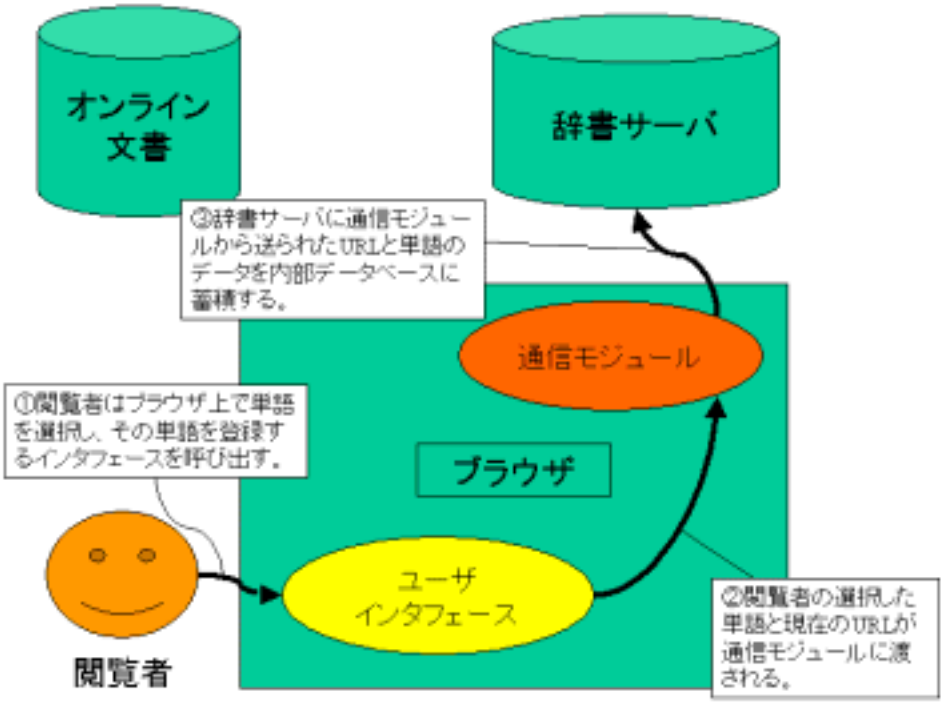

図 5

閲覧者が理解困難な単語を辞書サーバに登録するときの処理の流れ

（３）以後、オンライン文書の閲覧者は（１）において読み込まれたユーザインタフェースの機能を利用することにより、現在閲覧しているページのどの単語がどれだけ分かりにくいと登録されたかを、その単語が登録された回数を利用し、背景色の変化でその数を示す。この情報は、以後のアノテート等に利用される。
なお表示を変更するため、通信モジュールは表示変換サーバを介して辞書サーバにアクセスする。（図 6）

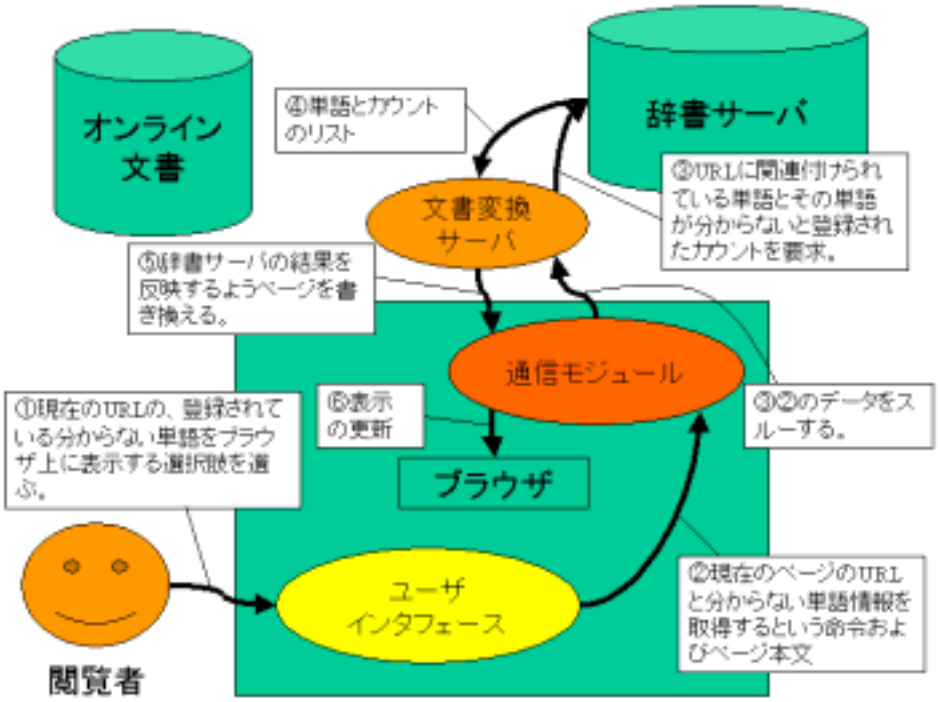

図 6

閲覧中のページにおける理解困難だとされている語を表示するときの処理の流れ

次に上記システムで登録されたデータを利用し、語義に関するアノテーション（理解困難な語に説明文等の登録）を行う手続きを示す。（図 7）

（1）閲覧者は読み込まれたインタフェースと通信モジュールを利用して、現在閲覧しているページのどの単語がどのくらい分かりにくいかを背景色等で知ることができる。アノテータは特定の単語にアノテートしたい場合、画面上の単語をマウスのドラッグ操作などにより選択し、ユーザインタフェース中のメニューの「説明文の登録」の項目を選択する。選択する単語は必ずしも、誰かが分からないと登録した語である必要はない。

（2）説明文入力用のフォームが画面上に現れ、アノテータは選択された単語の意味の説明文を入力し、通信モジュールを利用し、辞書サーバに送信する。

登録に関して、匿名によるノイズの増加を防ぐため、何らかのユーザ情報と一緒に登録される必要がある。

（3）辞書サーバは送信元のURLと単語、それに説明文、アノテータのユーザ情報、その単語を含む一文（例文として蓄えられる）を受け取り、それらを保存する。

（4）以後、オンライン文書の閲覧者は読み込まれたユーザインタフェースの諸機能を利用することにより、あるURLの特定の単語に関連付けられた説明文をさまざまな形式で閲覧することができる。次章で詳しく述べるが、たとえば、「ポップアップ」、「グロッサリー」、「挿入」、

「原文の言い換え」などが挙げられる。

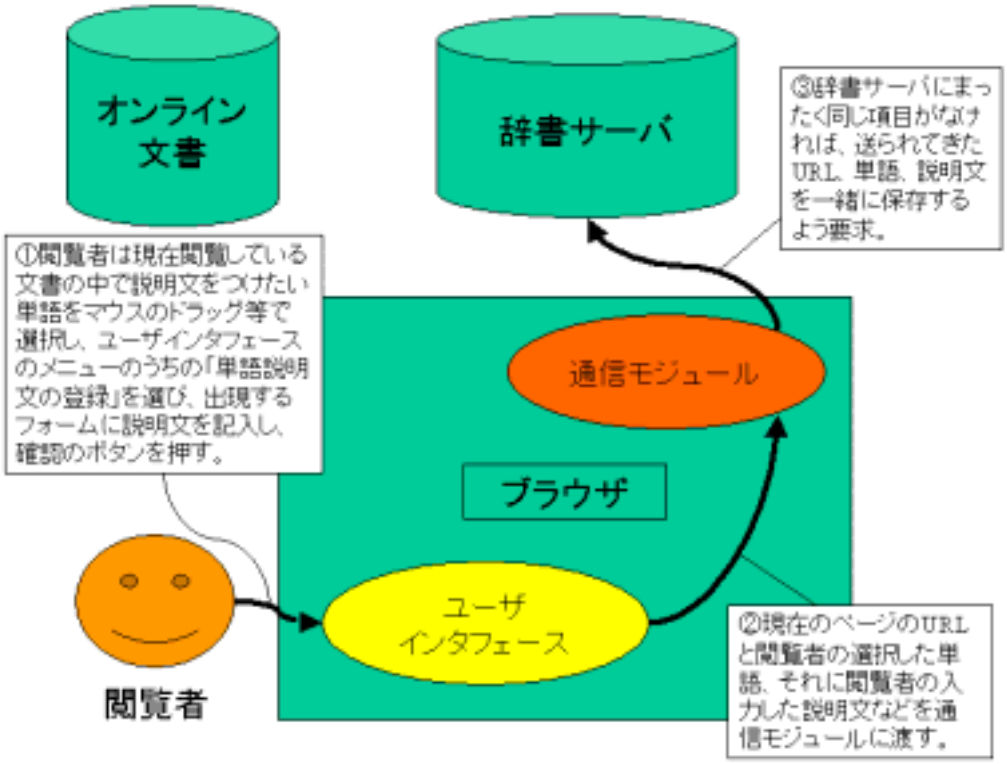

図 7

閲覧者が単語に説明文を登録する時の処理の流れ

上記システムによって蓄積されたデータは、いわばオンライン文書に関する共有の辞書であるといえる。次節ではこれらの蓄積されたデータを効率よく利用する仕組みとその応用に関して述べる。

