修士論文

閲覧時アノテーションを利用した Web ドキュメントの引用とその応用

350603294　林 亮介

名古屋大学大学院 情報科学研究科
メディア科学専攻

2008 年 1 月

閲覧時アノテーションを利用した Web ドキュメントの引用とその応用

350603294
林 亮介

本論文では，Web ドキュメントを対象とした新しい引用とその応用の仕組みを提案する．この仕組みでは，引用情報を引用先・引用元文書の部分要素に対するポインタと引用意図に関する属性を持つアノテーションとして XML 形式で定義し，引用先・引用元文書の部分要素間を繋ぐ双方向ハイパーリンクとして表現する．本仕組みは，引用者や閲覧者，引用分析を行う者にとって利点のある仕組みである．

次に，一般に文書を閲覧してから自身の文書を作成するまでに長期間が経過することを考慮して，閲覧時アノテーションと呼ばれる，人間が文書閲覧時に付与するメタデータを利用した引用支援手法を提案する．具体的には，文書閲覧時に閲覧時アノテーションを記録し，文書作成時に閲覧時アノテーションをトリガーにして，引用する予定の文書を検索するという引用文書検索手法を提案する．実験により，提案手法と一般的な引用文書検索手法を比較した結果，平均検索時間において提案手法が優れていることが分かった．また，閲覧時アノテーションに費やした時間と文書作成時に検索において短縮した時間を比較したところ，閲覧時アノテーションに費やした時間以上に検索時間を短縮できることが分かった．

さらに，提案手法を適用することで得られる引用情報である引用アノテーションを利用した，共引用の重み付けに基づく類似文書検索手法を提案する．従来の共引用による類似度の算出では，共引用の関係にある文書間の類似度は全て同じとされており，引用に関する意味的な情報は考慮されていない．そこで，引用アノテーションを利用して引用先文書の引用箇所間の距離と引用意図の関係を考慮し，共引用による類似度を詳細に重み付けする手法を提案する．実験により，引用先文書の引用箇所間の距離と引用意図の関係に応じた共引用により関連付けられる文書間の類似度を検証した結果，提案する重み付け手法は有効であることが分かった．

Quotation of Web Documents Using Reading Annotation and its Application

350603294
Ryosuke HAYASHI

In this thesis, we propose a new mechanism of quotation of Web documents and its applications. In our mechanism, we define a quotation information as an annotation to the documents described in an XML format. The quotation information includes the pointer to the internal element of the quoted document, the pointer to the internal element of the quoting document, and the attribute concerning the purpose of the quotation. And we represent the quotation information as a bidirectional hyperlink that connects the internal elements of the quoted document with that of the quoting document. Our mechanism has an advantage for document authors who quote online documents, for readers, and for researchers who perform citation analysis.

Also, we propose a method to facilitate the user to quote the document using reading annotation - metadata that we associate some attributes with any parts of the documents during reading them. Our proposed system records users' reading annotations and allows the user to retrieve them and quote parts of the document easily when the user writes a new document. We compared our method with a general retrieval method in some experiments. The results showed that our method was more effective than the general method in retrieval time.

In addition, we propose a method to similar documents based on co-citation extracted from quotation annotations - a set of quotation information accumulated in our system. Most of previous methods that have been proposed in citation analysis consider that all citations have the same similarities. Any semantic information on the quotation were not reflected in these similarities. We propose a method to consider semantic information on the quotation using our quotation annotation. Semantic information on the quotation include a distance between quotation parts and purposes of quotation. We experimented to show how effective our method is. Concretely, we classified co-citations based on their semantic information, and compared similarities between documents that have the relationship defined by a co-citation. The results of our experiment showed that our method was effective than the method that employed conventional measures.

目 次

第1章　はじめに

我々は文書を執筆する際，自己の主張に説得力を持たせるために関連する様々な文書の引用を行う．Web ドキュメントのように公開することを想定した文書においては特にその傾向が強い．引用とは，紹介，参照，論評やその他の目的で自己の著作物中に他者の著作物の原則として一部を参照あるいは掲載することである．引用の記述形式には，サイテーション（Citation）とクオテーション（Quotation）の二種類が存在する．一般に，サイテーションとは関係する文献や資料全体を参照記述する記述形式を指し，クオテーションは文献や資料の一部を一言一句そのまま記載する記述形式を意味する．

サイテーションは，引用者の言葉で要約して引用箇所を記述することが可能であるため，引用先文書の文脈に合わせて適切かつ簡潔に引用することが可能である．論文のような文書において見られるほとんどの引用の記述形式はサイテーションである．つまり，サイテーションは文字数の制限がある論文のような文書に適した記述形式であると言える．しかし，サイテーションによる記述には，引用元文書のどの主張部分に対して言及しているのか明示的に記述することができないという問題点や，引用者の主観により引用箇所を要約して記述するために引用元文書の著者の主張とは違った意味で引用してしまうという危険性を持つ．

一方，クオテーションの引用箇所には引用元文書の一部が一言一句そのまま記載されるため引用元文書のどの主張部分について言及しているのか明確であり，誤って文章を解釈し記述することもない．また，引用先・引用元文書の関係を文書の部分要素の粒度で詳細に記述することが可能であるため，サイテーションに比べクオテーションは引用者の主張を明確に記述することが可能な形式であると言える．しかし，クオテーションは論文のような文書においてほとんど見られないことからも分かるように，引用者にとって適切でない場合が存在する．例えば，引用したい主張が「システムの概要」のような引用したい主張部分が章全体のように文書の大部分にまたがる場合，クオテーションにより引用するのは冗長性が高い．つまり，クオテーションは引用先文書の文脈に合わせて適切かつ簡潔に引用することが困難な引用の記述形式である．

以上の考察から，サイテーションやクオテーションといった引用の記述形式は互いに一長一短であり，引用者にとってどちらも最適な記述形式とは言い難い．そこで本研究では，Web ドキュメントを対象とすることで，紙ベースのサイテーションやクオテーションといった両記述形式のメリットを併せ持つ Web ベースの新し

い引用の仕組みを提案する．具体的には，引用情報を引用先・引用元文書の部分要素に対するポインタ情報と引用意図に関する属性情報を持つアノテーションとして定義し，XML 形式で記述・管理することで紙ベースの引用の持つ問題点を解決する．提案する引用記述方式は，引用先・引用元文書の部分要素間を引用意図を持つリンクで繋ぐことから引用意図を持つ部分引用と呼ぶ．また，引用先・引用元文書のアノテーション情報として管理することから，引用意図を持つ部分引用を行うことで得られる引用情報を引用アノテーションと呼ぶ．提案する引用の仕組みは，引用者だけでなく閲覧者や引用情報の分析者にとっても利点のある仕組みである．

提案する引用の仕組みを実現するためには，まず Web ドキュメントの内部要素を扱う仕組みが必要である．現在，論文のような文書は主に PDF（Portable Document Format）を用いたフォーマットにより普及している．PDF は Adobe Systems 社によって開発された，電子文書のためのフォーマットである．レイアウトソフトなどで作成した文書を電子的に配布することができ，相手のコンピュータの機種や環境によらず，オリジナルのイメージを正確に再生することができる．しかし，PDF を用いた文書フォーマットは紙ベースの文書をコンピュータ上で表示，印刷することを前提に設計されたものであり，文書の内部構造に関する情報を保持していない場合が多い．また，内部要素に対する有効なアクセス手段も存在しないため，本研究で提案する部分引用を実現することが困難である．

そこで本研究では，文書やデータの意味や構造を記述するためのマークアップ言語の一つである XML（eXtensible Markup Language）を用いた文書フォーマットを提案する．XML は，内部構造に関する情報をタグとして保持することが可能であり，内部要素に対するアクセス手段として XPointer[2] と呼ばれる標準形式を利用することができる．また，XPointer 以外の既存の Web 技術との親和性も高く，Web ドキュメントを対象としている本研究に適した文書フォーマットであると考えられる．以後，この XML を用いた文書フォーマットを文書 XML と呼ぶ．

次に，文書 XML を作成し引用意図を持つ部分引用を行うためのシステムについて述べる．本システムは，ユーザの文書を作成する負担を軽減するための引用支援機能を併せ持つ．我々は文書を閲覧する際に，文章に対してしばしば下線引きやマーカ引きといったマーキングやメモ書きのようなコメントの付与を行う．閲覧する目的が明確な場合のマーキングやコメントは閲覧者が文書に対して行う自然なアノテーション行為である．本研究では，文書閲覧時に行われるこれらのアノテーションを総称して閲覧時アノテーションと呼ぶ．Schilit ら [3] は閲覧時アノテーションには文脈や個人の知識背景といった様々なコンテキストが含まれると述べている．例えば，研究活動の一環であるサーベイ時に行われる閲覧時アノテーションの場合，研究内容や背景に関するコンテキスト，つまり文書執筆時に利用することが可能な情報が多く含まれていると考えられる．つまり，閲覧時アノテーションには文書の引用に関連した情報が含まれていると考え，引用を行う前段階

の情報として利用する．具体的には，閲覧時アノテーションを記録し，文書作成時に検索・引用可能にすることで文書の引用支援を行う．閲覧時アノテーションの一つであるマーキングは文書中の部分箇所を特定する行為である．そのため，ユーザはマーキングされた部分を引用することで，文書引用時には被引用箇所を明示的に指定する必要がなくなる．また，閲覧時アノテーションを付与するためには必ず一度は文書を閲覧するため，閲覧時アノテーションには「孫引き」のような不適切な引用を防止する効果があると考えられる．さらに，閲覧時に考察したことや感じたことをコメントとして文書に付与することが可能であるため，引用する箇所や理由を忘れないという利点も考えられる．

また，閲覧時アノテーションには引用を目的としたもの以外にも様々な目的のものが存在する．例えば，専門用語の意味が理解できない場合にその専門用語をマーキングする場合があるが，マーキングした後に専門用語の意味を辞書で調べることが予想される．このように後に行う行為が予想される場合には，システマティックにオペレーションを対応させることが望ましい．つまり，マーキングを行うとシステムが自動的に辞書引きを行うような閲覧支援を行う必要がある．また，アノテーションとして付与された辞書情報をユーザ間で共有することで，次にその文書を閲覧したユーザは辞書を引く必要がなくなる．閲覧支援を行うことにより，ユーザは短時間で正確に文書の内容を理解することが可能になるため，間接的に文章の意味解釈の誤りを原因とした誤引用を防ぐことに繋がる．つまり，直接は引用を目的としない閲覧時アノテーションも，ユーザの引用行為を支援するために利用される．

次に，本システムを運用することで蓄積された引用アノテーションを用いた応用例を示す．計量書誌学（Bibliometrics）における主要な分析手法に引用分析（Citation Analysis）と呼ばれる手法が存在する．引用分析とは、文書間の引用・被引用関係を用いて文書間の関係を分析する研究である．引用分析において提案された有名な尺度に，共引用 [4]（Co-citation）と呼ばれる関係を用いた文書間の類似尺度が存在する．共引用の関係とは，同一の文書によって引用された文書同士の関係を指し，この関係にある文書同士には類似性があるとされている．しかし，従来の共引用による類似度の算出では，共引用の関係にある文書間の類似度は全て同じとされており，引用に関する意味的な情報は考慮されていない．

そのため，近年では引用に関する意味的な情報を文書中から抽出して引用分析を改良する試みがなされている．難波ら [7] は，引用先文書の著者がどのような意図で引用したかに着目している．例えば，引用には自己の論説の根拠を示すために行われる引用や他者の主張の問題点を指摘するために行われる引用が存在する．難波らは，書誌結合 [8]（Bibliographic Coupling）の関係を用いた文書間類似度を引用意図の違いを考慮して算出している．書誌結合の関係とは，共引用より古くより提案されている文書間の類似度指標で，ある同一の文書を引用している文書同士の関係を表す．難波らは，引用意図を考慮して文書間の関係を意味的に捉え

ることで，書誌結合のような類似尺度を高精度に算出することが可能になると述べている．例えば，引用元文書の主張を自己の論説の根拠にするために引用している文書と引用元文書の主張の問題点を指摘するために引用している文書では引用先文書の著者の立場が異なり，従って引用先文書で述べられている内容も異なるだろう．共引用の関係についても同様のことが言えるため，共引用の関係を用いて類似度を算出する上でも引用意図を考慮することは重要な要素の一つであると考えられる．

また，江藤 [9] は，文書は体系づけられて構成されるため意味的に近い内容のものはまとまって述べられるという仮説を立て，引用先文書の引用箇所間の距離が共引用における文書間の類似性に影響を与えると述べている．引用先文書の文脈情報を利用することで共引用による文書間の類似度は精度の高いものになると提案している．例えば，引用文書の冒頭の「はじめに」で引用された文書と末尾の「おわりに」で引用された文書の類似度と,「関連研究」について述べられている部分の同一パラグラフ内で引用された文書同士の類似度は異なるだろう．そのため，共引用の関係を用いて類似度を算出する上で，引用先文書の引用箇所間の関係を考慮することは重要な要素の一つであると考えられる．

しかし，現状では上記した引用情報を考慮した引用分析を行い，具体的な応用を実現することは困難である．なぜなら，引用情報がすべて文書中のテキストに埋め込まれてしまっているからである．文書中のテキストに対して自然言語処理を行い，引用情報を自動抽出する方法も考えられるが，引用箇所の文章記述のフォーマットも単一ではなく抽出することは難しい．そのため，あらかじめ応用を考慮した形で引用情報を取得・管理することが望ましい．

以上の現状の引用分析における考察を踏まえ，本システムを運用することで蓄積される引用アノテーションの有効性を示す．本手法では，引用情報を文書のアノテーション情報として文書テキストとは分けて管理しており，引用分析に利用することが可能である．そこで，引用アノテーションを用いた一つの応用例として，従来の引用情報では実現が困難であった類似文書検索手法を提案する．また，類似文書検索を実現することにより，ユーザの文書のサーベイ活動が活性化し，従来では発見できなかったような文書の引用を促進する．結果として，さらに有益な引用アノテーションが収集され，検索精度が高まるというポジティブなフィードバックサイクルが発生すると考えられる．

本研究では，提案する手法を評価するために二つの実験を行った．第一に，閲覧時アノテーションを用いた文書の引用支援手法の有効性を示すために被験者実験を行った．延べ人数 16 人の被験者に対して，文書閲覧時にアノテーションするグループとアノテーションしないグループに分かれてもらい，文書作成時に引用する文書を検索してもらった．その結果，検索に費やした平均時間において，文書閲覧時にアノテーションすることの優位性を示すことができた．また，閲覧時アノテーションに費やした時間と文書作成時に検索において短縮した時間を比較

したところ，閲覧時アノテーションに費やした 3 倍以上の時間を文書作成時に短縮できることが分かった．

第二に，引用アノテーションを用いた類似文書検索の有効性を示すために引用アノテーションを用いた共引用の文書間の類似度に関する実験を行った．具体的には，引用意図の関係と引用箇所の距離に基づき共引用を分類し，各共引用によって関連付けられる文書同士の平均類似度を比較した．その結果，どちらの分類方法についても良好な結果が得られ，提案する類似文書検索手法の可能性を示すことができた．

本論文は本章を含めて 7 章から構成される．第 2 章では，Web ドキュメントの最適な引用の仕組みについて，現状の引用に関する問題点を踏まえながら述べる．第 3 章では，閲覧時アノテーションの付与可能な文書閲覧インタフェースと閲覧時アノテーションを検索・引用可能な文書作成インタフェースを持つ引用支援システムについて述べる．第 4 章では，本システムを運用することで蓄積される引用アノテーションを用いた応用例として共引用に基づく類似文書検索手法を提案する．さらに第 5 章では，3 章で提案した引用支援手法と 4 章で提案した類似文書検索手法の有効性について実験・考察する．そして第 6 章では，本研究の関連研究について述べ，最後に第 7 章で本論文をまとめ，今後の課題を述べる．

第2章　Webドキュメントの引用

本章では，従来の文書を引用する際の記述形式について考察し，Web ドキュメントを対象とした引用の仕組みを提案する．また，従来の引用分析における問題点を共引用を例に述べ，提案する仕組みを適用することで得られる引用アノテーションと呼ばれる引用情報の有効性を考察する．

まず 2.1 節では，サイテーションとクオテーションと呼ばれる従来の引用の記述形式の問題点について述べ，求められる引用の記述形式を考察する．2.2 節では，2.1 節の考察を踏まえ，従来の記述形式であるサイテーションやクオテーションという枠に捉われない Web ドキュメントの引用の仕組みを提案する．2.3 節では，提案する引用の仕組みを実現するために，現在広く普及している PDF に基づく文書フォーマットの問題点について述べ，XML に基づく文書フォーマットを提案する．2.4 節では，共引用に基づく引用分析の現状と問題点について述べ，引用アノテーションの有効性について考察する．

2.1　サイテーションとクオテーション

引用とは，紹介，参照，論評やその他の目的で自己の著作物中に他者の著作物の原則として一部を参照あるいは掲載することである．現在，文章を引用する際の記述形式には，サイテーション（Citation）とクオテーション（Quotation）の 2 種類が存在する[1]．一般に，サイテーションとは関係する文献や資料自体を参照記述する記述形式を指し，クオテーションは書かれたものの一部を一言一句そのまま記載する記述形式を意味する．

現在，論文のような文書において見られるほとんどの記述形式はサイテーションである．サイテーションは引用箇所を引用者で要約して記述することが可能であるため，引用文書の文脈に合わせて適切かつ簡潔に引用することが可能である．つまり，サイテーションは文字数の制限がある投稿論文のような文書に適した記述形式であると言える．しかし，サイテーションによる記述にはいくつかの問題点がある．まず，図 2.1 のように，サイテーションによる引用は，引用元文書のどの主張部分について言及しているのか明示的に示されないため，引用元文書全体

[1]サイテーションは「引用」ではなく単なる「参照」と呼ばれることもあるが，本研究では引用の一種として捉える．

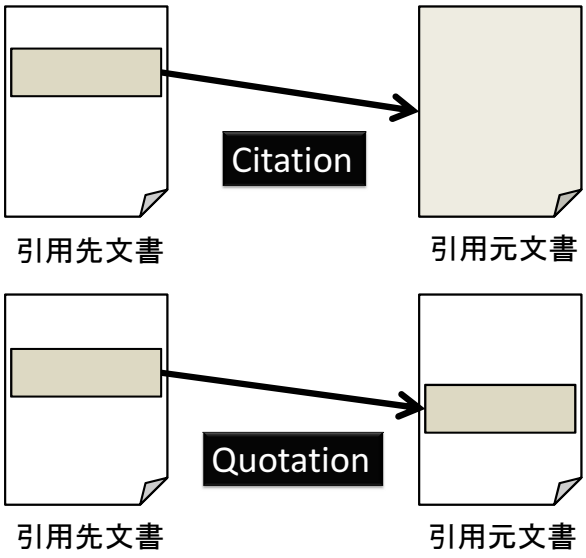


図 2.1: サイテーションとクオテーションによる文書間の関係

に対するポインタを表現していると言える．通常，引用元文書の出所に関する情報は，巻末の「参考文献」のような項目に著者，文献名，出典，日付を用いて記述されることが多い．しかし，文書によっては数十ページに及ぶものもあり，引用している主張部分がその一部の数ページということも往々に存在する．つまり，サイテーションのような記述形式では，引用元の出所を明確に示すことはできない．そのため，サイテーションは引用先文書と引用元文書の関係を曖昧に記述する結果となっている．また，サイテーションの引用先文書の引用部分は引用者により要約されるため，引用元文書の著者の意図とは違った意味で引用してしまう危険性がある．つまり，引用元文書の主張を誤って解釈・記述してしまう可能性がある．また，その誤りを訂正する方法も存在せず，孫引き[2]によってさらに誤った引用が行われる可能性がある．

一方，クオテーションの引用部分には引用元文書の文章の一部が一言一句そのまま記載されるため，引用元文書のどの主張について言及しているのか明確であり，また，誤って文章を解釈し記述することもない．つまり，図2.1のように引用先・引用元文書の部分箇所間の関係を記述している考えることができるため，サイテーションに比べクオテーションは引用先・引用元文書の関係を明確に記述することが可能な形式であると言える．しかし，クオテーションは現在の文書においてほとんど使用されないことからも分かるように，引用者にとって適切でない場合が存在する．例えば,「システムの概要」のような引用したい主張が章全体のように文書の大部分にまたがる場合，クオテーションにより引用するのは冗長であ

[2]孫引きとは，引用している文書を引用箇所のコメントどおり受け売りして引用することを指す.

る．また, 著作権法についての最高裁判所昭和 55 年 3 月 28 日の判例において，引用先が「主」，引用部分が「従」の関係にあることが引用の条件とされている．つまり，引用部分は質的にも量的にも従属的なものであり，本文を主とするならば引用部分は従たる関係でなくてはならず，引用部分の分量が多い文章は著作権法の観点から見ても好ましくないと言える．そのため現状では，一文のように非常に短い文章に限り，クオテーションによって引用されることが多い．また，引用元文書の記述が引用先文書で記述したい文脈に適さない場合も存在することから，クオテーションは引用先文書の文脈に合わせて適切かつ簡潔に引用することが困難な記述形式である．

上記のように，クオテーションの方がサイテーションに比べ，引用先・引用元文書の関係を詳細に記述することが可能であり，引用者の意図を明確に伝えることができる記述形式であると言える．しかし，クオテーションは引用先文書の文脈に適した形で簡潔に引用することが困難であるため，現状ではほとんど利用されていない．つまり，求められる引用の記述形式はサイテーションのように引用先文書の文脈に適した形の柔軟な引用が可能であり，クオテーションにように引用元文書の主張部分に対する明確なポインタの記述が可能である必要があると考えられる．

2.2 引用意図を持つ部分引用

前節で，サイテーションとクオテーションと呼ばれる引用の記述形式の持つ特徴について考察し，求められる引用の記述形式について考察した．クオテーションでしか引用元文書の主張部分を明示的に示すことができないのは，そもそも引用が紙上で行われることに起因している．つまり，サイテーションとクオテーションの持つ問題点は，引用者が文章記述により引用元を指し示すのでなく，引用元文書の被引用箇所を直接的なポインタにより指し示すことが可能な仕組みを提供することで解決できる．

そこで本研究では，従来のサイテーションとクオテーションといった紙ベースの引用の記述形式ではない，Web ドキュメントを対象とした引用の仕組みを提案する．具体的には，引用情報を Web の特徴であるハイパーリンクを用いて双方向リンクとして表現することで，紙ベースの引用記述の持つ問題点を解決する．提案する引用の仕組みは，ネルソンの提唱するトランスクルージョン [10]（Transclusion）の仕組みに類似しており，引用元文書の独自性などに影響を与えない仕組みである．

また，引用情報を引用先・引用元文書の両文書のアノテーション情報として XML 形式で記述・管理する．アノテーションとは，あるデータに対して関連するメタデータを注釈として付与することを言う．つまり，引用情報を文書の本文中にテキスト情報として保持するのではなく，文書の二次情報として扱う．アノテーション情報として，引用先文書の引用箇所と引用元文書の被引用箇所に対するポイン

タ情報を定義することで，サイテーションの特徴である引用箇所の自由記述を可能にし，尚且つクオテーションの特徴である引用元文書の主張部分を指し示すことを実現する．本研究では，このアノテーションを引用アノテーションと呼ぶ．

提案する引用の仕組みの持つ主な利点は以下のとおりである．

- 引用者にとってのメリット
  本仕組みは引用元・引用先に対するポインタ情報を保持しているため，引用箇所と被引用箇所の文章表現が異なる場合にも引用構造を失わない．そのため，引用者はサイテーションのように引用箇所を自身の言葉で要約して記述することが可能である．また，引用者は引用意図に関する属性情報（論拠の明示，論説の対比，問題点の指摘，先行研究の例示など）を引用情報に付与することが可能であるため，従来より明確に自身の主張を記述することができる．

- 閲覧者にとってのメリット
  通常，閲覧者は引用情報を引用先文書側からしか参照することができないが，本仕組みでは引用を引用先・引用元文書に対する双方向リンクとして表現するため，引用元文書側からの参照が可能である．また，引用をハイパーリンクとして表現するため，閲覧者は煩雑な作業をすることなく引用元・引用先文書の該当箇所にアクセスすることができる．

- 分析者にとってのメリット
  本仕組みでは，引用情報を原文情報とは別にアノテーションとしてXML形式で管理する．そのため，従来のように文書内のテキストを言語処理することなく，計算機を用いて引用情報を分析することが可能である．また，引用の構造情報と引用意図に関する情報を考慮することができる．

提案する引用の仕組みは引用元文書と引用先文書の部分要素間を引用意図によって繋ぐことから，引用意図を持つ部分引用の仕組みと捉えることができる．引用意図を持つ部分引用を行う具体的な方法については，次章でシステムと共に説明する．

2.3 XMLに基づく文書フォーマット

前節で提案した引用の仕組みを実現するためには，文書の内部要素に対するポインタを定義し，扱う仕組みが必要である．そこで本節では，XMLに基づく文書フォーマットを提案する．

現在，論文のような文書は主にPDF（Portable Document Format）を用いたフォーマットにより普及している．PDFはAdobe Systems社によって開発され

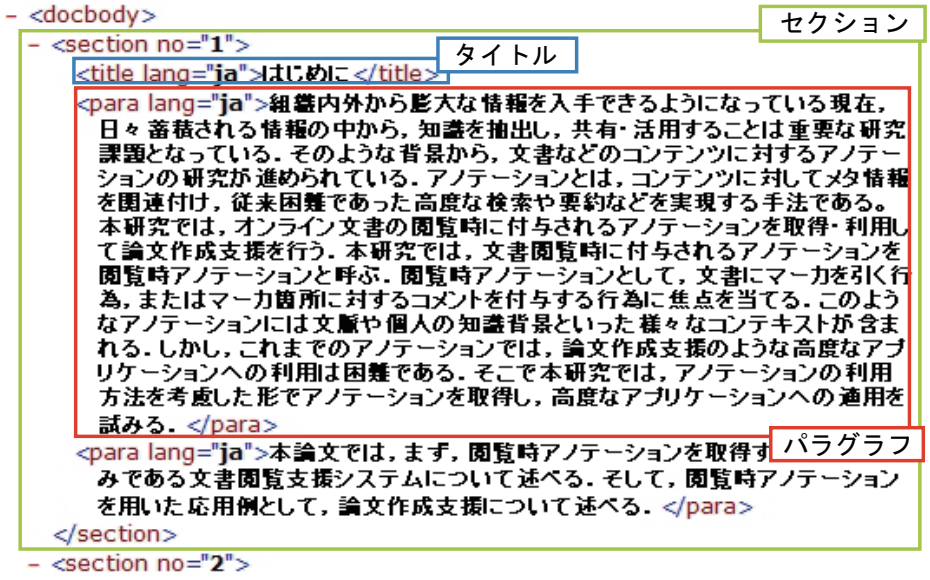


図 2.2: 文書 XML の記述例

た，電子文書のためのフォーマットである．レイアウトソフトなどで作成した文書を電子的に配布することができ，相手のコンピュータの機種や環境によらず，オリジナルのイメージを正確に再生することができる．しかし，PDF を用いた文書フォーマットは紙ベースの文書をコンピュータ上で表示，印刷することを前提に設計されたものであり，文書の内部構造に関する情報を保持していない場合が多く，また，内部要素に対する有効なアクセス手段も存在しない．ここでの内部構造に関する情報とは，セクションやパラグラフといった文書の意味的な構造情報を指す．そのため，本研究で提案する部分引用を実現することは困難である．

そこで本研究では，文書の内部構造に関する情報を扱うために，文書やデータの意味や構造を記述するためのマークアップ言語の一つである XML（eXtensible Markup Language）に基づく文書フォーマットを提案する．XML は，SGML（Standard Generalized Markup Language）をインターネットで活用しやすくするために，W3C（World Wide Web Consortium）にて策定された汎用の構造化データ記述言語である．XML では自由にタグを定義でき，文書中の文字列に意味付けができる言語構造を持つため，内部構造に関する情報をタグとして保持することが可能である．さらに，プログラムで自在に XML データを情報処理できるというメリットや，SGML の持つ複雑な印刷系のオプションなどを省略して言語仕様を規定しており，理解しやすく使いやすいというメリットがある．本研究では，この XML を用いた文書フォーマットを文書 XML と呼ぶ．

XML は Web との親和性が高い言語であり，Web における様々な既存の技術を利用することが可能である．本研究では，その中の一つである XPointer[2] を内部

図 2.3: 言語構造に関するタグの付与

要素を指し示すポインタ言語として利用する．XPointer とは W3C が提案している XML ポインタ言語であり，コンテンツの URI（Uniform Resource Identifier）とドキュメントの内部ノードを指し示す手段として標準化されている XPath[1] を記述することで，内部構造への言及を可能にしている．以下のように，コンテンツの URI である [Content] に続き，#xpointer 以降 [XPath] に XPath を記述することで，内部構造への言及が可能である．

[Content]#xpointer([XPath])

文書 XML の記述例を図 2.2 に示す．文書 XML では，文書構造を表すために様々な内部要素を表すタグが用意されている．例えば，セクションを表す要素として <section> 要素や <subsection> 要素，パラグラフを示す要素として <para> 要素，図を表す要素として <figure> 要素などが定義されている．

一方で，引用は文や句のようなパラグラフより小さい単位で行われる場合が存在する．例えば，文書に専門的な用語の意味を記述する際に，その用語が初めて使用された文書を引用することがある．この場合，引用先文書の引用部分は専門的な用語，つまり単語ということになる．本研究では，単語のような詳細な箇所を XPointer により指し示すために，パラグラフの内部に含まれる文字列に対して

形態素解析を行い，図 2.3 のように形態素に対して動的に言語構造に関するタグを埋め込む．動的にタグを埋め込むことで，パラグラフより詳細な単位に対するパスを XPath を用いて記述することができ，XPointer で指し示すことが可能になる．また，形態素に言語構造に関するタグを埋め込むことで，形態素単位以下の範囲の引用は構文的に誤っていることを機械的に認識することが可能となる．

また，文書 XML と PDF による文書フォーマットは排他的な関係ではないと考えている．つまり，Web 上で利用する際には文書 XML，紙として印刷して利用する際には PDF といったように，互いの特徴を活かし，互換性を持つべきである．そのため，相互にフォーマットを変換する仕組みがあることが望ましい．

2.4 従来の引用分析における問題点

本節では，従来より行われている引用分析の問題点について述べ，本研究で提案している引用の仕組みを適用することで蓄積される引用アノテーションの有効性について考察する．

引用は人間が意味を考慮して情報間の関連付けをする作業として捉える事ができる．計量書誌学（Bibliometrics）において，文書間の引用・被引用関係を用いて文書間の関係を分析する引用分析（Citation Analysis）という研究が古くから行われており，引用情報を活用した文書間の関係を測定する各種の尺度が提案されている．しかし，従来の引用分析においては，引用に関する意味的な情報は考慮されていない．

例えば，引用分析には共引用 [4]（Co-citation）と呼ばれる関係を用いた文書間の類似尺度が存在する．共引用の関係とは，同一の文書によって引用された文書同士の関係を指し，この関係にある文書同士には類似性があるとされている．例えば，図 2.4 のような文書間の引用関係があったとする．この場合，文書 A と文書 B が共引用の関係にあり，共引用された回数が 1 回としてカウントされる．この回数が指標値として利用され，文書同士の類似度が算出される．つまり，算出対象の文書同士を共に引用している文書の数によって類似度が決まる．共引用の関係は，CiteSeer[5] のような実用システムにおいて類似文書検索に利用されていることからも，その有用性は高い．