３．３　ページ辞書の構築

Webの発達によって、多くの人で知を共有するプロジェクトがいくつか生まれている。代表的なものはWebを利用する多くの人の知識を集めて、共有の辞書を作ろうという試みであるが、そういった試みは、一般的にひとつの大きな辞書の構築を目的としている。しかし、Web上に存在する単語は多岐に渡り、それらをひとつの体系的な辞書として整理するには大変な労力が必要になる。本研究では、すべてを網羅する総合的な辞書を構築するのではなく、ある限定された範囲において使用される辞書を考え、その辞書を共有できる仕組みを提案する。

一般にそれぞれのWebページはあるひとつのテーマに基づいていることが多いことから、限られ

た範囲ではあるが、辞書の適用範囲（見出し語と語義）をWebページ内に限定し、辞書を構築する。この辞書を「ページ辞書」と呼ぶことにする。イメージとしてはWebページとリンクした小さな辞書といった感じである。

前述のように、理解困難な語とその説明文登録システムはURLとそれに関するデータの対で保存されている。ゆえに、辞書サーバからURLをキーにして取得した、そのページ中の単語のデータがそのまま「ページ辞書」となる。

ページ辞書の各構成要素は以下のようになっている。
（１）単語（見出し語）
（２）語に対する説明文とそれを登録したアノテータ情報
（３）理解困難であると登録された回数を示すカウンタ
（４）その単語の現れているページ内での一文（例文として使用される）

（１）（２）（３）は前節で述べた要素である。（４）はその単語がどのように使用されるかを後で参照するために用いられ、説明文が登録される際に自動で登録される。こういった例文を蓄積することは、このページ辞書がそのページのみに依存したものではなく、より一般的な辞書として利用価値をもつために必要なものである。

３．４　ページ辞書の拡張

上記のように作成されたページ辞書は次のように拡張される。

（１）同じようなトピックに関するオンライン文書間で複数のページ辞書が存在する場合、それらのページ辞書をマージし、より広い範囲を持つ辞書を生成できる。
（２）新たに、あるオンライン文書中のある単語に説明文を登録したい場合、もし以前に同じ単語が違うURLで登録されており、同じ意味を示すものが存在するならば、登録する単語の説明文として、他のページ辞書の単語を参照してその説明文を使用できる。

ページ辞書をマージする仕組みは以下のようになる。

ページ辞書はWebページごとに関連付けられた、見出し語とその説明文の対の集合であらわされる。それらのページ辞書はおのおの独立に作成されるが、あるWebページに対するページ辞書が他のWebページも利用可能な場合がある。たとえば、同じ分野のページ群は一般的に同じような語彙を利用すると考えられるので、それらのページ群うちどれか一つがページ辞書を持っていたならば、他のページに転用が可能である。

このように、ページ辞書を有効に利用するためには同じようなトピック、分野を持つページを見つけ、関連付ける必要がある。その関連付けの発見には一般的に以下の方法が考えられる。

（1）ユーザが手動でそれらを関連付ける。
（2）ページ内の単語の文字列マッチング処理などで計算された、ページの類似度などで分別する。

（1）や（2）は次の問題点がある。
手動でのURLの関連付けは非常にコストがかかり、費用対効果が小さい。機械的なマッチングでは精度に問題があり、必ずしも正しい情報が取得できるとは限らず、最終的には人間が介入し修正などの作業が必要となる。

これらの問題点を踏まえ、ページ辞書を利用した次のような解決案が考えられる。

（1）ページ辞書が同じ語義、または後述の関連性のある語義を含んでいたり、語義に関する関係が見つからない場合は、ページ辞書の見出し語や説明文に含まれる単語の共通部分が大きいものを含むページを1つのページ群とする。
（2）同じアノテータがアノテートしたページ群は似たような分類であると仮定する。
（3）ページ辞書を見比べ、アノテータが手動で関連付けることもできる。
（4）ページ群を登録し、それに名前を付け、それらページ群の閲覧、分解、再構成等ができるようにする。

関連付けられたページ群のページ辞書がマージすることにより、専門用語辞書のような、より一般化された辞書としての利用も可能となる。

また、ページ辞書に含まれる語義（Webページ中の単語と説明文の組のこと）は必ずしも適切とは言えない場合も考慮し、語義に対して閲覧者が評価を行うことができる。たとえば、表示された説明文の脇に評価用インタフェースを用意し、閲覧者は5段階などでその説明文を評価し、その情報を共有サーバに送る。また、その情報はその説明文を付けた人に通知される。

この情報はページ辞書を常に適正なレベルに保つとともに、閲覧者にページ辞書の内容を提示する際の順序付けなどに利用される。

また、以下のような仕組みで語義と語義との関連性を作成、提示する。
複数のページ辞書を並べて表示し、その中で示される語義の内で関連付けを行うことができ、作業として、それぞれのページ辞書に対応するWebページを見ながら、ブラウザ上で行ったり、専

用のツールを利用する。これらの関連付けはページ辞書の管理者か任命された特定のアノテータが行うことにする。この語義間の関連性は、次章で示すように、Webページの類似度を計算し、自動的にコンテンツ統合を行うために利用される。

基本的には、2つの語義の関係をアノテータが決定し、その結果をテーブルに保存する。関係テーブルは以下のような構成になる。語義はあるURLにおける単語と説明文の対のことである。

（1） 語義1
（2） 語義2
（3） 語義1の、語義2に対する関係
　　（上位概念、下位概念、類義語、反意語、派生語）

次章では、本章で述べた仕組みによって作成されたページ辞書の応用について述べる。

第４章
辞書を用いたオンライン文書の加工

本章では、Webページをオンライン文書の代表例とみなし、前章で構築したページ辞書を利用した文書の加工について述べる。具体的には以下のことを実現する。

（１） ページ辞書とWebページを統合して、分かりにくい語の情報をユーザに閲覧可能にする。このとき、ページ辞書の内容の表示法としては、ポップアップ、グロッサリー、挿入、原文の書き換えがある。
（２） ページ辞書からWebページ間の類似度を計算して、複数のWebページを自動分類し、特定のトピックに関するページを統合して、複数ページのサマリーを生成する。
（３） ページ辞書から分野ごとの用語辞書を作成し、独立のコンテンツとする。

また、本章ではページ辞書を利用したWebページ間の類似度の計算方法についても説明する。

４．１　ページ辞書と Web ページの統合

ページ辞書とWebページを統合して、ブラウザ上で辞書内容を表示するプロセスは以下のようになる。（図 8）

（１）閲覧者は読み込まれたユーザインタフェースと通信モジュールを利用し、現在閲覧中のページに関連付けられたページ辞書を以下の方法で閲覧するようにメニューを選択し決定する。閲覧方法は以下の中から選ばれる。

①ポップアップ
閲覧中のページにおいて、ページ辞書に見出し語のある単語の上にマウスカーソルが置かれた時にその定義文を、ポップアップウインドウを開きその中に表示する。

②挿入
閲覧中のページにおいて、ページ辞書に見出し語のある単語について、定義文をその単語の直後に、括弧付けで表示する。

③グロッサリー
閲覧中のページに関連付けられたページ辞書の見出し語と説明文の対を一覧表にし、新たにウイ

ンドウを開いた中に表示する。

④原文の言い換え
原文の言い換えには2つの種類がある。1つは閲覧者の個人情報などを利用し、その人とってもっとも分かりやすいと思われるように原文の全体を書き換える言い換え、もう１つは閲覧者のインタラクション（マウスのクリックなど）に応じて適宜、部分部分、原文を言い換えるものである。前者を全文の言い換えと呼び、後者をインタラクティブパラフレーズと呼ぶ。

言い換えは、もともと言語処理的に非常に困難な作業である。そのため、原文の言い換えを行う場合、辞書サーバに登録されている単語や説明文、それに書き換えの対象となる文書についても同じく、統語的アノテーションを持つ必要がある。