従来の共引用による類似度の算出では，共引用の関係にある文書同士の類似度は全て同じとされており，引用先文書の引用意図や文脈情報のような意味的な情報は考慮されていない．しかし，引用者は様々な意図を持って文書を引用する [6]．例えば，先行研究に敬意を表して行われる引用と自己のシステムの有効性を示すために他者のシステムの機能と対比するために行われる引用では，引用により関連付けられる引用先文書と引用元文書の関係が異なるだろう．そのため，共引用の関係にある文書同士の関係も対象となる引用同士の関係に応じて異なってくると推測される．例えば，論文において「関連研究」に関する章で引用されている

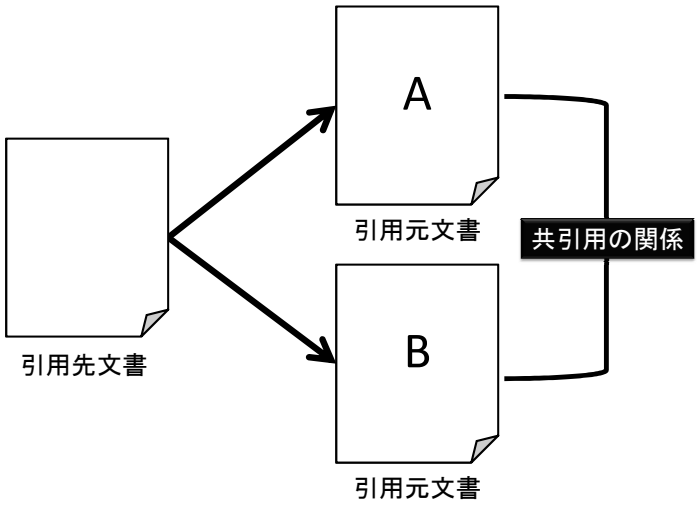


図 2.4: 共引用の関係

文書同士と冒頭の「はじめに」のような大きく論理を展開している章で引用されている文書同士では，類似度が異なると考えられる．

また，インパクトファクタ [11](Impact Factor) のような引用統計指標を算出する上でも，引用に関する意味的な情報を考慮することは有効である．インパクトファクタは，自然科学・社会科学分野の学術雑誌を対象として，その雑誌の影響度を測る指標であり，その雑誌に掲載された論文の平均的な被引用回数である．同じ分野における異なる雑誌の重要度を比較する場合に，インパクトファクタを用いるのは有効である．しかし，インパクトファクタでは引用がどういった文脈で行われているかといった意味的な情報が考慮されていない．例えば，その雑誌の掲載論文は肯定的に引用されることが多いのか，それとも批判的に引用されることが多いのかといったことが考慮されておらず，雑誌間の詳細な比較ができない．

しかし，現状では引用に関する意味的な情報を計算機上で機械的に扱うことは困難である．なぜなら，引用情報がすべて文書中のテキストに埋め込まれているからである．文書中のテキストを機械処理することで引用の意味を推定する方法も考えられるが，引用箇所の記述に対して自然言語処理を行い，抽出することは容易ではない．

そこで本研究では，構造化された文書 XML と引用アノテーションを利用することで，前述の引用分析における問題点の解決を試みる．前節で述べたように，文書 XML は文書の意味の単位であるパラグラフやセクションといった構造情報を保持しているため，文書の文脈情報を利用することが可能である．また，引用アノテーションは文書中のテキストから自動抽出するのではなく，引用先文書の著者

が引用する際に取得されるため，引用先文書の引用箇所，引用元文書の被引用箇所や引用意図に関する属性といった豊富な引用情報を保持している．さらに，引用情報を文書のアノテーション情報として本文のテキスト情報とは分けて管理しており，XML 形式で記述されているため引用分析に適用することが可能である．

2.5 まとめ

本章では，従来の引用の記述形式について考察し，Web ドキュメントを対象とした引用の仕組みを提案した．また，提案する引用の仕組みを適用することで得られる引用アノテーションの有効性について考察した．

まず，サイテーションとクオテーションと呼ばれる従来の引用の記述形式の問題点について述べ，求められる引用の記述形式を考察した．次に，従来の記述形式であるサイテーションやクオテーションという枠に捉われない Web ドキュメントの引用の仕組みを提案した．そして，提案する引用の仕組みを実現するために，現在広く普及している PDF に基づく文書フォーマットの問題点について述べ，その解決策として XML に基づく文書フォーマットを提案した．さらに，提案する引用の仕組みを適用し，蓄積される引用アノテーションを利用することで，引用分析をより高度に行うことが可能になることを述べた．

第3章　閲覧時アノテーションを利用した引用支援

前章では，Web ドキュメントを対象とした引用の仕組みを提案した．そこで，本章では提案する引用の仕組みを実装したシステムについて述べる．また，本システムは単に提案する引用の仕組みを実現するだけでなく，ユーザの引用行為を促進する機能を有することを示す．

まず，3.1 節で文書作成時における引用文書検索の問題点について述べ，その解決策として閲覧時アノテーションと呼ばれるユーザの自発的な行為から得られる情報を利用した引用文書検索手法を提案する．なお，文書作成時において引用すべき文書を検索することを引用文書検索と呼ぶ．3.2 節では，閲覧時アノテーションには，引用支援以外にも様々な副次的な利用方法が存在することを述べる．3.3 節では，提案している引用意図を持つ部分引用を行うために実装したシステムについて述べる．本システムは，大きく分けて閲覧支援機構，作成支援機構から構成される．システム構成について述べた後，各機構におけるインタフェースと主な機能について述べる．

3.1　閲覧時アノテーションを利用した引用文書検索

3.1.1　従来の引用文書検索における問題点

一般に，文書作成時に引用する文書は過去に閲覧したことのある文書であり，閲覧してから長い期間が経過していることが多い．例えば，研究活動の場合，他者の論文をサーベイしてから自己の論文を執筆するまでには，約 1 年間が経過していることも往々にしてあるだろう．そのため，既読文書を後で容易に検索するためには，文書をフォルダ分けしたり，文書にタグを付与したりして管理する必要がある．そのため，EndNote[12] のような文書管理を行う文書作成支援システムが提供されている（図 3.1 参照）．EndNote では，文書情報を個人ライブラリに保存し，管理することができる．そして，自身の文書を作成する際に，ライブラリに保存した文書情報を検索し，引用元文書として引用することが可能である．

本研究では，文書を引用するために行われる検索は引用者の主観的な部分が大きいため，主観的な情報によって文書を検索可能であることが重要であると考え

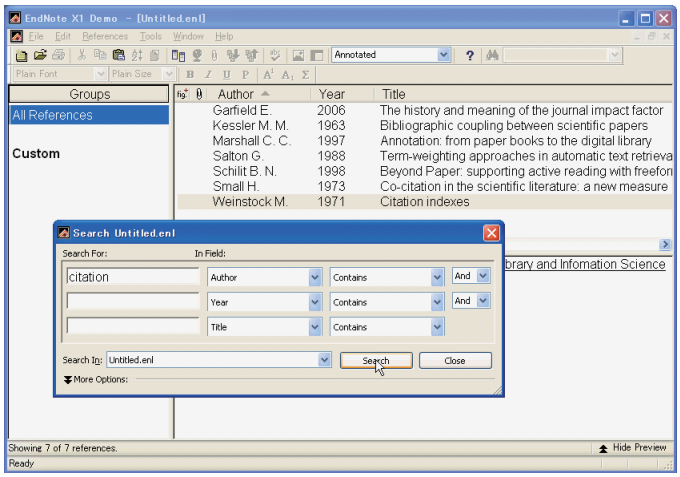


図 3.1: EndNote 画面例

ている．例えば，過去に閲覧した「アノテーション」に関する文書を引用したいと考えたとする．その場合，おそらく「アノテーション」を検索のキーワードとして検索するだろう．しかし，その文書中では「アノテーション」という用語は出現せずに,「注釈」と記述されていたり，もしく単純に「メタデータ」と記述されていたりして，ユーザの検索要求にヒットしないことがある．そこで，概念辞書のようなものを用いて検索キーワードである「アノテーション」を「注釈」や「メタデータ」に拡張して，検索するような解決法が考えられるが，今度は不要な文書まで検索でヒットしてしまい，その中から絞り込む必要が出てきてしまう．このような場合，語の表記ゆれなどが問題なのではなく，ユーザ自身が閲覧した際にどの文書を「アノテーション」に関する文書として捉え，どのように解釈したかということが問題となる．しかし，現状ではそのような主観的な情報を閲覧履歴から機械的に抽出するのは困難であるため，ユーザ自身が付与する仕組みが必要である．そのため，少なくとも既に閲覧した文書に対しては，ユーザの解釈に関する情報をユーザ自身によって付与可能にし，その情報をトリガーとした文書検索を可能にすべきである．

EndNote では，文献管理を行うためにユーザが自ら文献情報を記入することが可能である．文献情報の中には,「Research Notes」と呼ばれる主観的な情報を文書に関連付けて記述でき，その情報をトリガーに検索することが可能である．しかし，EndNote の文献情報はあくまでも文書全体に関連付けて記述するものである．一般に，引用とは，紹介，参照，論評その他の目的で自己の著作物中に他人の著作物の原則として「一部」を参照あるいは掲載することを指す．そのため，引用するためには自分の記憶を頼りに検索された文書の中から引用に該当する「一部」を再度探し出す必要がある．しかし，数十ページにも及ぶような文書の場合，その作業は非常に煩雑なものであり，ユーザにとって大きな負担になる．そのた

め，引用文書検索において主観的な情報を文書の部分要素に関連付けて記述可能にし，その情報を検索可能にすることが必要である．

3.1.2 閲覧時アノテーションの検索

本研究では，文書の閲覧者が行う自発的なアノテーション行為に着目し，そのアノテーション行為より得られる情報を文書の部分要素に関連付けられた主観的な情報として利用を試みる．一般に，文書を閲覧する際には，文書中で重要だと感じた箇所に対してマーキングしたり，考察や感想をコメントとして記入する．このような行為は，文書の原文に対して二次的な情報を付与していくことから文書に対するアノテーション行為として捉える事ができる．本研究では，これらマーキングやコメントといったアノテーションは文書閲覧時に付与されることから閲覧時アノテーションと呼ぶ．なお，一般的な情報検索システムにおけるインデックス情報のような機械的に生成されるメタデータも原文に対する二次情報であることからアノテーションと呼ばれることがあるが，ここでのアノテーションには含まないこととする．ここでのアノテーションとは，原文の著者を含むすべてのアノテーション行為をおこなう人間が人為的に作成する二次情報を指し，機械的に付与される類の二次情報とは分けて考える．

Schilit ら [3] はアノテーションには文脈や付与したユーザの知識背景といった様々なコンテキストが含まれると述べている．例えば，研究活動の一環であるサーベイ時に行われるアノテーションに着目すると，閲覧者の研究内容や背景に関するコンテキストのような主観的な情報が多く含まれていると考えられる．また，閲覧時アノテーションは文書中の重要箇所を記録するという目的で行われ，引用する文章の記録に利用されることがある [13]．そのため，閲覧時アノテーションを引用する前段階の情報として利用することは人間の直感に適していると考えられる．そこで，引用文書検索を行う際に検索対象とし，マーキング箇所の文章を引用可能にすることでユーザの引用支援を行う．

アノテーションの形式には，下線引き，ハイライトマーカによるマーキング，アスタリスク，囲み文字やコメントといった様々な種類が存在する．紙に比べ Web においては，下線や囲み文字に比べてハイライトマーカによりマーキングすることが多い [14]．そこで本研究では，ハイライトマーカによるマーキング（以後，簡略化してマーキングと呼ぶ）とコメントを閲覧時アノテーションとして扱う．

コメントには部分箇所と関連付けられないもの（文書の余白に書かれるメモ書きのようなもの）も存在するが，本研究ではマーキング箇所に関連付けられたコメントのみを記録の対象とする．なぜなら，コメントを記述するトリガーとなった文書の該当箇所を明示的に記録しておくことで，再び理解する際の文脈情報として利用できるからである．しばしば，後で参照するために付けておいたコメントを実際に読み返したとき，その内容が理解できないことがある．その時自分が

表 3.1: マーキング属性

属性名	説明
引用箇所記憶	引用する箇所を記憶する為のマーキングに付与する属性
筆者の主張	筆者の主張部分で重要な箇所に付与する属性
文書エッセンス	該当文書の本質的な箇所に付与する属性
客観的事実	客観的な事実 (実験データ等) で重要な箇所に付与する属性
論理の展開	論理を展開している重要な箇所に付与する属性
並列関係	並列な要素に対して付与する属性
英単語	分からない英単語に付与する属性 (英和辞書を自動で引くオペレーションが対応)
専門用語	分からない専門用語に付与する属性 (専門用語辞書を自動で引くオペレーションが対応)
その他	上記以外のマーキングに付与する属性

何を考えていたのか，そもそも自分がどういうつもりでその情報を書き留めたのか，その当時の思考の文脈や意図が思い出せず，結局ほとんど役に立たないという場合である．これは，コメントが記述されたその場の状況・思考の文脈が，コメント単体からは読みとれないことから起こりうるケースである．コメントは自分が必要と感じた情報を記録しておくものだが，その場の思考の流れを完全に明記することは実際には不可能である．そのため，コメントを記述する当人にとって当たり前のことはわざわざ書き留めないため，コメントに含まれる内容はどうしても断片的な記述となる。結果として，後でコメントを参照しても理解の手助けとなる文脈情報に欠け理解することが難しくなり，コメント情報単体では役に立たないことが多い．

本研究では，マーキングやコメントにアノテーションした意図に関する属性情報を付与することが可能である．各閲覧時アノテーションに付与可能な属性を表 3.1，3.2 に示す．マーキングにおいてはマーキング色に対応付けて属性情報を付与できる．そのため，色を用いてマーキング箇所の情報を分類しておくことで，様々な利用場面に応じた検索が可能である．例えば，自己の主張の新規性を示すために，他者の主張と対比して文章を書きたい場面がある．その場合，文書の閲覧時にあらかじめ対比したい主張部分をマーキングし，自己の主張との差分をコメントし，コメントに問題点属性を付与しておく．すると，実際に自己の文書を作成する際に，コメントの問題点属性を指定して検索することで，目的に合った箇所のみを収集することが可能である．つまり，アノテーション属性とキーワードを指定することにより，ユーザの文脈に適した閲覧時アノテーションを検索することが可能である．さらに，検索された閲覧時アノテーションの一覧リストから，文

表 3.2: コメント属性

属性名	説明
アイディア	アイディアをコメントした場合に付与する属性
疑問点	疑問点をコメントした場合に付与する属性
問題点	問題点をコメントした場合に付与する属性
まとめ	まとめをコメントした場合に付与する属性
感想	感想をコメントした場合に付与する属性
言い換え	言い換えをコメントした場合に付与する属性
補足	補足する情報をコメントした場合に付与する属性
英語の和訳	英語の和訳をコメントした場合に付与する属性
用語説明	用語の意味をコメントした場合に付与する属性
その他	上記以外のコメントに付与する属性

章の流れや組み合わせを考えながら，文章を構成していくことができる．

さらに，前章で述べた引用の仕組みを考慮し，検索されたマーキングによって指定された部分要素をそのまま引用することを可能にする．閲覧時アノテーションの一つであるマーキングは文書の部分要素を特定する行為であるため，引用者は文書引用時にマーキング箇所を引用することで，被引用箇所を明示的に指定する必要がなくなる．さらに，引用するには必ず閲覧時アノテーションを付与する必要があるため，必ず一度は文書を閲覧する．そのため,「孫引き」のような不適切な引用を防止する効果があると考えられる．

3.2 閲覧時アノテーションを利用した副次的な応用

閲覧時アノテーションを利用することで可能になる応用は引用支援以外にも存在する．そのような副次的な応用として，閲覧時アノテーションの提示による文書の閲覧支援，閲覧時アノテーションの付与に基づく専門用語辞書の構築が挙げられる．

これらの応用は文書の閲覧を効率的かつ適切に行うように誘導することで，ユーザは短時間で正確に文書を理解することが可能になり，間接的に文章の意味解釈の誤りを原因とした誤引用を防ぐことに繋がる．そのため，閲覧時アノテーションはこれらの利用法においても，広義にユーザの文書の引用を支援していると言える．