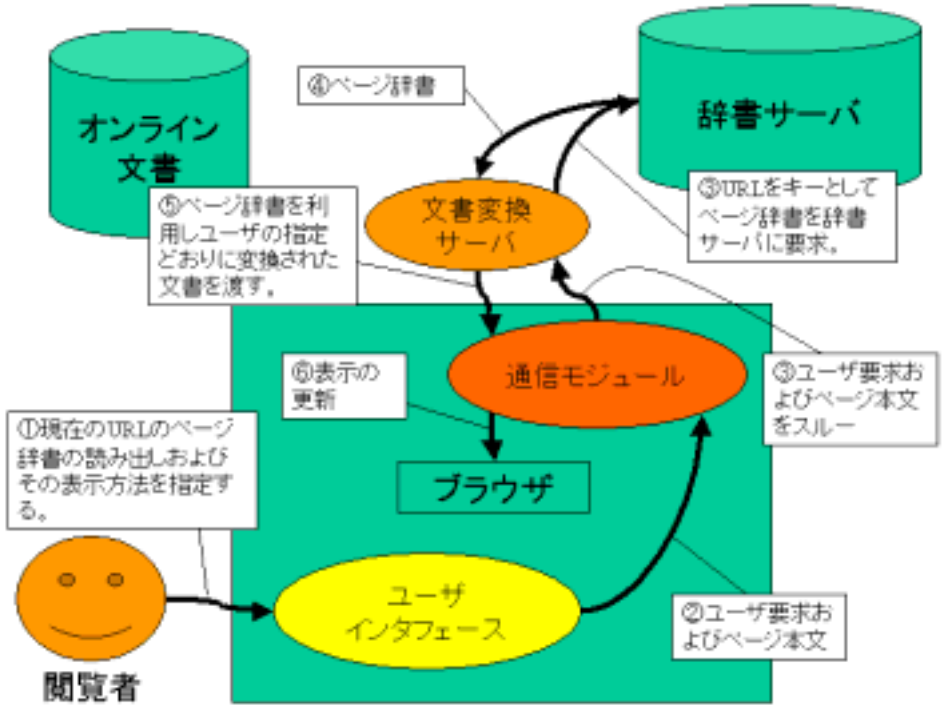

図 8
ページ辞書の情報を閲覧者に表示する
（ポップアップ表示、説明文の挿入、グロッサリー表示）ときの処理の流れ

なお、原文の言い換えの処理の流れは他の3つと異なる。言い換えについての詳細、原文の言い換えの処理の流れ、インタラクティブパラフレーズの説明は後節で述べる。

（２）以上の言い換えを除く3つの表示方法のどれかが選ばれると、その情報と閲覧中のページ

本文は通信モジュールを通じて文書変換サーバに渡される。
文書変換サーバは辞書サーバにアクセスし、ページ辞書を取得した後、閲覧者の要求に応じて、渡された本文を書き換える（ポップアップの挿入、挿入文の追加、グロッサリーウインドウの追加）。

通信モジュールに変換された文書を渡し、通信モジュールはブラウザにその文書を表示させる。

４．２　言い換え

前節で述べたように、言い換えは多くの言語的操作が必要とし機械的実現が非常に困難である。たとえば、単語をその説明文にそのまま置き換えたのでは全く文との親和性がない。これは当然で、辞書の定義文はその単語の説明であって類義語などの「等価な表現」ではないからである。

ゆえに定義文を原文に埋め込む際、定義文を適切な形に変換する必要性が生じる。これは言い換えであり、過去にいくつか研究がなされている。佐藤らは、複合名詞の言い換え[10]や、サ変名詞の言い換え[7]、格変換による言い換え[8]を提案し実装している。彼らは言い換えを以下の3つのクラスに分類している。

（１）構文的言い換え
言葉に関する知識によって実現可能な言い換え。単語を同義語や類義語に置き換える言い換えや、構造のマッピングに基づく言い換えをこのクラスに分類する。
（２）意味的言い換え
参照表現などを、それが指す内容で置き換える言い換えがこのクラスに含まれる。また、省略されているものを意味的に補う言い換えもこのクラスに含める。
（３）語用論的言い換え
（１）（２）以外のより複雑な言い換え。ある状況において同じ効果を持つような文に言い換えるものがこれに含まれる。

文の統語情報に関する付加情報（統語的アノテーション）を利用することにより以上のクラスの（１）と（２）が実現できる。統語的アノテーションには、係り受け、品詞、活用、照応（指示詞の参照）の情報が含まれる。

今回統語的アノテーションを持つテキストと統語的アノテーションを持つ辞書を利用することによって、一般的な言い換えを実現するルールを試作した。今回試作したルールは日本語の文法依存であるが、適応するルールを変更することにより、特定の文法理論に依存しない言い換えのシステムが実現できる[13]。

４．２．１　言い換えルール

ここでの言い換えは、言い換えられる単語と、それを言い換える定義文の置き換えとして実現される。その両者は、統語的アノテーションによりXML形式のタグ要素（以後エレメントと呼ぶ）として扱われる。なお、使用するタグはJUMAN[14]とKNP[12]によって得られた結果を人手で係り受けを修正しGDAタグに変換したものであり、品詞体系は茶筅[11]に合わせたものである。
たとえば以下は、「もみ合い」という名詞エレメントが「売りと買いが両方あり小幅の値動きを繰り返す状態」という定義文エレメントで言い換えられる際のそれぞれのエレメントの例である。

- 言い換えられるエレメントの例

```
<n 読み="モミアイ" 基本形="もみ合う" 品詞="動詞-自立" 活用="五段・ワ行促音便-連用形" sense="momiai">もみ合い</n>
```

- 言い換えるエレメントの例

```
<su id=”momiai”><vp><vp><adp><np syn="p"><np 読み="ウリ" 基本形="売る" 品詞="動詞-自立" 活用="五段・ラ行-連用形">売り</np><ad 読み="ト" 品詞="助詞-格助詞-一般">と</ad><np 読み="カイ" 基本形="買う" 品詞="動詞-自立" 活用="五段・ワ行促音便-連用形">買い</np></np><ad 読み="ガ" 品詞="助詞-格助詞-一般">が</ad></adp><np 読み="リョウホウ" 品詞="名詞-一般">両方</np><v 読み="アリ" 基本形="ある" 品詞="動詞-自立" 活用="五段・ラ行-連用形">あり</v>、</vp><adp><np><adp><np 読み="コハバ" 基本形="小幅" 品詞="名詞-形容動詞語幹" 活用="五段･ラ行-連用形">小幅</np><ad 読み="ノ" 品詞="助詞-連体化">の</ad></adp><np>値</np><n>動き</n></np><ad 読み="ヲ" 品詞="助詞-格助詞-一般">を</ad></adp><v 読み="クリカエス" 基本形="繰り返す" 品詞="動詞-自立" 活用="五段・サ行-基本形">繰り返す</v></vp><n 読み="ジョウタイ" 品詞="名詞-一般">状態</n></su>
```

具体的には原文テキストの各アノテーションエレメントが語義に関する情報を持っているとき、その情報に基づいて辞書から定義文エレメントを取得し、それら2つのエレメントについて変換ルールを適応する。ルールには2種類考えられる。1つはどのようなエレメントの間の言い換えでも適用されるグローバルルールと、エレメントのタグ、属性によって適用するルールが違うローカルルールである。

以下に今回試作した言い換えのルールを示す。なお、EとはEntryのことで置き換えられる対象のエレメントを指しDはDefinitionのことで定義文のエレメントを指す。

グローバルルール

（１）DにEが含まれている場合は言い換えない
（理由）語の説明に、説明する語が含まれている場合明らかにより簡単な表現になっていないため。また、辞書としてそういった記述は不自然なため。
（例）日本－日本という国。

（２）括弧付きで補足されているテキストエレメントは言い換えない。
（理由）定義文に括弧付きで説明されている部分は補足的であり、そういった情報はすでに原文において説明されていることが多いため。
（例）信託する－（責任や任務を）信用して委託する

（３）言い換え後２重否定になるようなものは言い換えない。
（理由）置き換え後、原文よりも複雑な構造になると考えられるため。
（例）「無視してはならない」の「無視」が「注意を向けない」という定義文を持つ場合、ないが重なるので言い換えない。
（×注意を向けないのではならない）

（４）Eにかかっている格助詞で、Dにかかっているものと同じ物があればその格助詞に係るものは削除する。
（理由）定義文ではどういった文脈でその語が用いられているかを示すために[～などを]とか「～などが」のような格助詞句を持つ場合が多い。しかし、そのような句は原文の中にすでに存在する場合が多く言い換え後冗長になったり、意味が分かりにくくなるため。
（例）「圧迫と偏狭を除去する」の「除去する」を「不要なものを取り除く」という定義文で言い換える場合「不要なものを」は言い換えに使わない。
（×圧迫と偏狭を、不要なものを取り除く）

（５）Dでシチュエーションの例示を行う句があればその部分は言い換えない。
（理由）（４）と似ているが、語の使われる文脈を示すために「において」や「において」などが使われることがある。これらも原文中にすでに自明である場合が多いため言い換えに使用しない方がよい。
（例）[成果－行為などの結果として生じたよい事柄]のうち「行為などの結果として」は言い換えに使わない。

ローカルルール

ルールに使われる記号は以下の意味で使用される。

N ::= n(名詞) | np(名詞句)
V ::= v(動詞) | vp(動詞句)
AD ::= ad(副詞・形容動詞・名詞) | adp(副詞句・形容動詞・名詞句)
AJ ::= aj(形容詞) | ajp(形容詞句)
なお、助詞はadとし、助詞を主辞に持つ句はadpとする。

ルール表に示されるN-Nなどは置き換えられるエレメント(E)がNで置き換えるエレメント(D)がNであるというような置き換えの対を示す。以下ルール表。

N－N	(1) そのまま置き換える (例)惨禍→わざわい、恵沢→恵み 複合名詞の場合は置き換えてしまうとつながりが悪くなり意味が取りづらくなる。その為に辞書引きによる置き換えは行わず、名詞間のつながりが分かるように助詞を挿入することが有効である（表題を言い換えるの論文）。 （１）N1-N2→N1のN2といいかえる。 日本国民→日本の国民 （２）N1-N2-N3 N2がサ変名詞の場合、N1をN2するN3と言い換える。
N－V	(1)Dを連体化し、「こと」をつけたもので置き換える。 (例)協和(心を合わせ仲良くする)→心を合わせ仲良くすること
V－N	(1)Dに係るvpがあればそのvpをEの活用で置き換える。
V－V	(1)Eの活用をDの活用に適応する。 (例)確認する→確かに認める。 (2)Dの表層格が、Eに係るADPの格を持たなければ置き換えない。 サ変名詞＋するの場合、ヴォイスやアスペクトなども考慮する必要がある[7]。
AD－N	(1)Dに係っているADを活用させ置き換える。 (2)Dに係っているADがない場合は置き換えない。
AJ－N	(1)Dに係っているAJを活用させ置き換える。 (2)Dに係っているAJがない場合は置き換えない。

４．２．２　全文の言い換え

原文の言い換えは全文の言い換えとインタラクティブパラフレーズの2つに分類されると述べた。まず前者の流れを述べる。（図 9）

（１）閲覧者は全文の言い換えをユーザインタフェースで選択する。
（２）通信モジュール、文書変換サーバへと要求が渡り、文書変換サーバは辞書サーバに閲覧中のURLに関連付けられたページ辞書で語義・統語的アノテーションを持つものを要求し、取得する。ここで、語義・統語的アノテーションとは、統語的アノテーションに語義情報を加えたものである。
（３）文書変換サーバは閲覧中のURLに関連付けられた語義・統語的アノテーションが存在すれば、それらのアノテーションを取得する。
（４）文書変換サーバは閲覧者の以前の言い換え履歴等を含むユーザプロファイルを要求、取得する。
（５）語義・統語的アノテーションを持つ定義文、文書をもとに、文書変換サーバは語義アノテーションのついているすべての単語について言い換えを行う。
（６）文書変換サーバは言い換え語の文書を通信モジュールに渡し、ブラウザに更新された文書が表示される。

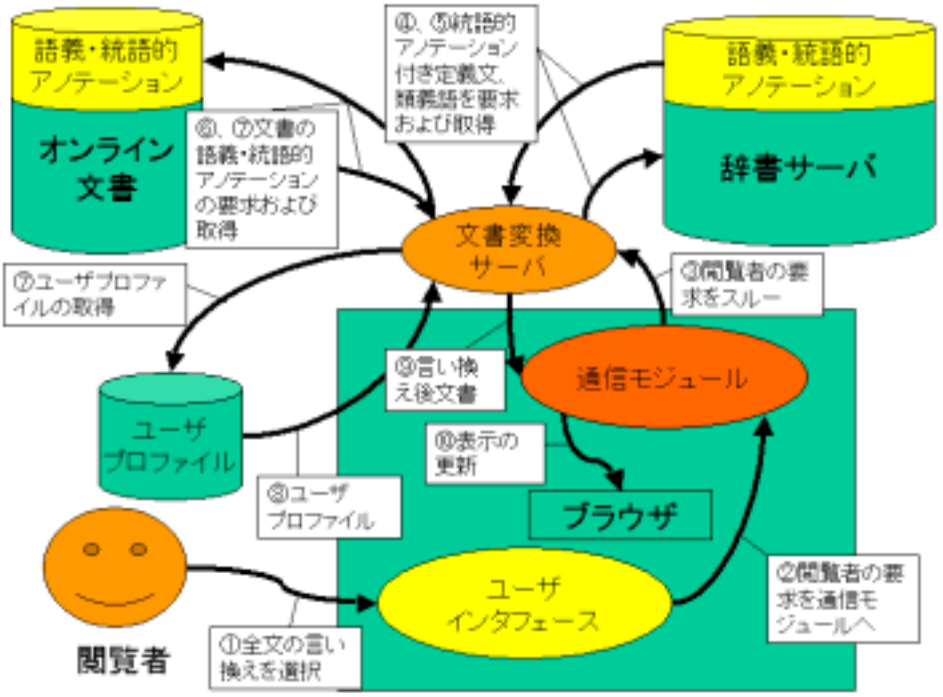

図 9
全文の言い換えを行うときの処理の流れ

４．２．３　インタラクティブパラフレーズ

前述した全文の言い換えには以下の点が問題となりうる。

（１）文書中の難解な単語すべてについて言い換え、ユーザの意図を反映していない。
（２）難解な単語すべてをその定義文に従って言い換えているため、冗長性が増し、逆に理解しにくくなる場合がある。
（３）言い換え後の表現にさらに難解な単語が存在した場合、それ以上言い換えることができない。

もちろん、ページ辞書に含まれる説明文が非常に良質なものであり、言い換えに適切なものであれば、全文の言い換えを行ってもよい結果が得られると思われるが、必ずしもページ辞書に登録される定義文は言い換えを前提として登録されているわけではないため、上記問題点は考慮されなければならない。そこで以下の手段でも言い換えが可能である。

（１）単語の言い換えをユーザの選択というインタラクションに従って行えるようにする。（セレクティブなパラフレーズ）
（２）言い換え後の単語についても言い換えを可能にする。（インクリメンタルなパラフレーズ）
（３）ユーザの操作履歴を保存して、最近言い換えたものを直後に再び言い換えないようにする。

以上の言い換えをインタラクティブパラフレーズと呼ぶ。なお言い換えに定義文のみを用いるのではなく同義語（関連付けられたものがあれば）も言い換え候補として利用するので、閲覧者はより広い選択の幅を持つことができる。

処理の流れは2段階ある。
閲覧者がインタラクティブパラフレーズをリクエストし、インタラクティブパラフレーズ用に文書が変更されるまでが第1段階、閲覧者がブラウザ上で分からない単語を選択し、言い換えをリクエストし結果を受け取るまでを第2段階とする。それぞれの流れを示す。

第1段階の流れは以下のようになる（図 10）。

（１）ユーザは読み込まれたユーザインタフェースを用いて、インタラクティブパラフレーズを選択する。
（２）通信モジュール、文書変換サーバにその要求が渡り、文書変換サーバは文書サーバにURLをキーとして文書の語義・統語的アノテーションを取得し、語義の付けられている単語については背景色を変更するよう文書を更新し、その部分について以後ユーザのインタラクションを受け

られるようにする。
（3）インタラクティブパラフレーズ用の文書がブラウザ上に表示される。（ユーザのインタラクション待ちの状態になる。）

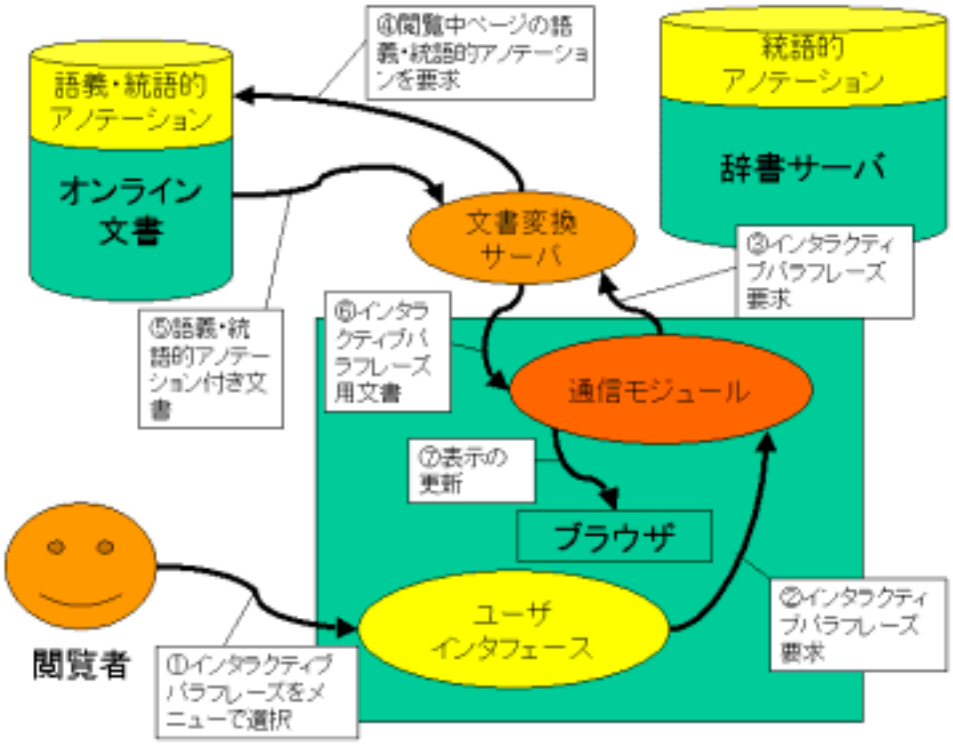

図 10
インタラクティブパラフレーズの準備を行う処理の流れ

第2段階の処理の流れは以下のようになる（図 11）。

（1）閲覧者は単語あるいはテキスト領域を画面上で特定する。
（2）通信モジュール、文書変換サーバがその要求を受け付け、閲覧者が選択した単語あるいは領域に含まれる（複数の）単語について辞書サーバに問い合わせをする。辞書サーバは定義文のリストと、関連付けられたものがあれば類義語のリストを返す。
（3）文書変換サーバは閲覧者のプロファイルを要求する。このプロファイルには過去においてその閲覧者がどのような言い換えをおこなったかなどの履歴情報が含まれているテーブルである。
（4）文書変換サーバはユーザプロファイルを参照し、言い換えに使用する単語を選別する。ただし、ユーザが一つの単語のみを選択した場合は、プロファイルに関わらず言い換える。
（5）アノテーションを持つ類義語リストと定義文を言い換え、結果を通信モジュールに返すと