3.2.1 文書の閲覧支援

まず，閲覧時アノテーションを他者と共有することで未読文書の閲覧支援が可能になることについて述べる．閲覧時アノテーションを共有し，他ユーザの付与した閲覧時アノテーションを文書本文と共に閲覧することで，未読文書の内容の理解が促進されると考えられる．例えば，同じ研究室の同じプロジェクトに属している二人のユーザが存在したとする．ユーザ同士は同じプロジェクトに属しているため，類似した背景知識を持っているだろう．そのため，同一文書中の同様の箇所を分かりにくいと感じたり，着目したりすることが予測される．そこで，一方のユーザが文書を閲覧する際に，分かりにくいと感じた用語や文章に後で読む時に理解し易くするようなマーキングをし，コメントとして解説文を記述しておく．すると，後で閲覧するユーザはその情報を見ることで容易に文書を理解することが可能になる．

しかし，一般に閲覧時アノテーションを共有する際の問題点として，ユーザ同士の背景知識や文脈が異なるためアノテーションの意味を互いに理解できないことが指摘されている．この問題点を解決し，背景知識が異なるユーザ同士がアノテーションの意味を正確に共有するためには，アノテーションにどのような種類の情報なのかということを明記しておく必要がある．そこで本研究では，閲覧者がどういう意図でマーキングやコメントを行ったのかという情報をアノテーションの属性情報として表示色と対応付けて取得・共有する．しかし，色はそのユーザの経験に基づき情報の種類を判別することに利用されることが知られている [13]．例えば，閲覧文書の著者にとってクリティカルな主張部分の記録に赤色を利用するユーザもいれば，黄色を利用するユーザもいるだろう．そのような視覚における直感的な部分の個人差を損なわないようにするために，マーキング色と属性情報は一対一の固定対応ではなく，ユーザが自由に指定することが可能であるべきである．つまり，共有すべきは属性情報であって色ではないということを踏まえてシステムの設計を行う必要がある．

松岡ら [15] は，論文のアブストラクトに対するマーキングシステムの実運用を行い，そのデータからマーキング箇所にはその論文における重要語を多く含むという結果を得ている．そのため，多くのユーザがマーキングしている箇所はその論文にとって重要な箇所である可能性が高い．つまり，マーキングを共有することによって得られる統計情報をユーザに提示することで，その論文の重要箇所を選択的に閲覧できる可能性があるだろう．

また，過去にユーザ自身が付与した閲覧時アノテーションを提示することで，既読文書の再閲覧を効率的に行うことが可能である．過去にユーザ自身が付与した閲覧時アノテーションには，過去に閲覧した際の文脈を想起させる効果がある．一般に，文書を読んでから自身の文書を書くまでには長期間経過していることが多いことを述べた．そのため，文書を書く際に過去に読んだことのある文書を再び

閲覧しても，文書の内容をなかなか理解できないことがある．そのような場合に，文書に付与されているマーキングやコメントは文書と共に閲覧することで，文書を読んだ際に考えていたこと等が想起され，内容を再理解する際の手掛かりとして利用される．例えば，閲覧時に重要箇所をマーキングにより記録しておくことで，再閲覧の際には文章を流し読みするだけで，大まかな内容を把握できるだろう．

3.2.2 専門用語辞書の構築

ユーザは特に意識することなく閲覧時アノテーションを付与していくことで，専門用語辞書が構築される．例えば，文書閲覧時に分からない専門用語が出現したら，その用語説明を文書にコメントすることがあるだろう．そのような場合に，コメントに用語説明属性を付与し，その用語の説明を記述することで，マーキング箇所とコメント属性から関連付けられているコメントはマーキング箇所に対する用語説明だということが機械的に判別可能である．その情報を利用することで用語と記述された用語説明は自動的に専門用語辞書の項目に追加登録できる．つまり，ユーザは文書上でコメントとして用語説明の付与を繰り返すことで，専門用語辞書を拡張していくことが可能である．また，コメントを付与する過程において，文章と専門用語辞書の項目を繋ぐリンクが取得されるため，用語説明の付与されている文章は辞書の用例として利用できる．

一方，構築された専門用語辞書は，文書の閲覧・作成において様々な用途に利用することができる．例えば，マーキング属性に辞書引き機能を対応付けておくことで，分からない専門用語をマーキングすると自動的に専門用語辞書の項目を引き，表示することが可能になる．また，文書作成時に利用することも考えられる．一般に，文書に専門用語を記述する場合，その用語の説明を記述するだろう．閲覧時に利用・拡張した専門用語辞書を利用することで容易に用語の説明を引用することが可能になる．用語説明を引用した際，論文と辞書項目にリンクを張ることで，ユーザ自身の言葉で記述するより，用語説明に客観性と正確性を持たせることができる．

3.3 引用支援システム

提案している引用意図を持つ部分引用を行うための引用支援システムについて述べる．本システムは，3.1 節，3.2 節で述べた機能を有する．本システムは，大きく分けて閲覧支援機構，作成支援機構から構成される．システムの構成について述べた後，各機構におけるインタフェースと主な機能について述べる．

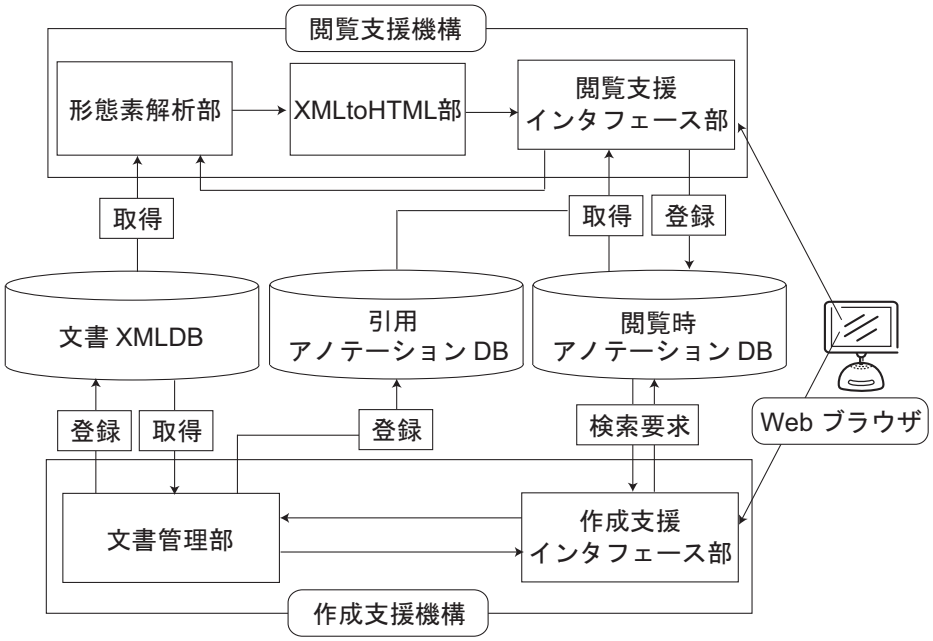


図 3.2: システム構成図

3.3.1 システム構成

本システムのシステム構成図を図3.2に示す．本システムはWebアプリケーションとして実装されており，専用のクライアントソフトを必要としない．ユーザはWebブラウザを利用してサーバにアクセスすることによって，文書の閲覧・作成などの操作を行うことができる．また，Webアプリケーションとして構築することで，不特定多数のユーザから閲覧時アノテーションや引用アノテーションを収集し，共有することが可能になる．ユーザは閲覧支援機構を通じて文書の閲覧と閲覧時アノテーションの付与を行い，作成支援機構を通じて文書の作成と引用アノテーションの付与を行う．

収集されたデータは閲覧時アノテーションデータベース，文書XMLデータベース，引用アノテーションデータベースの各データベースに保存される．データベースには，PostgreSQL 8.1を利用した．また，図3.2から分かるように，文書情報と引用情報は分けて管理される．前章でも述べたように，これは本研究では引用情報は引用・被引用文書の両文書へのアノテーション情報であるという考え方に基づいて設計されているからである．また，藤田 [16] が述べているWeb文書の信頼性や保存性のような問題点を考慮して，本システムでは文書情報の管理も行う．

3.3.2 閲覧支援機構

閲覧支援機構は，主に文書をユーザに提示し，その文書に対する閲覧時アノテーションを取得する機能を持つ．文書 XML と閲覧時アノテーションをユーザにブラウザ上で提示するまでの流れを以下に示す．

1. 形態素解析部により文書 XML に形態素タグを付与
2. XMLtoHTML 部により文書 XML を HTML に変換
3. 閲覧支援インタフェース部により HTML に閲覧時アノテーションを統合し表示

まず，本機構の形態素解析部では要求のあった文書 XML を取得し，文書 XML に含まれる文字列に対して形態素解析を行う．そして，形態素単位に分割し，言語構造に関する情報を形態素にタグとして埋め込む．このタグを本研究では形態素タグと呼ぶ．形態素解析器には，Sen[17] を利用した．Sen は Java で書かれた日本語形態素解析システムであり，MeCab[18] の Java 移植版である．形態素タグに分割することで，パラグラフ以下の詳細な部分要素に対するアノテーションを可能にすると同時に，形態素タグ以下の選択を不可にすることでマーキングにおける選択範囲の誤りを防止する．また，マーキング範囲を XPointer を用いて指し示すことが可能になり，閲覧時アノテーションの再現が可能になる．

次に，形態素タグの付与された文書 XML を XMLtoHTML 部で HTML に変換する．この際に，文書 XML における各要素の XPath を算出し，HTML タグの ID 属性に埋め込む．HTML における XPath ではなく，XML における XPath を利用するのは，オリジナルの文書に対するポインタを残しておくためである．つまり，スタイルシートには依存しない仕組みになっており，例えば，文書をページ単位に分割して一部分だけ表示するような場合にもポインタを一意に決定することが可能である．

最後に，変換された文書の HTML を閲覧支援インタフェース部が表示する．その文書に該当するアノテーションが存在する場合は，HTML 上に統合し表示する．ユーザがリアルタイムに付与する閲覧時アノテーションに関しては Javascript を用いて動的に表示する．付与されたアノテーションは XMLHttpRequest[20] を用いてサーバと HTTP 通信を行い，閲覧時アノテーションデータベースに登録する．登録する際に，マーキングの色をユーザプロファイル[1]を用いて対応付けている属性に変換し登録する．属性として保存することで，他のユーザとの閲覧時アノテーションの意味の共有を促進しており，他のユーザが該当するマーキングを閲覧する際には，ユーザ自身の対応付けている色で表示することが可能である．本シス

[1]ユーザプロファイルとは，マーキングの色と属性情報を関連付けているユーザ別の情報を指す．XML 形式で管理されており，ユーザは Web ブラウザを通じて変更することが可能である．

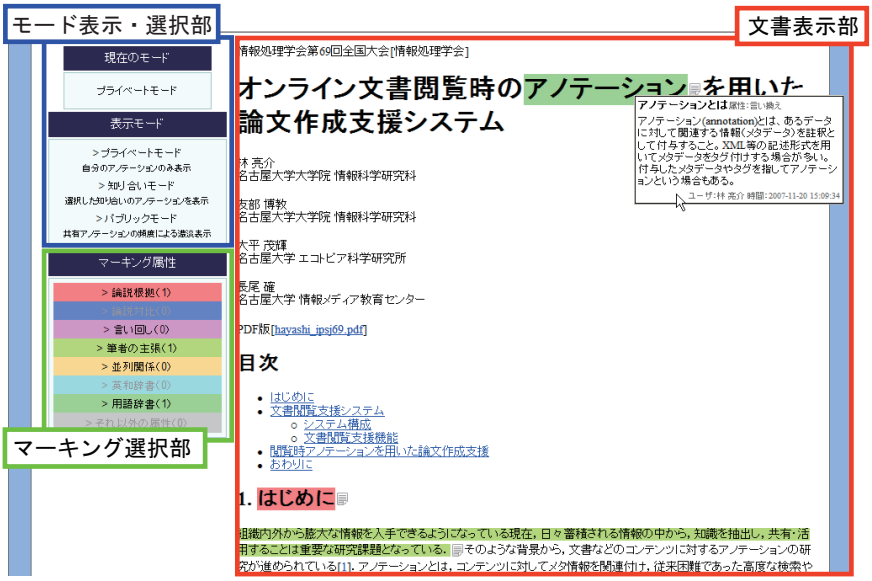


図 3.3: 閲覧支援機構のインタフェース

テムは，マーキングの色を共有するのではなく，マーキングの属性を共有する仕組みになっている．

また，閲覧支援インタフェース部は，文書が一定サイズより長い場合には，ページを分割して表示する．従来の文書の表示形式では，紙ベースで考える際の文字数等で形式的にページ分割を行っていたが，本システムでは意味のまとまりであるセクションを単位とした意味的なページ分割を行う．そのため，ユーザがページを前後して文書を読むような状況が減少すると考えられる．

本システムの，閲覧支援機構のユーザインタフェースを図 3.3 に示す．本インタフェースは以下に示す 3 つのコンポーネントから構成される．

- モード表示・選択部
- マーキング選択部
- 文書表示部

以下で，各コンポーネントの詳細について述べる．

モード表示・選択部

上側のウィンドウで現在のモードを表示し，下側のウィンドウで表示モードを選択することで文書表示部に表示するアノテーションを選択できる．表示モードには，プライベートモード，知り合いモード，パブリックモードの 3 種類が存在す

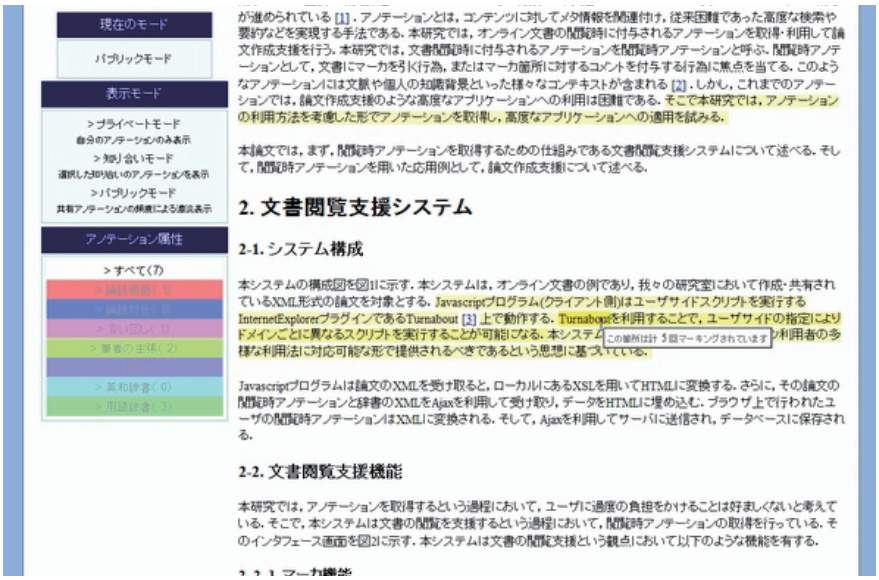


図 3.4: パブリックモード

る．適切にモードを切り替えることで効率的な閲覧が可能になる．例えばプライベートモードは，既読文書を閲覧する際の効率的な理解のために適しており，知り合いモードやパブリックモードは未読文書を閲覧する際の文章理解の支援になると考えられる．以下，各表示モードについて述べる．

- プライベートモード
  プライベートモードは，閲覧ユーザ自身の閲覧時アノテーションのみを表示するモードである．このモードのときに限り，右側の文書表示部で自由に閲覧時アノテーションを付与することが可能である．文書に閲覧時アノテーションを付与する方法については，文書表示部の項目で述べる．

- 知り合いモード
  知り合いモードは，表示したいユーザを指定することで，そのユーザの閲覧時アノテーションのみを表示するモードである[2]．このモードのときは，他者のマーキング情報を自分用にコピーすることができる右側の文書表示部に表示される他者のマーキング上で右クリックし，出現したメニューの「同意する」項目をクリックすることで，該当する他者のマーキングを自身のマーキングに取り入れることが可能である．