同時にユーザプロファイルに言い換えた元の単語を記録する。
（６）通信モジュールは変更された文書をブラウザに表示する。

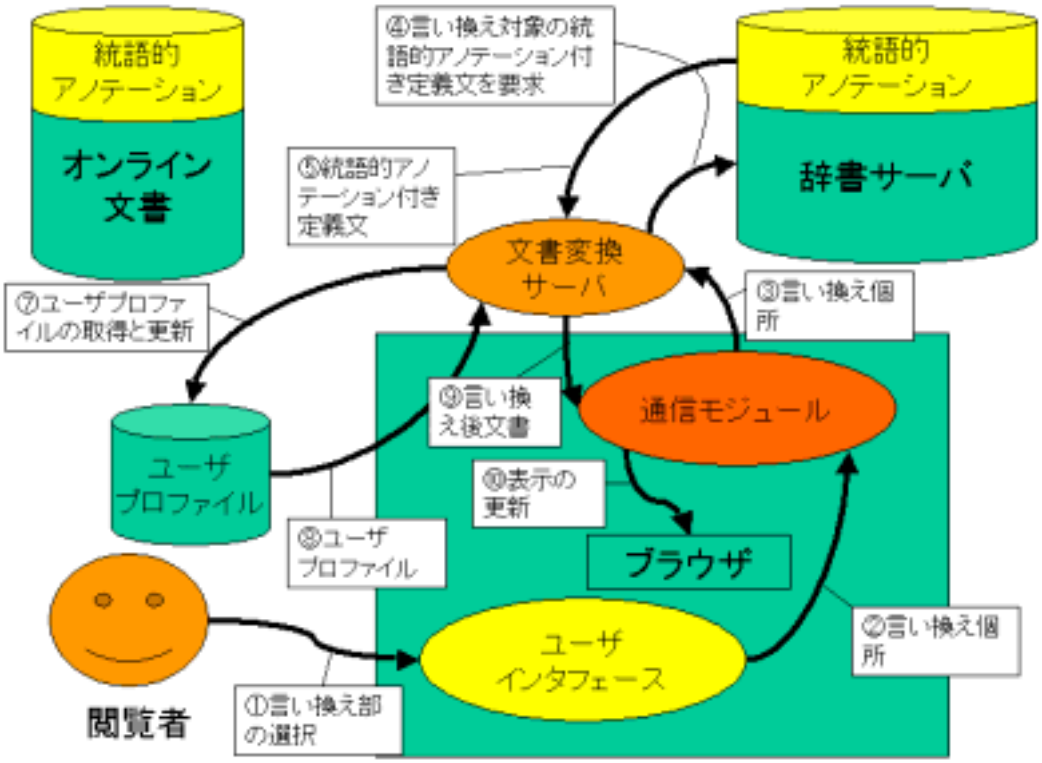

図 11
インタラクティブパラフレーズの実際の処理の流れ

インタラクティブパラフレーズはユーザのインタラクションに応じて原文の書き換えを行うものであるが、そのインタラクションの手法にも次の3つの分類が考えられる。

（１）クリックパラフレーズ
（２）リストパラフレーズ
（３）リージョンパラフレーズ

以下それぞれについて説明する。

①クリックパラフレーズ
マウスで分からない単語エレメントをユーザがクリックすると、その単語の類義語または定義文を利用して言い換える。分からない語をクリックするという自然な動作で言い換えを行うことができる。

②リストパラフレーズ
マウスで分からない単語エレメントをユーザがクリックすると言い換え可能な単語の一覧がリスト表示され、ユーザがそのうちの一つを選択するとその候補を利用し言い換える。言い換え語の候補があらかじめ一覧できるので、繰り返し言い換えする手間を省くことができる。

③リージョンパラフレーズ
ユーザがマウスのドラッグ操作により、文書内の領域を選択し、それに含まれるすべての単語に対しユーザ履歴を参照して言い換える単語を選別し、言い換えを行う。一度に複数の言い換え候補を選択することができるため、ユーザの手間を減らすことができる。

４．３　Web ページ間類似度の計算

ページ辞書に基づくWebページ間の類似度の計算法は以下のようになる。
ここでは簡単のために2つのWebページ間での計算手法を提示する。

対象とするページ辞書は片方を $PD1$、もう一方を $PD2$ とし、3つの2次元テーブルを用意する。それぞれ、見出し語類似テーブル、語義類似テーブル、説明文類似テーブルと呼ばれ、順に I (Index)，C (Concept)，D (Definition)で表される。

見出し語類似テーブルは、2つのページ辞書間の見出し語の類似度を示すもので $PD1$ と $PD2$ について、Iij は以下のような値を持つ。

$$PD1(i) = PD2(j) \text{ のとき } Iij = 1$$
$$\text{それ以外 } \quad Iij = 0$$

語義類似テーブルは、2つのページ辞書間の、語義の類似度を示すもので、$PD1$ と $PD2$ について、Cij には $PD1$ の i 番目、$PD2$ の j 番目の語義の類似度を示す数値が入る。語義の類似度は関連付けツールなどを使って人がつけた、上位・下位関係、派生語関係、類義語関係などを数値化したものがはいる。もし $PD1i$ と $PD2j$ が語義に関連をもつなら、Cij はその関連に応じた数値、関連が付けられてないなら0となる。人手で関連をつけるため、情報に信憑性が高く、語義類似テーブルは他のテーブルに比べ重きをおかれるべきである。

説明文類似テーブルは、2つのページ辞書間の、説明文の類似度を示すものである。説明文は形態素解析されており、不必要な助詞などの情報が除かれ、活用形のあるものは原形に合わせられている。$PD1$ と $PD2$ において、Dij には $PD1$ の i 番目、$PD2$ の j 番目の説明文の要素の文字列マッチングを行い値は以下のようになる。

$$PD1(i) = PD2(j) \text{ のとき } Dij = 1$$
$$\text{それ以外} \quad Dij = 0$$

単語の類似、語義の類似、説明文の類似はそれぞれページ間の関連について持つ重要度が異なる。一般的に語義の類似、単語の類似、説明文の類似の順で重要であることが多いと考えられる。よって、ページ間の類似度は以下の式で求められる。

$$S(similarity) = A\sum_{i=1}^{n}\sum_{j=1}^{m} Iij + B\sum_{i=1}^{n}\sum_{j=1}^{m} Cij + C\sum_{i=1}^{n}\sum_{j=1}^{m} Dij$$
$$(B \geq A \geq C)$$

上記式により求められたページ間の類似度には次のような応用が考えられる。

（1） 関連するページ群の発見
（2） 関連するページ群の内容をマージすることによる知識統合

第５章
実装

５．１　実装環境

第３章と第４章で述べたシステムの実装環境について述べる。
図 12 は本システムの構成要素を示す。
以下に示す表では、構成要素それぞれが何を用いて実装されているかを示す。

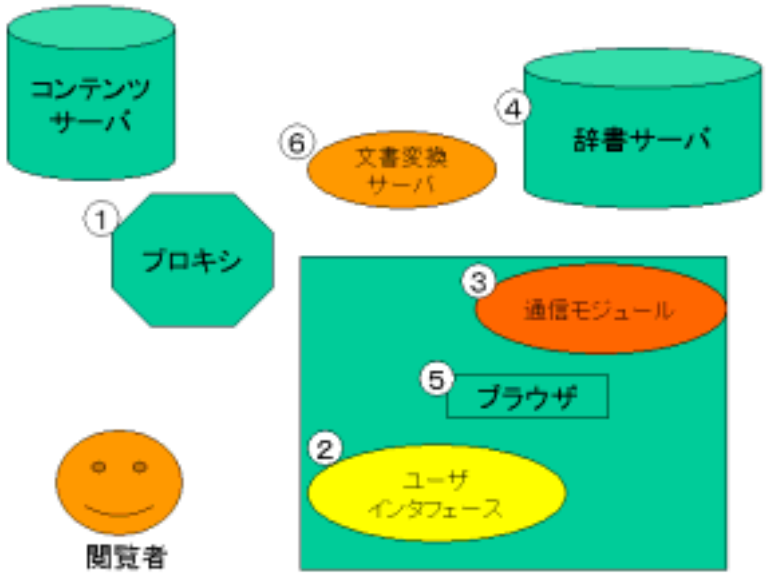

図 12
システム構成要素

①プロキシ	IBM Websphere Transcoding Publisher (http://www.jp.ibm.com/software/websphere/)
②ユーザインタフェース	JavaScript によるメニュー（図 13 参照）
③通信モジュール	通信機能を持つ JavaApplet
④辞書サーバ	IBM DB2 Enterprise Edition (http://www.jp.ibm.com/software/data/udb/)
⑤ブラウザ	Microsoft Internet Explorer 5.0
⑥文書変換サーバ	Java アプリケーション(JDK1.3)

図 13

ユーザインターフェースとして読み込まれるメニュー

５．２　実行画面例

前節で示したメニューを操作することによって、閲覧中の文書に対して様々な処理を行うことができる。ここでは、それらの処理を画面例を通して説明する。

５．２．１　初期画面

図 14 は閲覧文書の例であり、文書が最初に読み込まれたときの様子である。
以後、この文書に対して処理を行うこととする。
見出しの段落が拡大されているが、この領域を対象として説明を進める。
文書として日経新聞のウェブサイト(http://www.nikkei.co.jp)のマーケット欄の記事を使用した。

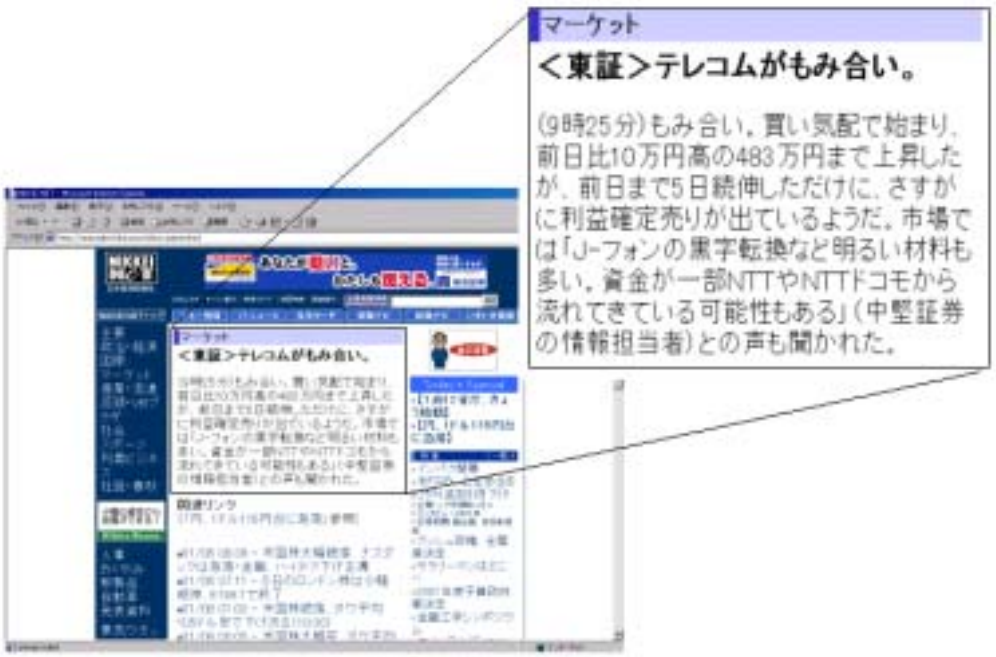

図 14　閲覧文書が最初に読み込まれたときの状態

５．２．２　理解困難語の登録

閲覧者が理解困難である単語を登録する際はメニューより「register as difficult」を選択する。(図 15)

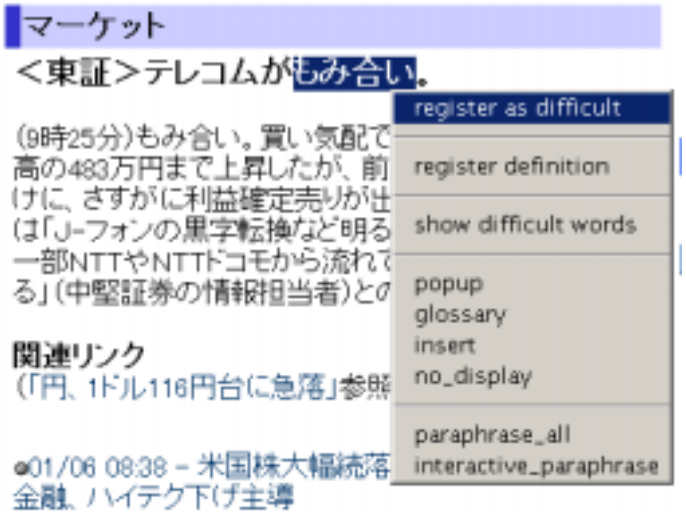

図 15　理解困難語の登録

閲覧者が理解困難だと登録した語がどれであるかはメニューより「show difficult words」を選択すると、理解困難だとされた語の背景色が赤く変化する。理解困難であると多くの人に登録された単語は背景色がより赤く表示される。（図 16）

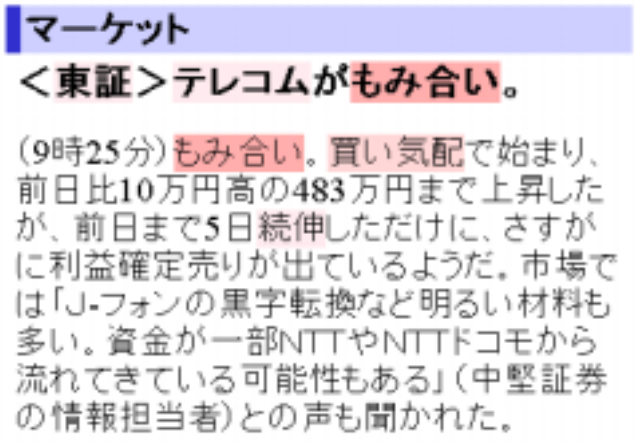

図 16

理解困難な語の表示

５．２．３　説明文の登録

単語に説明文をつけるにはメニューより「register definition」を選択する。すると単語入力フォームが出現するので、入力エリアに説明文を入力し、ＯＫを押すと辞書サーバの方に送られる。（図 17、図 18）