- パブリックモード

[2]本システムでは，登録されているユーザ情報を検索し，知り合いモードで閲覧したい他ユーザの ID をシステムに登録しておくことができる．また，相互に登録し合っているユーザの場合には，文書作成時に共著者を選択する際に利用することが可能である．

パブリックモードは，該当文書に付与・共有されているマーキングを頻度別に色を濃淡表示するモードである（図 3.4 参照）．最も多くユーザがマーキングしている箇所を最も濃い色でマーキング表示し，そこからマーキングしているユーザ数に応じて段々と薄い色のマーキングで表示する．

マーキング選択部

マーキング選択部は，右側の文書表示部に表示したいマーキングの属性を選択するためのコンポーネントである．

表示モードがプライベートモードと知り合いモードの場合には，マーキングの属性名をクリックすることで表示をオンとオフで指定することができる．例えば，「筆者の主張」属性[3]のマーキングのみを表示したいというユーザの要求にも対応することが可能である．オンの場合，属性名は黒文字で表示され，オフの場合は灰色文字で表示される．属性名の隣の括弧内の数字は文書に付与されている該当属性のマーキング数を示している．また，マーキング選択部に表示しているマーキング属性と色の対応関係は，文書表示部に表示されているマーキング属性と色の対応関係と同様である．

表示モードがパブリックモードの場合には，頻度別に表示したい個別の属性か，またはすべてのマーキングを選択することができる．文書表示部に頻度別表示されている項目は黒文字で，それ以外の項目は灰色文字で表示される．

文書表示部

文書表示部は，文書本文の表示，マーキング，コメントといった閲覧時アノテーションの表示を行う．また，ユーザは文書表示部上でマウス操作することで，閲覧時アノテーションを付与することが可能である．

まず，文書表示部において，マーキングを付与する手順を以下に示す．

1. マウスドラッグによりパラグラフ内の文字列を選択 (図 3.5 の 1)
2. 右クリックメニューの「選択文字列をマーキング」を選択 (図 3.5 の 2)
3. マーキング属性を選択し「決定する」ボタンをクリック (図 3.5 の 3)

付与されたマーキングは図 3.5 の 4 ように表示される．

本システムでは，マーキングに対して「公開」・「非公開」を設定することが可能である．「公開」にするとすべての表示モードに反映されるが，「非公開」にするとプライベートモードのみに反映される．

[3]「筆者の主張」属性とは，閲覧者が筆者の主張で重要だと思って行ったマーキングに付与する属性である．

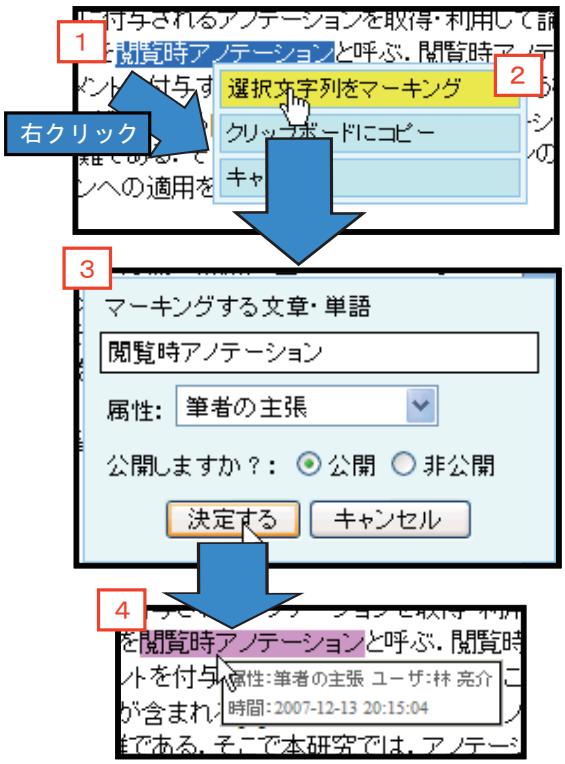


図 3.5: マーキングアノテーションの手順

また，オペレーションが対応付けられているマーキング属性の場合，マーキングが付与された時点でオペレーションが起動する．例えば,「英和辞書」属性の場合,自動的に goo 辞書 [19] の該当する用語ページが別ウィンドウで表示される．また,マーキングされた文字列上にマウスポインタを置くことでマーキングしたユーザ名，日時，属性情報がポップアップ表示される．

さらに，本システムではマーキング箇所に対してコメントを付与することが可能である．コメントを付与する手順は以下の通りである．

1. マーキングされた文字列上で右クリック (図 3.6 の 1)
2. 出現するメニューの「新規コメントの作成」を選択 (図 3.6 の 2)
3. タイトルと本文を記入，コメント属性を選択し「決定する」ボタンをクリック (図 3.6 の 3)

記述されたコメントはマーキングの横にアイコン表示され，アイコン上にマウスポインタを置くことで図 3.6 の 4 のようにコメント情報（タイトルと本文），属性情報，コメントしたユーザ名，日時がポップアップ表示される．また，ポップアップ表示されたウィンドウの色は属性情報に対応している．付与したコメント

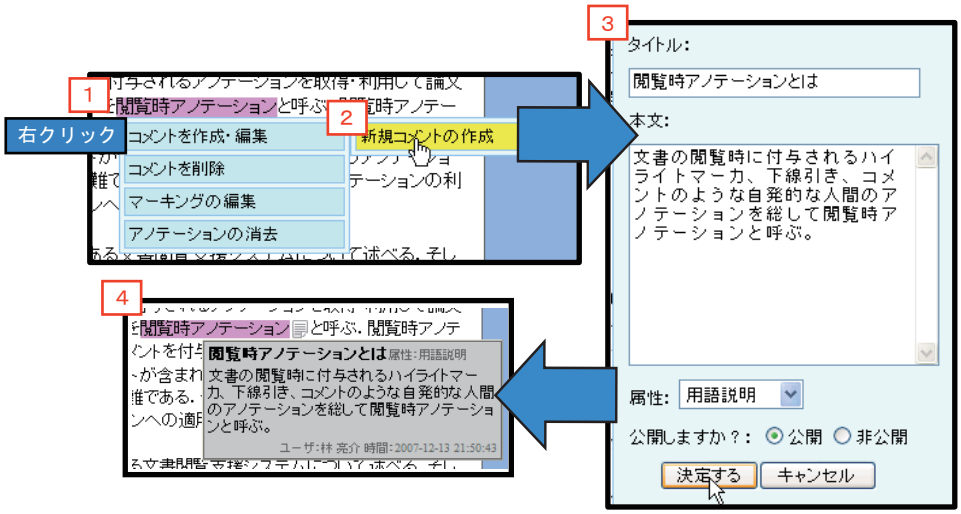


図 3.6: コメントアノテーションの手順

は編集・削除することができ，また，一つのマーキング箇所に複数のコメントを付与することも可能である．また，マーキングと同様にコメントに対しても「公開」・「非公開」を設定することが可能である．

3.3.3　作成支援機構

作成支援機構は，Web ドキュメントを作成・編集する機能を持つ．本機構は，Web ドキュメントの作成・編集にための以下のインタフェースを持つ．

- BasicWriter
- OutlineWriter
- ContentWriter
- Preview

BasicWriter はタイトルや著者情報といった文書の基本情報の編集，OutlineWriter は文書の章構成といったアウトラインの編集，ContentWriter はパラグラフや図の挿入，引用といった文書のセクション内部の編集を想定して実装されている．また，Preview は完成形のフォーマットに流し込んで，実際にどのように見えるのかをチェックするためのインタフェースである．以下では，BasicWriter，OutlineWriter，ContentWriter の各インタフェースの詳細について述べる．

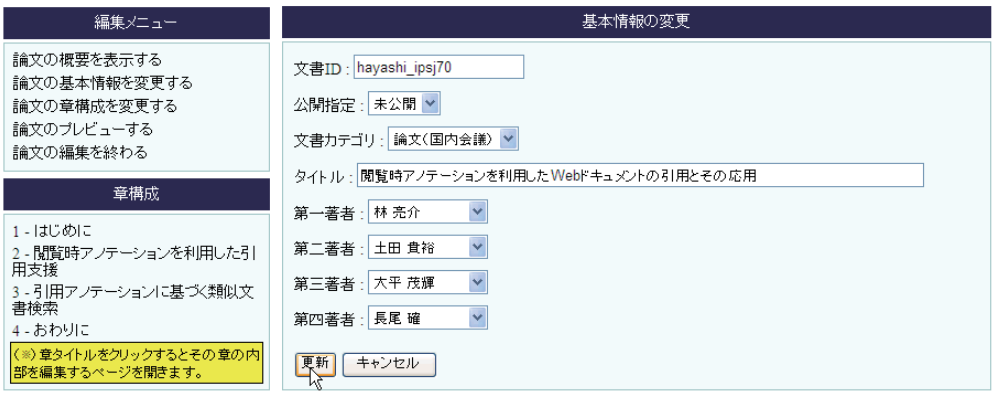


図 3.7: BasicWriter インタフェース

BasicWriter

BasicWriter を図 3.7 に示す．BasicWriter は文書に関する基本情報を編集するためのインタフェースである．以下に，BasicWriter において編集可能な項目を挙げる．

- 文書カテゴリ
- 公開指定
- タイトル
- 著者情報

文書カテゴリとは，Web ドキュメントがどのような文書（例えば論文など）なのかを表す情報である．文書カテゴリをプルダウン形式で選択することで，文書の出力フォーマットが決定する．また，公開指定は文書を他者に公開する・しないの 2 択の選択形式である．編集中の場合には，他者に公開しない「非公開」を選択する．本システムでは,「公開」を選択すると編集することができなくなる．本研究では，Web ドキュメントを一般の出版物と同様に捉えており，一度公開した文書を修正したい場合は文書自体を修正するのではなく，アノテーションを用いて訂正するべきだという考え方に基づいて設計している．タイトルはテキストフィールドによる自由記述形式で編集可能である．また，著者情報の編集はセレクトボックスのプルダウン形式で選択する形式になっている．セレクトボックスの選択候補には事前に登録したユーザ名が列挙される．作成者はユーザ名を選択するだけで，ユーザの所属や連絡先といった著者情報を自動的に文書に挿入することができる．

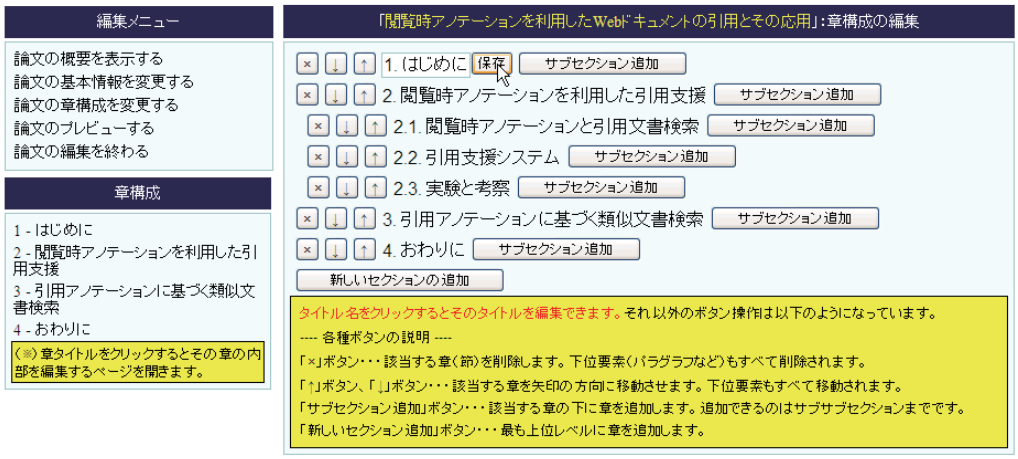


図 3.8: OutlineWriter インタフェース

OutlineWriter

OutlineWriter を図 3.8 に示す．OutlineWriter は章構成といった文書のアウトラインを編集することが可能であり，従来より提案されているアウトラインプロセッサ（Outline Processor）のような特徴を持つインタフェースである．OutlineWriter 上で可能な編集操作は，セクションやサブセクションといった章単位の追加・削除・入れ替えやセクションタイトルの編集である．

本システムは，論文のような長い文書を書くことを想定して設計されている．一般に長い文書を書く場合，大雑把な文書構成を決めてから，見出しを付けていき，ブロック（ノード）単位で細部についての記述を追加していくことが多い．そのため，OutlineWriter で文書全体のツリー構造を見ながら章構成を決めていき，構成要素の内部を編集する場合には後述する ContentWriter で記述する流れを想定している．

本インタフェースにおける編集操作の情報は Javascript+XMLHttpRequest による非同期通信でサーバに送信され，保存される．また，編集操作に応じてページ遷移しないため，Web 上で文書を編集する際に起こりがちなブラウザの「戻る」ボタンや「更新」ボタンによって編集情報が消えてしまうという問題に対するロバストネスが高められている．

ContentWriter

ContentWriter を図 3.9 に示す．ContentWriter はセクション単位で文書のセクション内部を編集するためのインタフェースである．具体的には，セクション内

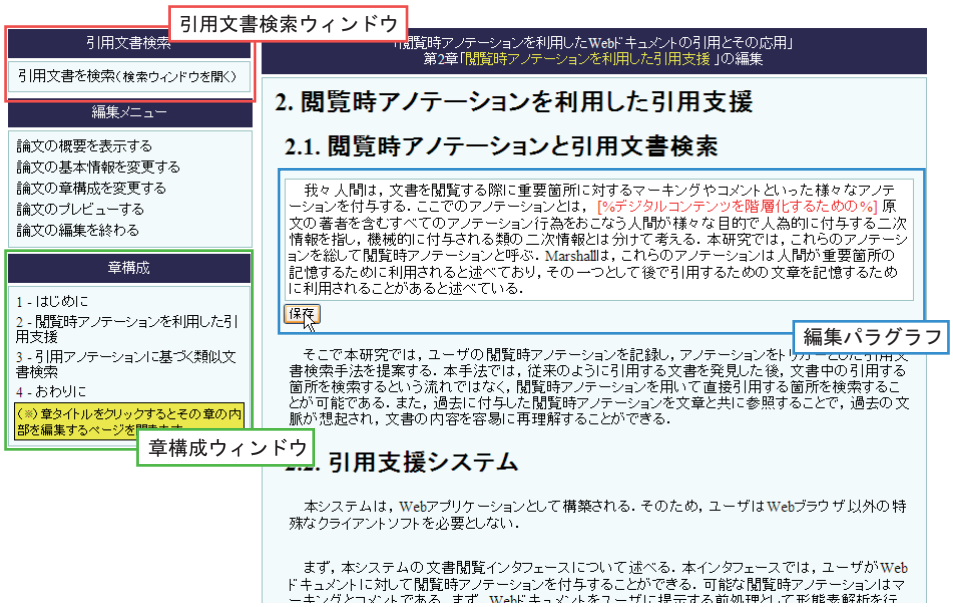


図 3.9: ContentWriter インタフェース

部のオブジェクト（パラグラフや図など）を追加・削除といった編集操作が可能である．また，本インタフェースにおいてユーザは文書の引用が可能である．

本インタフェースでは，簡単なマウス操作によりオブジェクトを追加・編集することが可能である．オブジェクトを追加する場合には，追加したい位置[4]で右クリックして追加メニューを表示させる．メニュー項目の中から，追加したいオブジェクトに項目を選択することでオブジェクトを追加できる．追加可能なオブジェクトは,「パラグラフ」,「リスト」,「図」などのセクションの内部要素である．

オブジェクトを編集したい場合には，そのオブジェクトをクリックすると編集可能になる．また，そのオブジェクト上で右クリックして，編集メニューを表示させることができ，メニューの中から，項目を選択することで対象オブジェクトの削除や移動といった編集が可能である．