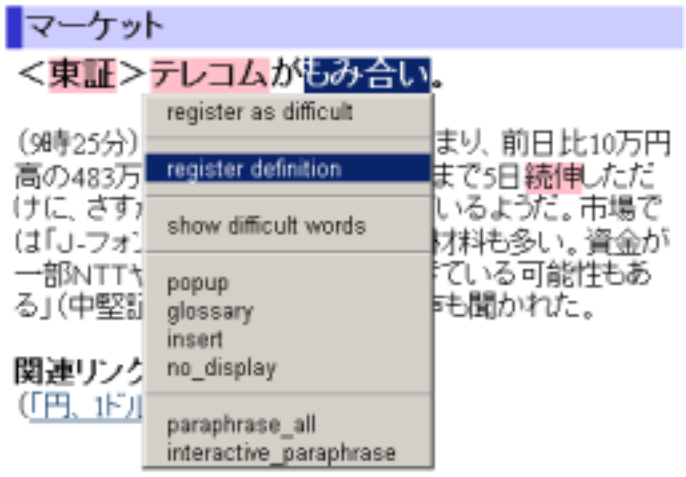

図 17

説明文の入力を選択する様子

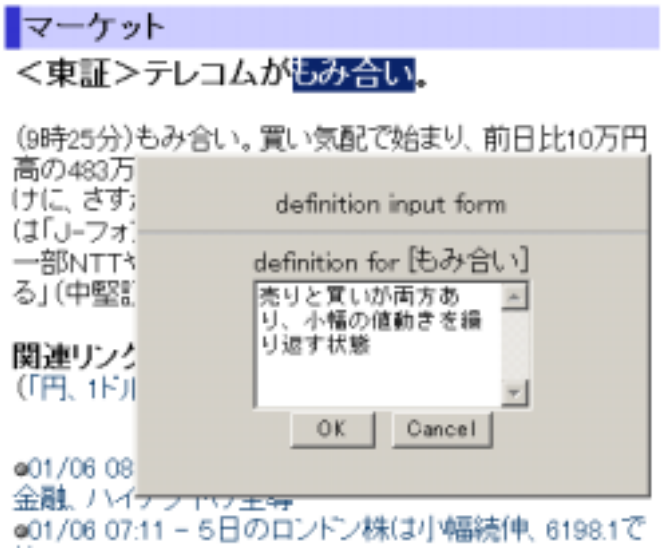

図 18

説明文の入力フォームに入力する様子

５．２．４　ポップアップ

メニューで「popup」を選択すると、閲覧文書中の単語で説明文がアノテートされているものに対しポップアップを挿入して表示する。（図 19、図 20）

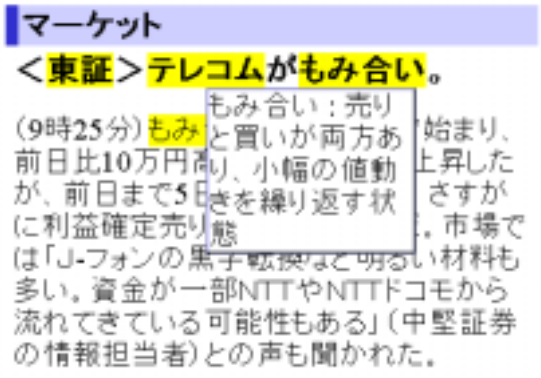

図 19

「もみ合い」の説明文のポップアップ表示

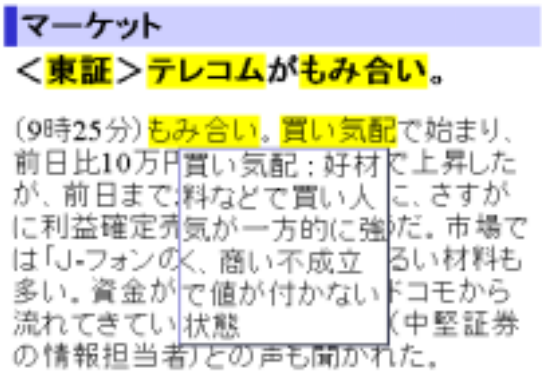

図 20

「買い気配」の説明文のポップアップ表示

５．２．５　挿入

メニューで「insert」を選択すると、閲覧文書中の単語で説明文がアノテートされているものに

対し説明文を括弧付けで挿入して表示する。（図 21）

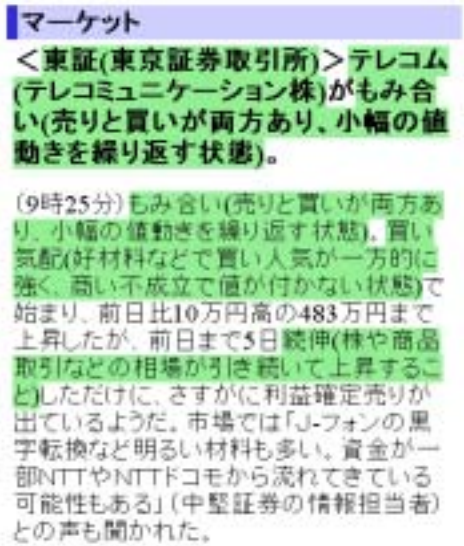

図 21 説明文の挿入

５．２．６　グロッサリー

メニューで「glossary」を選択すると、閲覧文書中の単語で説明文がアノテートされているものに対し説明文をグロッサリーウインドウに表示する。（図 22）

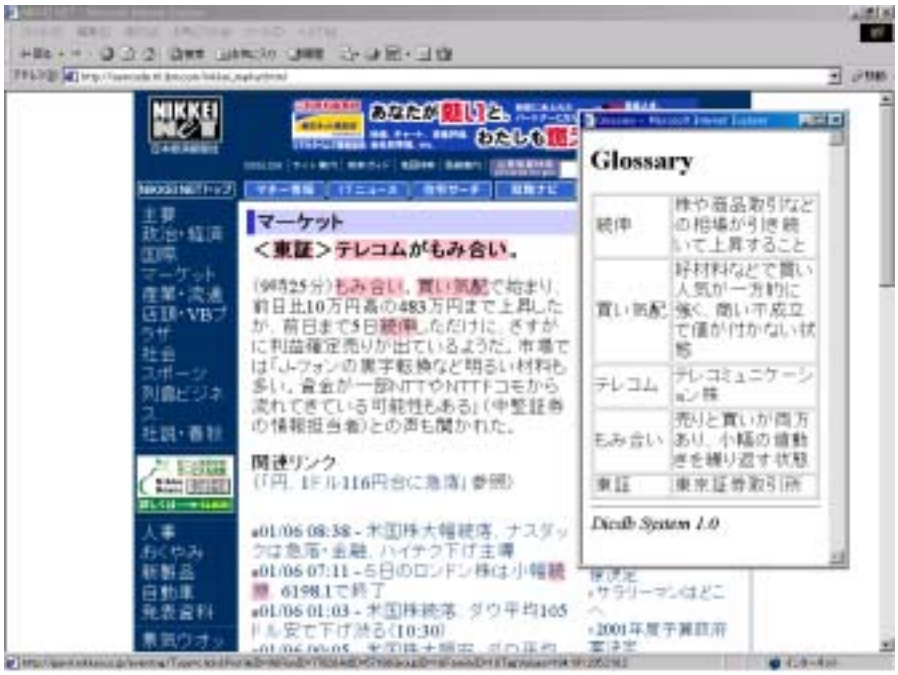

図 22 グロッサリーウインドウの表示

５．２．７　全文の言い換え

メニューで「paraphrase_all」を選択すると、閲覧文書中の単語で説明文がアノテートされているものすべてに対し言い換えを実行する。（図 23）

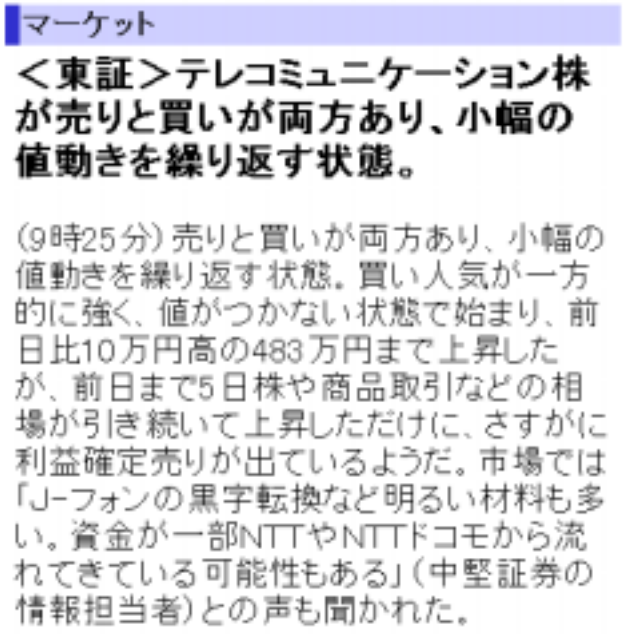

図 23　全文の言い換え

５．２．８　インタラクティブパラフレーズ

メニューで「interactive_paraphrase」を選択すると、インタラクティブパラフレーズモードに切り替わる。（図 24）

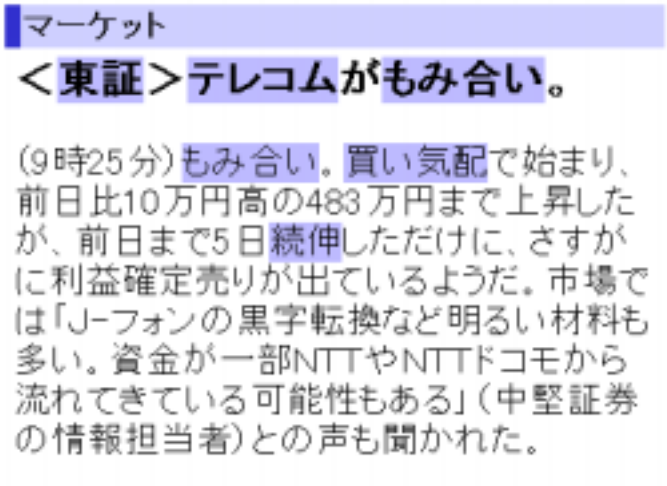

図 24　インタラクティブパラフレーズモード（青い個所が言い換え可能）

インタラクティブパラフレーズモードでは言い換えは「クリックパラフレーズ」、「リストパラフレーズ」、「リージョンパラフレーズ」の３種類が可能である。それぞれについて説明する。