また，ContentWriter 上では文書の引用が可能である．本システム上で行う引用の手順をフローチャートにて図 3.10 に示す．

まず，ContentWriter の左上の引用文書検索ウィンドウの「引用文書を検索」ボタンをクリックすると，図 3.11 のような過去に付与した閲覧時アノテーションのリストが出現する．閲覧時アノテーションをキーワードと属性で検索して，もし閲覧時アノテーションの中から引用文書の引用箇所を発見した場合は「この箇所を引用」ボタンをクリックする．そして，開いた引用ウィンドウ上で引用意図に関する属性を選択し,「引用する」ボタンをクリックすると引用文章がクリップボー

[4]本インタフェース上でマウスオーバーすると青色の横線が表示される位置．オブジェクトとオブジェクトの間など．

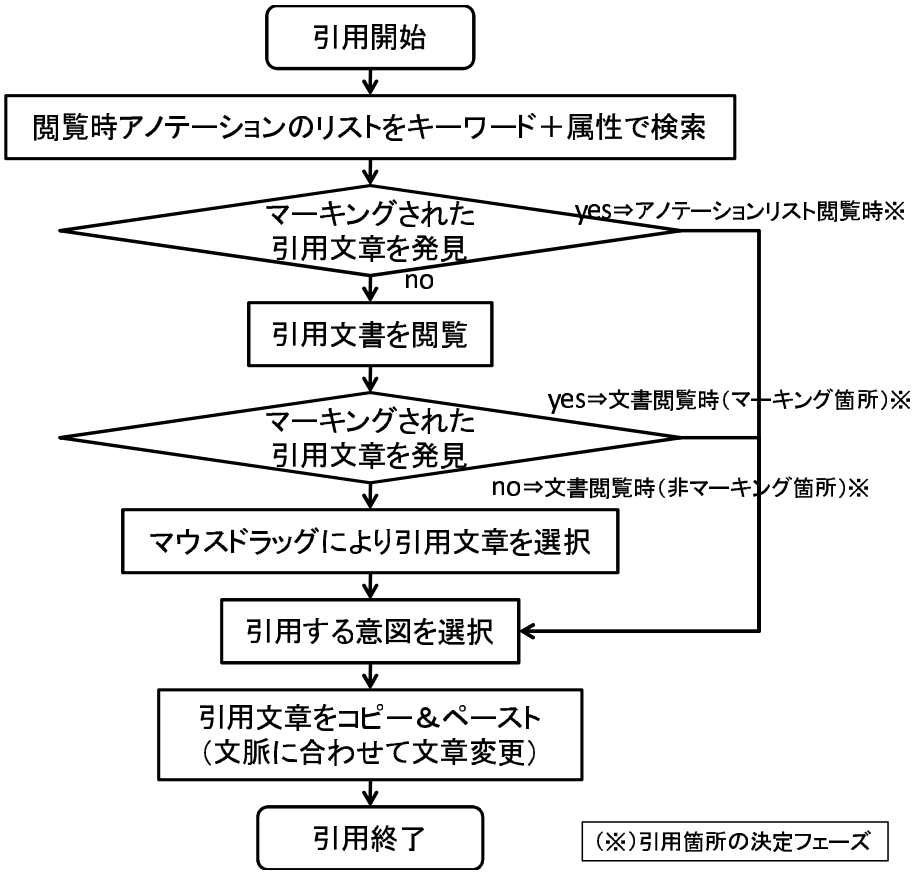


図 3.10: 引用の手順

ドにコピーされる．付与することの可能な引用意図に関する属性情報を表 3.3 に示す．

一方，閲覧時アノテーションの中に引用したい箇所が見つからなかった場合やマーキング箇所の前後の文脈に関する記憶が曖昧で引用箇所と断定しきれない場合は，再度アノテーションした文書にアクセスし閲覧することができる．文書を閲覧して，もし，マーキング箇所が引用したい箇所であった場合は，右クリックメニューから「マーキング箇所を引用」を選択する．すると，引用ウィンドウが開くので，引用意図に関する属性を選択し,「引用する」ボタンをクリックすると引用文章がクリップボードにコピーされる．また，マーキング箇所以外の文章を引用したい場合は，マウスドラッグにより引用したい箇所を選択し，右クリックメニューから「選択箇所を引用」を選択する．あとは同様に引用意図に関する属性を選択し,「引用する」ボタンをクリックすると引用文章がクリップボードにコピーされる．

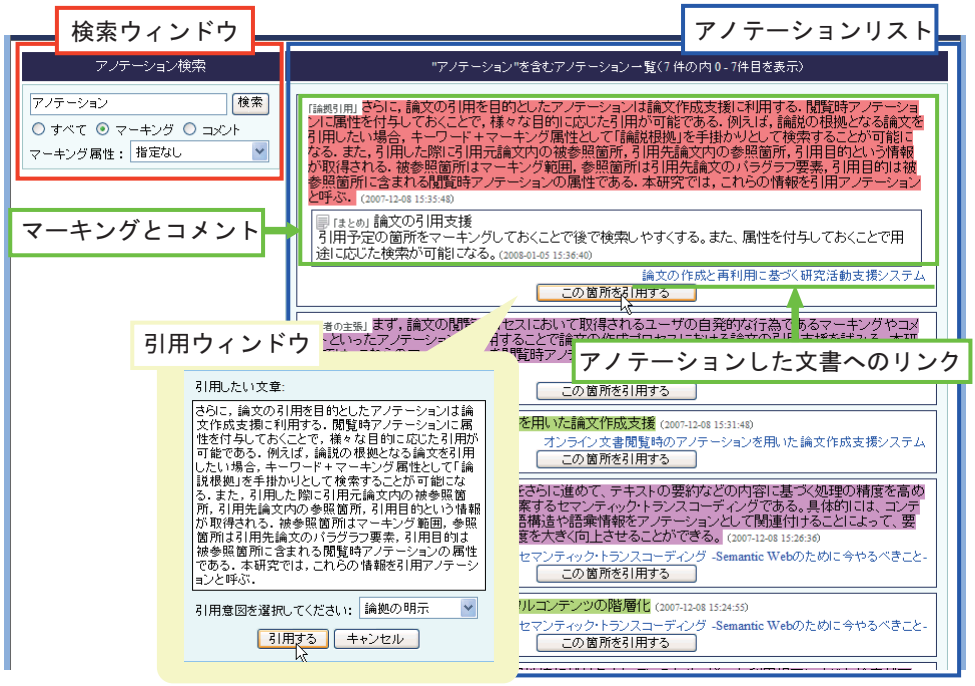


図 3.11: 閲覧時アノテーションのリスト表示

クリップボードにコピーされた文章を ContentWriter 上の引用したい箇所にペーストすることで引用が完了する．ContentWriter において，引用した文章は赤色の文字列で表示され，その部分を編集することで引用した文章を自身の文書の文脈に合わせることが可能である．

3.4 まとめ

前章では，Web ドキュメントの新しい引用の仕組みを提案した．そこで，本章では実際に提案する引用を行うために構築したシステムについて述べた．また，本システムは単に提案する引用を可能にするだけではなく，いくつかの引用支援機能を有することを述べた．

まず，文書作成時において引用すべき文書を検索する際の問題点について考察し，その解決策として閲覧時アノテーションと呼ばれるユーザの自発的な行為から得られる主観的な情報を用いた引用文書検索手法を提案した．また，閲覧時アノテーションは引用文書検索以外にも，様々な応用に利用することが可能になることを述べた．最後に，引用意図を持つ部分引用を行うための引用支援システムについて，システム構成について説明した後，閲覧支援機構，作成支援機構におけるインタフェースと主な機能について言及した．

表 3.3: 引用属性

属性名	説明
論拠の明示	自己の主張の論説の根拠を示すための引用
論説の対比	自己の主張と他者の主張を対比するための引用
問題点の指摘	他者の主張の問題点を指摘するための引用
先行研究の例示	先行研究を例示するための引用
その他	上記した以外の目的の引用

第4章　引用アノテーションを利用した類似文書検索

前章では，閲覧時アノテーションを利用することで可能になる文書の引用支援について述べ，引用支援システムを提案した．本章では，引用支援システムを運用することで得られる引用アノテーションを利用した応用について述べる．

まず，4.1 節で本システムを運用することで蓄積される引用アノテーションは，従来の引用情報に比べて有用性の高い情報を含むことを述べる．また，引用アノテーションが蓄積されることで得られる文書ネットワークについて言及する．4.2 節で，本システムにおける類似文書検索と呼ばれる検索の仕組みの必要性について考察した後，4.3 節では，引用アノテーションを利用して，従来より提案されている共引用と呼ばれる類似度指標を詳細に扱う方法を提案する．さらに，4.4 節では，引用支援システムに組み込む類似文書検索機構について述べる．

4.1　引用アノテーションの有用性

前章で述べた引用支援システムを利用してユーザが文書の引用を行うことで，引用アノテーションが取得される．引用アノテーションには，引用先文書の引用箇所情報，引用元文書の被引用箇所情報，引用者の引用意図に関する属性情報などが含まれる．引用アノテーションは図 4.1 のように XML 形式で機械的に処理可能な形で管理される．上記した引用情報以外にも，XML には引用したユーザの ID 情報や時間情報が含まれる．

引用アノテーションは，引用先・引用元文書の関係を部分要素レベルで保持している．そのため，文書の閲覧者は容易に引用の意図を理解することが可能になる．例えば，閲覧者が文書閲覧中に引用されている文書に対して興味を持ったとする．現在では，まず「参考文献」のような巻末に付与されている引用文献リストから出典を確認し，Web ドキュメントであれば URL にアクセスし，論文であれば掲載されている学会誌などにアクセスし文書をダウンロードするだろう．そして，引用元文書を閲覧し，引用あるいは参照されている部分を探す必要があるだろう．しかし，本手法では引用先文書の引用箇所から引用元文書の被引用箇所に対して直接的にハイパーリンクが張られているため，そのような煩雑な作業をす

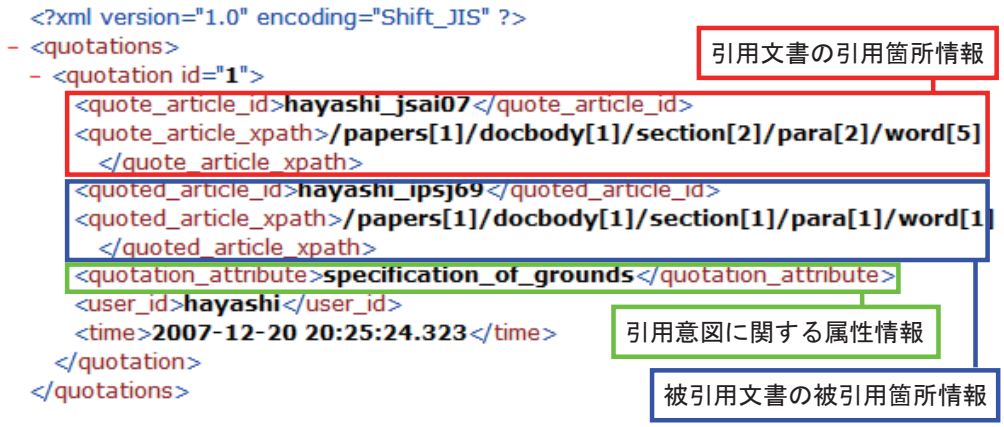


図 4.1: 引用アノテーション XML

る必要はない．さらに，引用意図に関する属性情報を参照することで，引用元文書にアクセスすることなく，どのような意図により行われた引用であるか把握することが可能である．

また，引用アノテーションは文書の部分要素間を双方向に繋ぐ属性付きリンクと捉えることが可能であり，この双方向リンクを用いることでユーザの文書のサーベイ活動を支援することができる．例えば，ある研究分野において重要とされる古典論文を閲覧したとする．その際，引用アノテーションにより「この箇所を引用して書かれた論文が存在する」という情報およびその引用意図が文書の部分要素のメタ情報として表示される．つまり，通常引用元文書の方向へ向かってサーベイを行うが，ユーザはこのリンクを辿ることで引用先文書の方向へ向かう関連文書のサーベイを行うことが可能になる．また，引用元文書の著者は自身の文書がどのように引用されているのかトレースすることが可能である．つまり，自身の文書の一部を誰が，いつ，どのような意図で引用しているか確認することができる．もし，引用先文書の著者の考えに興味を持った場合には，引用先文書の著者とコンタクトをとり，互いに意見を交換することで，問題に対する理解をさらに深めることができるだろう．

引用アノテーションが蓄積されると，図 4.2 のような文書の部分要素間を引用意図に関する属性によって関連付けられた Web ドキュメントのネットワークが構築される[1]．このネットワークは，従来の引用情報に基づくネットワークに比べて，多くの意味的な情報を内包している．例えば，互いの Web ドキュメントの部分要素間の関係として引用の構造が分かるため，文書間の関係の推定などに文章の文脈情報を利用することが可能になる．また，従来の引用情報は参照先を指し示す以外の情報を持たないが，このネットワークのリンクは引用意図に関する属性情

[1]図 4.2 では，引用意図に関する属性を矢印の線種で表している．

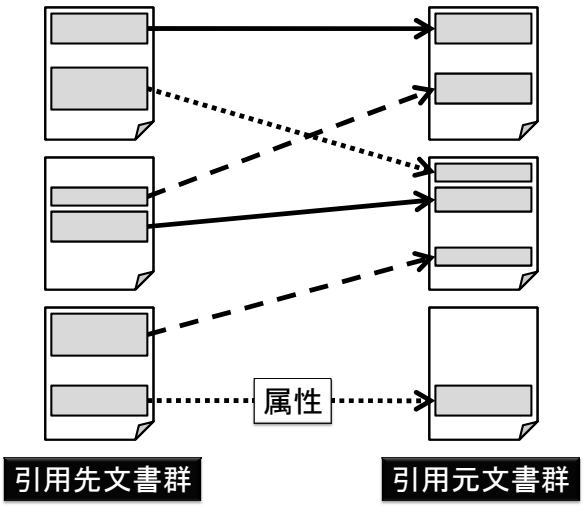


図 4.2: 引用アノテーションに基づく文書ネットワーク

報を保持している．そのため，このネットワークを利用することで従来より意味的に高度な応用を実現することが可能になるだろう．

4.2 類似文書検索の必要性

前章で提案した引用支援システムは，既読文書を効率的に検索し，引用することが可能な仕組みである．しかし，ユーザが引用すべき文書を過去に閲覧しているかどうかは定かではなく，引用すべき文書を見逃している可能性がある．そのため，ユーザの要求に合わせて適切に閲覧文書を検索する仕組みが必要である．

現在，Web ドキュメントの検索システムとして様々なものが実現されているが，主にキーワードをクエリとして検索結果を絞り込む類のものが多い [5][21][22]．キーワードにより絞り込むシステムの場合，対象が論文のような専門的な文書であるほど絞り込みキーワードを必要とし，ユーザに専門領域に関する知識が少ない場合には多くの負担を強いる．また，分野によって同じ概念を違う語で表記されることがあり，そのような違いをユーザが考慮することは困難である．

そのような背景から，ユーザのキーワード入力による負担を軽減するために，類似している文書を提示し，ユーザの検索の支援を行う試みがなされている．類似文書をユーザに提示することで，絞り込みの手間を省くことや，適切な文書であるが絞り込みキーワードが含まれていないために除外されてしまうような文書の発見が可能となる．類似文書検索における文書間類似度の算出では，従来より提案されており，ある程度の精度が示されているベクトル空間型モデル [23]（Vector

Space Model）を利用するのが一般的である．しかし，ベクトル空間型モデルでは，文書に含まれる単語単位のマッチングにより類似度を算出するため文書の表層的な情報しか考慮することができない．また，文章を単語単位に分割して処理するため，文章の文脈情報を考慮することができないといった問題点がある．

そこで本研究では，引用情報を利用した類似文書検索手法に着目し，引用アノテーションに含まれる意味的な引用情報に基づく類似文書検索手法を提案する．引用は人間が文書の意味を考慮して文書間を関連付ける行為であり，ベクトル空間型モデルでは考慮されていない文書間の関係を取得できる可能性がある．また，引用アノテーションに基づく類似文書検索の仕組みを適用することで，ユーザが従来では発見できなかったような文書の引用を促進する．そのため，さらに多くの有益な引用アノテーションが収集され，類似文書検索の検索精度が高まるというポジティブなフィードバックサイクルが発生するだろう．

4.3　引用アノテーションを利用した共引用に基づく文書間類似度

2.4 節では，引用分析の分野において共引用と呼ばれる文書間の類似尺度が提案されていることを述べた．また，従来の共引用による類似度の算出では，共引用の関係にある文書同士の類似度は全て同じとされており，どのような意味において文書同士は共引用の関係にあるのかということは考慮されていないことについて言及した．

そこで本節では，引用アノテーションを適用することで，引用に関する意味的な情報を考慮した，共引用に基づく類似文書検索手法を提案する．具体的には，共引用の関係に基づく文書間類似度を，共引用の関係にある引用同士の関係に応じて詳細に重み付けする手法を提案する．この手法では、以下の 2 つの指標を用いる

- 引用意図の関係に基づく重み付け
  文書がどのような意図によって引用されたかによって共引用の関係は異なると考え，引用意図の関係によって共引用を分類し重み付けする．例えば，自己の論説の根拠を示すために行われる引用と他者の主張の問題点を指摘するために行われる引用では引用意図が異なり，それに伴い共引用の関係も変わってくる．本研究では，同様の引用意図を持つ引用同士によって共引用の関係にある文書間には，引用意図の異なる引用同士による共引用の関係にある文書間に比べて，文書内容が類似する可能性が高いという仮説を立てる．例えば，関連研究を述べる際に，類似システムについて述べられている文書を複数引用して自身のシステムと比較することがあるが，引用されている文書同士は類似している可能性が高いだろう．この場合，どの文書を引用した

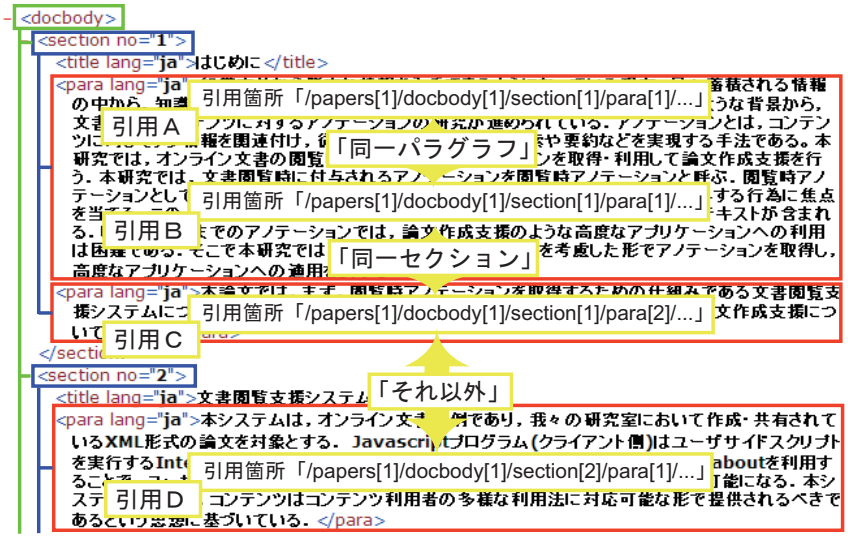


図 4.3: 引用箇所間の距離

意図も同じであり，システム同士を対比するためである[2]．そこで，本システムに蓄積された引用意図に関する属性情報である「論説の根拠」,「論説の対比」,「問題点の指摘」,「先行研究の例示」の 4 種類に関して，共引用に利用される引用同士が同じか異なるかで分類し，重み付け．もし，引用意図が同じであれば異なる場合に比べて高い重みを与える．

- 引用箇所間の距離に基づく重み付け
  引用文書の文脈情報を利用して引用箇所間の距離を算出し，距離に応じて共引用の関係を分類し重み付けする．通常，文書にはセクションやパラグラフといったまとまりが存在しており，そのまとまりに応じて記述される意味的な内容が異なる．そのため，引用箇所間の距離が近いほど引用されている文書同士は意味的に近い内容である可能性が高い．例えば，引用文書の冒頭の「はじめに」で引用された文書と末尾の「おわりに」で引用された文書の関係と,「関連研究」について述べられている部分の同一パラグラフ内で引用された文書同士の関係は異なるだろう．そこで，本手法によって提案している文書 XML の構造情報を利用して，引用先文書における引用箇所間の距離を考慮する．共引用に利用される引用が行われている引用箇所間の距離を「同一パラグラフ」,「同一セクション」,「その他」に分類する．例えば，図 4.3 のような引用 A，引用 B，引用 C，引用 D の引用箇所の XPath が図の通りであった場合，引用 A と引用 B の関係は「同一パラグラフ」，引用 B と引用 C は「同一セクション」，引用 C と引用 D は「それ以外」に分類される．本

[2]本システムにおいてシステム同士を対比するために引用に付与される属性は「論説の対比」である．

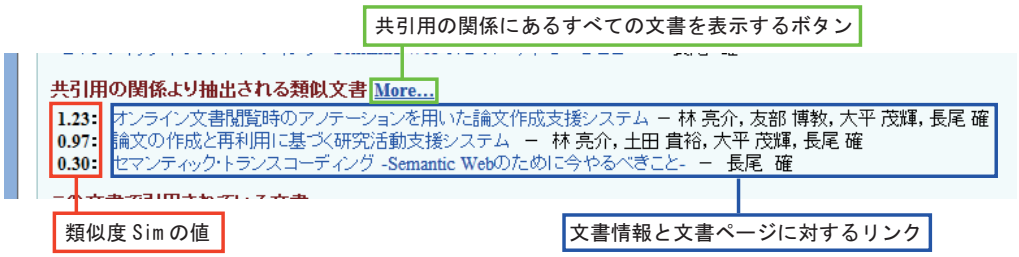


図 4.4: 類似文書の提示

手法では,「同一パラグラフ」,「同一セクション」,「それ以外」の順に高い重みを与える.

上記の 2 つの指標に基づき共引用の関係を重み付けし，類似文書検索に利用する文書 D_1 と D_2 の類似度 $Sim(D1, D2)$ の算出式を以下に示す.

$$Sim(D_1, D_2) = \sum_{i=1}^{N} CocitedW_i \tag{4.1}$$

$$CocitedW_i = DistanceW_i * IntentW_i \tag{4.2}$$

式 (4.1) における $Sim(D_1, D_2)$ は，共引用の関係に応じて与えられる重み $CocitedW_i$ の総和であり，N は文書 D_1 と D_2 が共引用されている回数である．そのため，$CocitedW_i$ の値が大きく，かつ，共引用されている回数 N が大きいほど文書同士の類似度が高くなる.

また，式 (4.2) における $DistanceW_i$ は共引用の関係にある引用箇所間の距離を考慮した値で，最小値 0 から最大値 1 の間で引用箇所間の距離に応じてヒューリスティックに与えられる．また，$IntentW_i$ は共引用の関係にある引用同士の引用意図に関する属性が同じなら高く，異なるなら低くなるように，最小値 0 から最大値 1 の間でヒューリスティックに与えられる重みである．$CocitedW_i$ を利用することで，引用の引用意図に関する属性情報と引用箇所間の距離を考慮し，共引用に基づく文書間類似度を詳細に重み付けすることが可能である.

4.4　類似文書検索機構

類似文書検索機構は，引用支援システムに組み込まれるモジュールである．本機構では，前節で提案した類似度 $Sim(D_1, D_2)$ の値に応じて，類似文書をランキングして表示する.

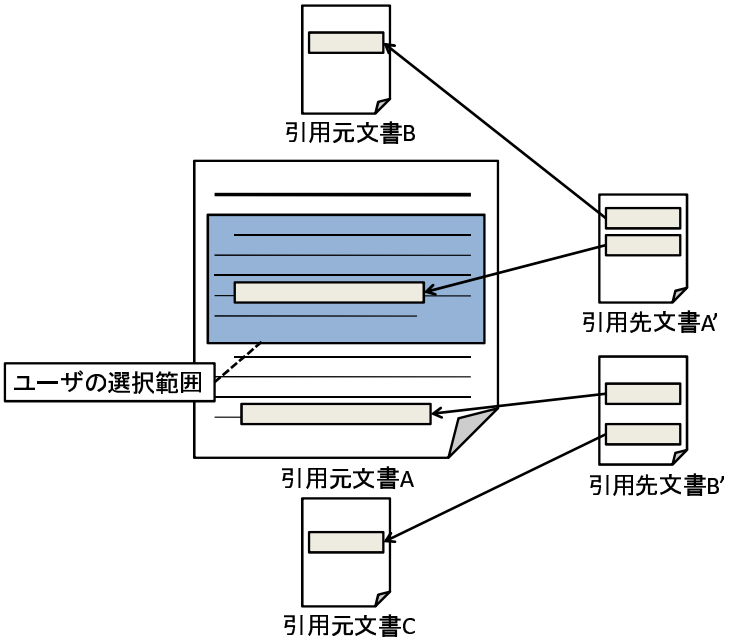


図 4.5: 類似度算出に利用される共引用

引用支援システムにおいて類似文書を検索する方法は二通りある．まず，一つ目の方法は，文書の概要ページ[3]において類似文書を検索する方法である．図 4.4 のように，文書の概要ページにおいて，類似度の高い順に最大 3 本の文書情報が表示される．表示される文書情報は該当する文書のタイトルと著者名，文書へのリンク，算出された類似度の値である．「More…」と表示されている部分をクリックすると共引用の関係にある文書情報がすべて表示される．また，文書へのリンクの上にマウスカーソルを置くと，類似度の算出に利用された共引用している文書に関する情報として，文書のタイトルと著者名，文書に対するリンク，$DistantceW_i$ の値，$IntentW_i$ の値がポップアップして表示される．

二つ目の方法は，文書閲覧中に類似文書を検索する方法である．マーキングをするのと同様に範囲選択し，右クリックメニューから「類似文書検索」をクリックすることで類似文書がポップアップウィンドウにてユーザに提示される．提示される情報は，上記した一つ目の方法と同じであり，該当する文書のタイトルと著者名，文書へのリンク，算出された類似度の値である．但し，この検索方法の場合は類似度の算出方法が一つ目の方法と異なる．この検索方法の場合，選択された範囲に被引用箇所が含まれる引用を用いた共引用のみで類似度が算出される．も

[3]概要ページとは，ユーザが文書を閲覧する際に，本文にアクセスする前にアクセスするページである．概要ページには，該当する文書のタイトル，著者情報，付与されている閲覧時アノテーション情報，参考文献リスト，該当文書を引用している文献リストなどが表示されており，ユーザは概要ページを閲覧することで文書の内容をある程度把握できる．

し，選択された範囲に被引用箇所が存在しない場合は，類似文書は存在しないとされ，右クリックメニューに「類似文書検索」項目は表示されないようになっている．例えば，図 4.5 のような引用関係があったとする．引用先文書 A’ により引用元文書 A と引用元文書 B は共引用の関係であり，引用先文書 B’ により引用元文書 A と引用元文書 C も共引用の関係である．一つ目の方法の場合は，この 2 つの共引用を用いて類似度が算出され，ランキング表示される．しかし，二つ目の方法の場合は選択範囲に被引用箇所がある引用先文書 A’ による共引用しか利用されない．つまり，結果として類似文書として提示されるのは引用元文書 B のみである．

また，範囲選択だけでなく，セクションやサブセクションのような文書の構成単位を選択して，類似文書検索することも可能である．セクションのような文書の構成単位を選択する場合には，文字列を選択しない状態で文書表示部で右クリックすると，閲覧している文書の章構成が木構造で表示される．表示された木構造からセクションやサブセクションを選択することで，類似文書がユーザに提示される．類似度の算出方法は範囲選択する場合と同様で，選択されたセクションやサブセクションに被引用箇所が含まれる引用を用いた共引用のみで類似度が算出される．また，ユーザに提示される類似文書に関する情報も同様である．

二つ目の方法のようにユーザの選択した範囲に含まれる引用のみを利用することで，ユーザの要求に適した類似文書のみを提示することが可能になる．例えば，ユーザが文書のシステムの説明部分を選択したならば，ユーザはその文書の研究背景や目的のような思想的な主張に興味があるのではなく，実装されたシステムに興味があると推測される．そのため，思想的な主張が類似している文書を提示するのではなく，実装されたシステムが類似している文書を提示する必要があるだろう．類似システムについて述べているような文書の場合，文書のシステムの説明部分を引用している可能性が高い．共引用の関係にあるもう片側の引用元文書にも同様のことが言えるため，選択された箇所に含まれる引用のみを利用することで，ユーザの要求に適した文書のみを提示することが可能である．さらに，提示された文書にアクセスした場合には，類似度算出に利用された引用の被引用箇所がハイライト表示される．つまり，文書閲覧中に類似文書検索を行うトリガーとなった被引用箇所と関連している箇所をユーザは知ることが可能である．二つ目の類似文書検索の仕組みは，一つ目の方法では考慮できないユーザの詳細な要求に対応することが可能な仕組みであると言える．

4.5　まとめ

本章では，引用支援システムを運用することで得られる引用アノテーションの応用例を提案した．

引用支援システムを運用することで蓄積される引用アノテーションは，ユーザにとって従来の引用情報に比べ有用性の高い情報を含んでいた．また，引用アノ

テーションが蓄積されることで得られる文書ネットワークは従来の引用情報に比べて意味的に高度な応用を実現できることを述べた．次に，本システムにおける類似文書検索の必要性について言及した．さらに，引用に関する意味的な情報を考慮して共引用の関係を詳細に重み付けする手法を提案した．本手法は，引用意図の関係と引用先文書の引用箇所間の距離を利用して文書間類似度を算出する手法であった．最後に，前章で提案した引用支援システムに組み込まれる類似文書検索機構について述べた．引用支援システムにおいて，ユーザは二種類の類似文書検索を行うことが可能であり，文書閲覧中の類似文書検索の仕組みはユーザの詳細な要求に対応できる仕組みであった．

第5章　実験と考察

本章では，まず，3 章で提案した，引用支援システムにおける閲覧時アノテーションを用いた Web ドキュメントの引用支援手法の有効性に関する実験について述べる．次に，その実験において得られた引用情報を用いて，4 章で提案した共引用に基づく類似文書検索手法の有効性に関して行った実験について報告する．

5.1　引用支援に関する実験

3 章で提案した閲覧時アノテーションを用いた Web ドキュメントの引用支援システムの有効性に検証するために実験を行った．以下，まず実験方法について述べ，さらに実験結果と考察について述べる．

5.1.1　実験方法

閲覧時アノテーションを用いた引用支援の有効性を示すために被験者実験を行った．被験者を A と B の 2 つのグループに分け，以下の 2 通りのやり方で文書を作成してもらった．A グループは，文書を閲覧する際にアノテーションを付与して，文書を作成する際の引用文書検索に利用して文書を引用してもらった．B グループは，文書を閲覧する際にアノテーションを行わずに，文書を作成する際に適切な文書を引用してもらった．また，本実験では研究活動における論文の執筆を想定して実験を行った．

本実験に参加してもらった被験者は情報科学に関する専門的な知識を持っている本研究室の人であり，参加した被験者の延べ人数は 16 人であった．本実験において被験者に提示した Web ドキュメントは，本研究室で公開している情報科学に関する専門的な日本語論文[1]であり，その総数は 57 本であった．

本実験においてユーザが文書を引用する手順を図 5.1 に示す．まず，最初に書くべき文書の内容をある程度考えてもらうのは，研究活動において特定の研究テーマがある状態で文書をサーベイするのが一般的であるという理由に基づいている．また，文書を閲覧してから作成するまでに一定期間空けて実験を行う．これも，関