（１）クリックパラフレーズ

言い換え可能であると表示された個所をクリックすると、段階的に言い換えが可能である。（図 25、図 26）

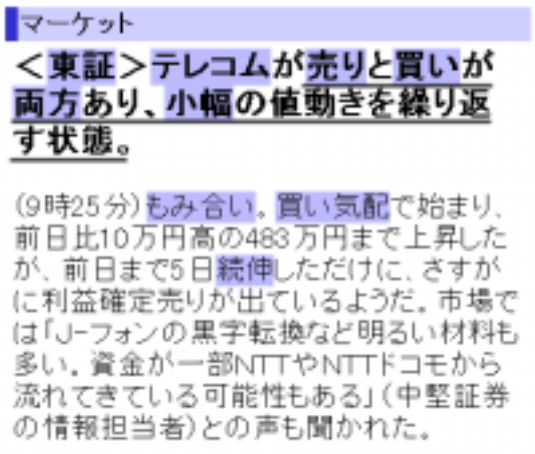

図 25　「もみ合い」を言い換えた場合

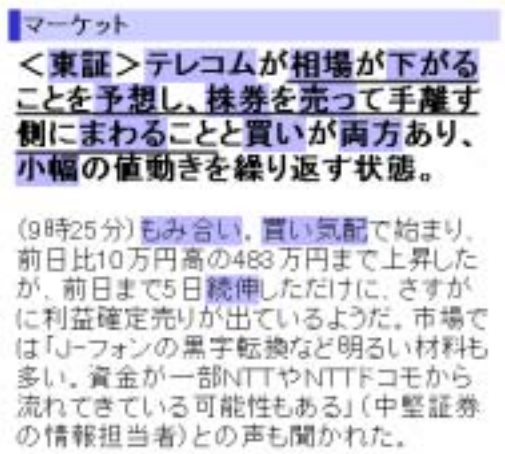

図 26　さらに、「売り」を言い換えた場合

（２）リストパラフレーズ

言い換え候補のリストを表示させ、その中から選択することもできる。（図 27）

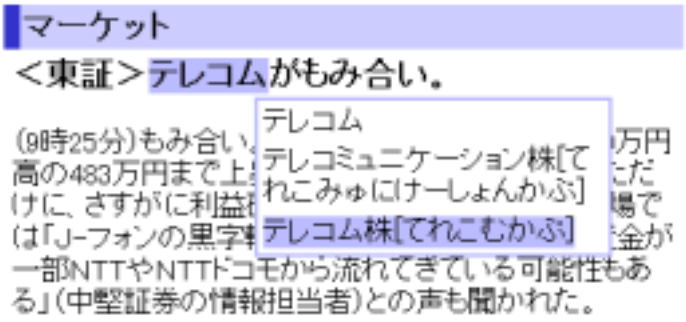

図 27　リストパラフレーズ

（3）リージョンパラフレーズ

マウスのドラッグで範囲指定した領域を一度に言い換えることもできる。（図 28、図 29）

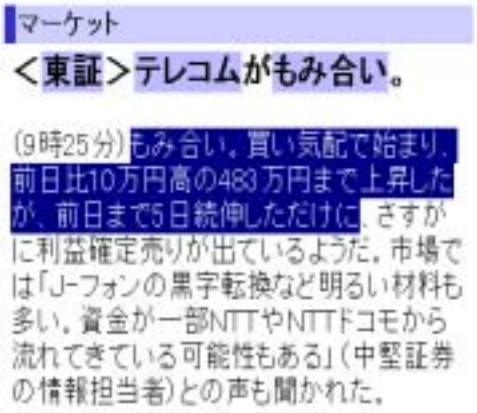

図 28　マウスのドラッグで領域選択

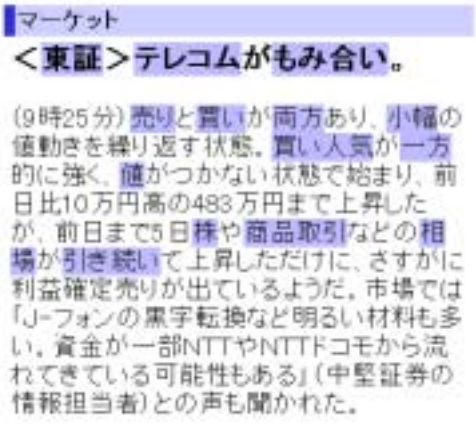

図 29　リージョンパラフレーズ

第６章
まとめと今後の課題

６．１　まとめと考察

説明してきたように、オンライン文書中の単語について、分からない個所を共有、かつ説明文を共有しページ辞書を構築していくことで、大勢の間での知識の共有が可能になり、アノテーションのコストも抑えられる。

さらに、構築されたページ辞書を閲覧者にさまざまな工夫をして見せることによりページ辞書をより活用でき、ユーザの理解を促進することができる。ページ辞書を持つページが今後増えていき、Webページの関連付けが進んでくると、ページ辞書のカバーする範囲は広くなり、より実用性のある、しかも動的に構築される辞書として重要な役割を果たすと考えられる。

今後オンライン文書がますます増加していく中で、本研究により、背景知識を持たない人であっても理解できる文書が増え、オンライン文書がより効率的に利用されるようになると考えられる。

６．２　今後の課題

今後は本研究の延長として以下の問題について考慮する予定である。

（１）言い換えの改善
言い換えに関しては、まだ結果が不自然な点が多く見られた。言い換えを行うために必要な情報として主に係り受け、活用、品詞情報を用いているが、より高度な言い換えのためには、文書内容に関するさらに詳しいアノテーションが必要になるだろう。

（２）ユーザ情報を考慮した表示
現在は語彙アノテーションが付与された単語すべてについて表示を行っている。しかし、それではユーザにとって既知の単語にまで表示してしまいオーバーヘッドが多く、処理にも時間がかかってしまう。ユーザ履歴などのユーザ情報を蓄積する仕組みを作り、ユーザが不必要だと考える表示については行わないようにする必要がある。

（３）提示された文書の分かりやすさ・難しさの定量的評価

文書の分かりやすさの指標には、含まれる語の難しさ、文の複雑さ（係り受けの複雑さ）、文の結束性、首尾一貫性、レイアウトの複雑さ、などが挙げられる。それらを利用して、文書の難しさの指標について多くの研究がなされているが、まだ決定的なものは見つかっていない。そのため、ポップアップ、挿入、グロッサリー、原文の言い換えの操作によって原文がどれだけ閲覧者にとって分かりやすくなっているかの評価は困難である。

（４）単語語義付けの自動化

オンライン文書中の単語に自動で語義情報を付けることは、アノテーションのコスト削減に大いに役立つ。たとえば、Web を利用して用語の説明を自動生成する研究も始められている[9]。人手では辞書登録に限界があるため、できるだけ自動化すべきであろう。語義を決定するには語義の曖昧さを解消せねばならない。こういった研究は word sense disambiguation という分野において多くなされており、その成果を利用したい。

（５）スケーラビリティ

ある少数のＷｅｂページ間のページ辞書は比較的簡単にマージされ得るが、関係するオンライン文書が多くなるにつれて、辞書の項目も飛躍的に増え、同じ見出し語の時の処理や、同義語の時の処理などを考えると、自動的なマージは困難である。しかし、辞書利用の効率化のため、辞書をうまくマージするルールなどの作成が必要である。また、大規模な実験を行っていないため、実際の運用の際、辞書サーバ、プロキシサーバ、文書変換サーバをどのように分散化し、安定したシステムを構築すべきか確かでない。

（６）ページ間類似度の計算式の評価

今回示したページ間の類似度を計算する式は試験的なものであり、まだシステム自体が作成されたばかりで、アノテートされているページが少なく、計算式の妥当性を検証するまでにいたっていない。

（７）著作権

オンライン文書をネットワーク上で加工・改変することは著作権上問題が起こり得る。これを回避するには trans publishing[6]のように元ページの情報へのポインタを必ず埋め込む手法を利用したり、「かな棒くん(http://www.kanabo.net/)」「ひらがなナビィ(http://www.flm.co.jp/kids/)」のようにトランスコーディングをクライアントサイドで行い著作権問題を回避する必要があるだろう。

以上の問題に対処することによって、より実用的で、閲覧者にとって利用しやすいが構築できるであろう。

謝辞

慶應義塾大学教授の石崎俊先生、IBM 東京基礎研究所の長尾確氏には研究の指針ならびに有益なアドバイスを頂きました。感謝いたします。また、プロキシサーバの実装については、ＩＢＭ東京基礎研究所学生研究員の細谷慎吾氏に協力していただきました。

参考文献

[1] **Extensible Markup Language(XML)**
http://www.w3.org/XML.

[2] **Jeff Heflin, James Hendler.**
Semantic Interoperability on the Web.
Extreme markup Languages 2000.

[3] **Masahiro Hori et al.**
Annotation-based Web Content Transcoding.
In Proceedings of the Ninth International WWW Conference 2000.

[4] **Hiroshi Maruyama, Kent Tamura, and Naohiko Uramoto.**
XML and Java Developing Web applications.
Addison-Wesley, 1999.

[5] **Katashi Nagao et al.**
Semantic Transcoding: Making the World Wide Web
more understandable and usable with external annotations.
TRL Research Report RT0386. IBM Tokyo Research Laboratory, 2000.

[6] **Theodor Holm Nelson**
Transcopyright: Dealing with the Dilemma of Digital Copyright.
Educom Review, 32:1 (January/February 1997), 32-5.

[7] **近藤恵子，佐藤理史，奥村学.「サ変名詞＋する」から動詞相当句への言い換え.**
情報処理学会論文誌 Vol.40 No.11, pp.4064-4074, 1999.

[8] **近藤恵子，佐藤理史，奥村学． 格変換による単文の言い換え**
情報処理学会研究報告 NL-135 pp.119-126, 2000.

[9] **桜井裕，佐藤理史． ワールドワイドウェブを利用した用語検索の実現**.
情報処理学会研究報告 2000-NL-137 pp.23-29, 2000.

[10] **佐藤理史，論文表題を言い換える．**
情報処理学会論文誌 Vol.40 No.7, pp.2937-2945, 1998.

[11] **奈良先端技術大学茶筅開発部 日本語形態素解析システム「茶筅」**
Version 2.0 for Windows 1999.
http://cl.aist-nara.ac.jp/lab/nlt/chasen/.

[12] **日本語構文解析システム** KNP
http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/

[13] **東中竜一郎，長尾確．アノテーションに基づく知的文書変換．**
情報処理学会研究報告 2000-ICS-120 pp.33-40, 2000.

[14] **松本裕治，黒橋禎夫，山地 治，妙木 裕，長尾 真．**
日本語形態素解析システム JUMAN **使用説明書 version 3.3 1997**.
http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html.