[1] http://www.nagao.nuie.nagoya-u.ac.jp/papers/papers.xml

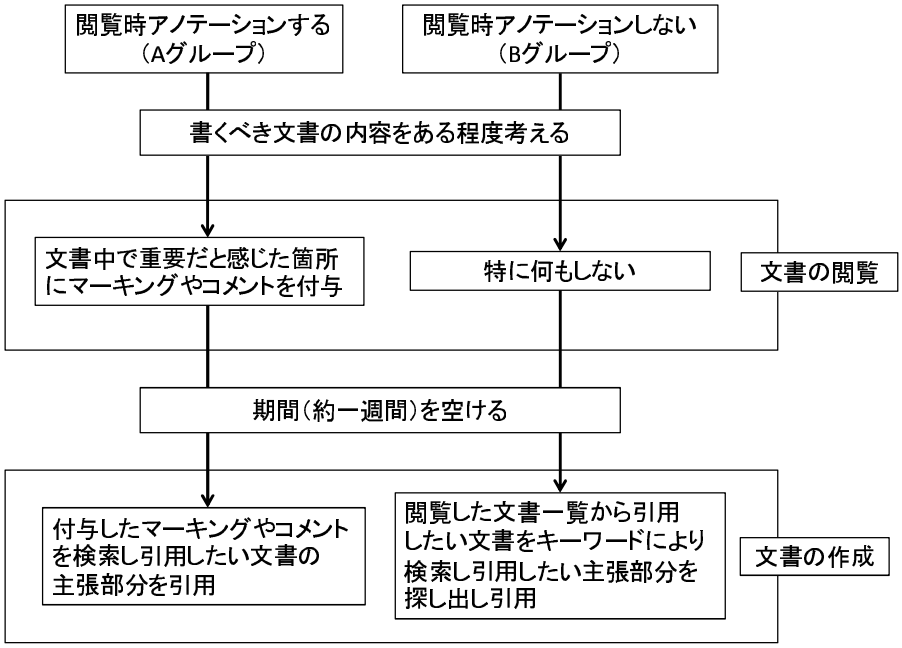


図 5.1: 実験の手順

連研究をサーベイしてから論文を執筆するまでには，ある程度の期間が経過していることが一般的であるという研究活動を再現するためである．

また，本実験で引用対象としている文書は過去に一度は閲覧したことのある文書であり，過去に閲覧したことのある文書を検索・引用するような状況を想定している．例えば，文書を書いている際に引用文書が少ないと感じて新たな文書を検索・引用することがあるが，そのような閲覧したことのない文書の引用は今回は対象としない．そのため，A グループの場合はアノテーションをした文書のみを引用文書検索の対象とし，B グループの場合は閲覧履歴を用いて閲覧した文書のみを引用文書検索の対象とした．

引用支援手法の評価方法について述べる．まず，前途の二つのグループにおける引用コストを比較する．引用コストは，一回当たりの引用文書検索に費やした時間とした．つまり，引用文書検索に費やした時間が短いほど引用コストが小さく，引用支援が有効に活かされていると考えた．次に，閲覧時アノテーションを付与するのに費やした時間と閲覧時アノテーションを付与することで短縮できた時間の関係を考察する．閲覧する目的が明確な場合は，閲覧時アノテーションは自然な行為であるが，本システムにおいてアノテーション属性を選択・付与することはユーザに負担を強いる．そのため，閲覧時アノテーションの付与に費やした時間と短縮できた時間は費用対効果の関係にあるとして考察する．

5.1.2 実験結果と考察

表 5.1: グループ別データ

	作成された文書数	引用文書検索回数	引用数
A グループ	10	59	55
B グループ	6	26	21
合計	16	85	76

本実験において作成された文書数と，行われた引用文書検索回数と引用数をグループ別に表 5.1 に示す．引用文書検索回数と引用数が異なるのは，一度引用しようとして検索したが，最終的には引用しなかった文書が存在したためである．本実験では，最終的には引用しなかった文書の検索時間もデータとして利用した．本実験においては，文書 1 本を作成するにあたり，平均すると約 5.3 回の引用文書検索が行われた．

表 5.2: 閲覧時アノテーションに関するデータ

	マーキング	コメント	合計
アノテーション数	372	26	398
平均時間（秒）	5.3	65.1	9.3

また，本実験において付与された閲覧時アノテーションに関するデータを表 5.2 に示す．アノテーション数とは，今回の実験において行われた全ユーザの閲覧時アノテーションの合計数であり，平均時間とはユーザが閲覧時アノテーションの付与に費やした平均単位時間である．本実験では，マーキングに比べてコメントの付与された数は少なかった．これは，今回のような短期間の実験では，マーキングにより引用箇所を記録し閲覧するだけで文書内容が想起することができたためと考えられる．また，マーキングが付与されていた箇所の内，引用された箇所は約 14%であり，コメントが付与されていた箇所の場合は約 15%とほとんど違いがなかった．しかし,「アイディア」属性のコメントが付与されていた箇所は 7 箇所の内 4 箇所，つまり約 57%が引用されており，非常に強い関連性が見られた．そのため，閲覧時アノテーションのリストを表示する際に「アイディア」属性のコメントが付与されている箇所を上位に表示するといったように，ランキング方法を今後改良していく必要があるだろう．

A グループと B グループの引用コストを比較した結果を図 5.2 に示す．A グループは B グループに比べて，一回当たりの引用文書検索に費やした時間は 101.3 秒

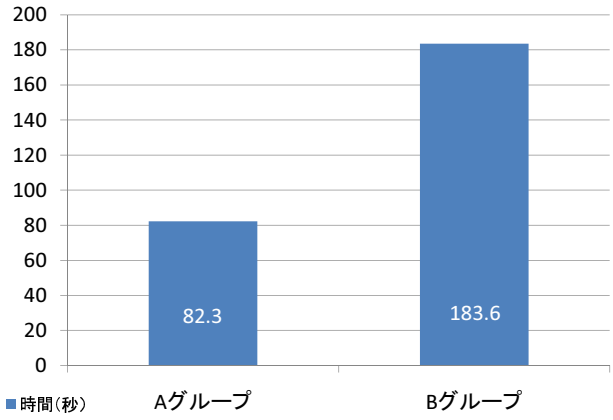


図 5.2: 引用コストの比較

短く，B グループの検索時間の半分以下であった．以上の結果から，引用コストという観点から本手法の優位性を示すことができた．

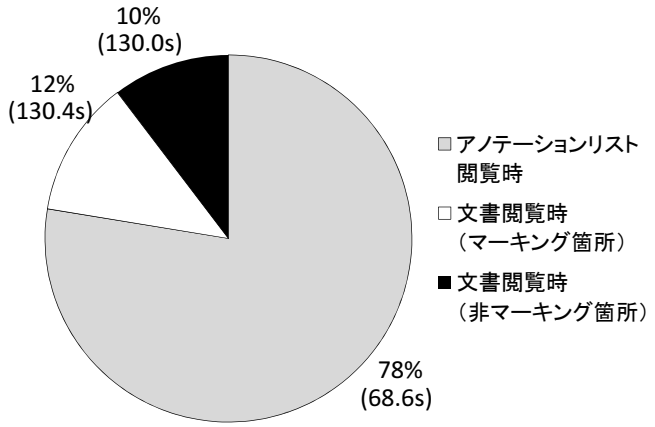


図 5.3: 引用箇所の決定フェーズの割合

A グループの引用コストについて詳細に見ていくと，ユーザの引用文書検索時における引用箇所の決定フェーズによって平均検索時間に偏りがあることが分かった．ここで議論する引用箇所の決定フェーズは，図 3.10 のフローチャートにおける 3 つの引用箇所の決定フェーズに対応している．引用箇所の決定フェーズの割合とそのフェーズにおける平均検索時間を図 5.3 のグラフで示す.「文書閲覧時（マー

キング箇所)」と「文書閲覧時（非マーキング箇所)」に引用箇所を決定した場合の平均検索時間が 130.4 秒と 130.0 秒であり，B グループの平均検索時間から 50 秒程度しか短縮できていないが,「アノテーションリスト閲覧時」に引用箇所を決定した場合の平均検索時間の平均値は 68.6 秒であり，B グループの平均検索時間と比べ約 120 秒短かった．そのため，本実験の結果に関しては，閲覧時アノテーションのリスト表示の効果が顕著に表れた結果と考えられる．

また,「文書閲覧時（マーキング箇所)」と「文書閲覧時（非マーキング箇所)」には，ほとんど平均検索時間に差は見られなかった．そこで「文書閲覧時（非マーキング箇所)」で引用した箇所の周辺にユーザによるマーキングが存在するか調べた.「文書閲覧時（非マーキング箇所)」で引用したユーザは計 4 名であり，すべてのユーザが引用箇所の近辺にマーキングを付与しており，マーキング範囲を若干修正して引用していたことが分かった．また，アンケートにおいて,「文書閲覧時（非マーキング箇所)」に引用箇所を決定したユーザ 4 名中 3 名がマーキング箇所を引用しなかった理由を,「自分の文章の文脈に合わせるために範囲を修正したかったため」と回答していた．もう 1 名に関しても，マーキングした箇所を含む範囲を引用していた．つまり，本実験においては「文書閲覧時（マーキング箇所)」と「文書閲覧時（非マーキング箇所)」のどちらの場合においても，文書にアクセスしマーキング近辺の文章を閲覧してから引用するという流れは変わらないため，検索時間に違いが出なかったものと思われる．

一方で，今回の実験では,「アノテーションリスト閲覧時」に引用箇所を決定した割合が 8 割近くに上っており，ほとんどのユーザがマーキング箇所の文章を閲覧するだけで，引用箇所を決定できたことを示している．また，マーキングした箇所を引用した割合[2]は約 9 割に上り，多くのユーザが文書を閲覧する際に引用すべき箇所を決定できたと言える．しかし，長期的かつ創造的な活動である研究活動における論文執筆のような文書作成の場合，この割合が下がることは容易に予想される．そのため，今後は長期的な実験を行い，本手法の有効性を検証する必要があるだろう．

次に，A グループにおける閲覧時アノテーションを付与した数と検索時間の関係を見ると，閲覧時アノテーションを付与した数が多いと検索時間が長くなる傾向が見られた．アノテーション数と検索時間をユーザ別にプロットしたグラフを図 5.4 に示す．これは，閲覧時アノテーションが多くなるにつれて，閲覧時アノテーションリストの中から必要なアノテーションを検索する必要が出てくるためである．本実験では，閲覧時アノテーションをフラットにリスト形式で並べて表示したため,「閲覧時アノテーションを付与している箇所が該当文書中のどの辺りであるか分かりづらく，一度文書にアクセスしないと引用箇所を決定できない」という意見があった．そのため，増田ら [24] が提案しているコンテンツ検索の仕組みのように，まず文書全体の俯瞰から閲覧時アノテーションの付与されている箇所

[2]「アノテーションリスト閲覧時」と「文書閲覧時（マーキング箇所)」の割合の和

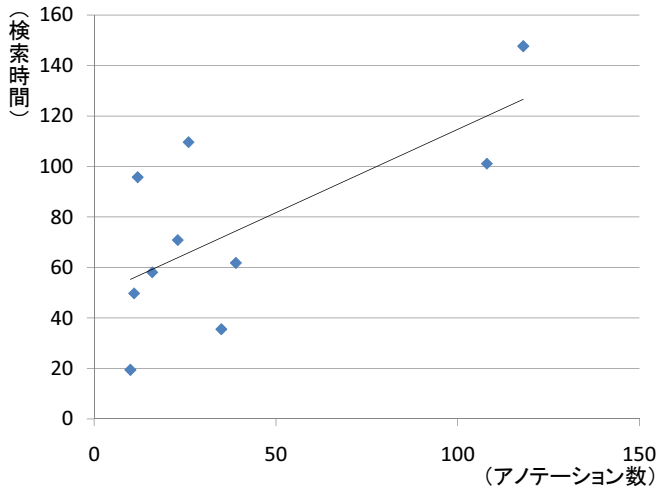
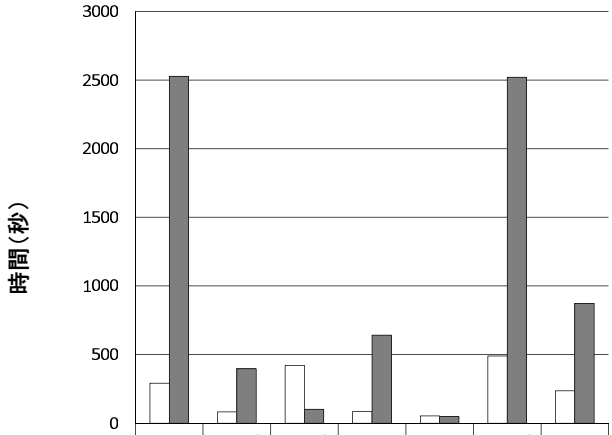


図 5.4: アノテーション数と検索時間

を検索していく仕組みが必要であると考えられる．

また，閲覧時アノテーションを多く付与していたユーザのアンケートに,「閲覧時アノテーションをユーザ自ら整理し，効率よく検索できる仕組みが欲しい」という意見があった．例えば，閲覧時アノテーションに対して検索用のタグを付与して整理する仕組みなどを提供する必要があるだろう．一方で，本システムはキーワードと属性による閲覧時アノテーションの検索機能を有していたが，検索が行われた回数は計 58 回と少数回であった．なお，属性を指定した検索に関しては一度も行われなかったが，現段階では閲覧時アノテーションの数が少なくキーワードによる検索で十分であったからであると考えられる．

次に，閲覧時アノテーションを付与するのに費やした時間と閲覧時アノテーションを付与することで短縮できた時間の関係を考察する．ここで，ユーザが閲覧時アノテーションを付与するのに費やした時間をアノテーション時間と呼ぶ．また，本実験では 6 名のユーザに A グループと B グループの両グループにおいて文書を引用してもらった．その 6 名のユーザにおいて，B グループ時の引用コストから A グループ時の引用コストを引いた値に，A グループ時に引用文書検索を行った回数を掛けた値を閲覧時アノテーションを付与することで短縮できた時間と考え，ここでは短縮時間と呼ぶ．短縮時間は閲覧時アノテーションを付与することで短縮できたと考えられる時間であるため，アノテーション時間と短縮時間は費用対効果の関係にある．ユーザ別のアノテーション時間と短縮時間の関係を図 5.5 に示す．図 5.5 では，アノテーション時間を左側（塗りつぶしなし）の棒グラフ，短縮時間を右側（塗りつぶしあり）の棒グラフで表し，右端にアノテーション時間と短縮時間の平均値を示す．ユーザ別に引用コストを見たところ，A グループ時に B グループ時以上の引用コストを費やしたユーザはいなかったため，短縮時間が



	user1	user2	user3	user4	user5	user6	avg.
□アノテーション時間	291.6	82.9	420.2	83.9	53.9	489.7	237
■短縮時間	2527.7	396	102.3	642.1	49	2520	873

図 5.5: アノテーション時間と短縮時間

マイナスとなるユーザはいなかった．また，最も短縮時間の大きかったユーザは2527.7 秒短縮しており，逆に最も小さかったのは 49.0 秒であった．平均アノテーション時間は 237.0 秒，平均短縮時間は 873.0 秒であるため，アノテーション時間の 3 倍以上短縮しており，本手法が有効であることを示している．

また，個別に見ていくと，6 名中 4 名はアノテーション時間に比べて短縮時間が大きかったが，2 名はアノテーション時間に対して短縮時間が小さかった．特に，ユーザ 3 はアノテーション時間が 420.2 秒に対して短縮時間は 102.3 秒であり，特にコストパフォーマンスが悪かった．ユーザ 3 のアノテーション時間の内訳を見ると，コメントに費やした時間が 239.8 秒と閲覧時アノテーション全体の時間の半分以上を占めていた．そのため，ユーザ 3 のコストパフォーマンスが悪かった要因としては，コメントの付与に大きな時間を費やしたことが挙げられる．しかし，コメントの付与されていた部分や周辺は引用しておらず，従ってユーザ 3 の検索時間の短縮には貢献していなかったと考えられる．表 5.2 からも分かるように，マーキングは 1 回平均 5.3 秒かかることに比べ，コメントは平均 65.1 秒と多くの時間を要するにも関わらず，今回の実験では検索に利用されることも少なく，検索におけるコメントの有用性が低かった．しかし，マーキングに比べてコメントは文章に対する主観的な情報を自然言語により豊富に記述することが可能であるため，閲覧してから長期間経過した後でも，コメントを閲覧することでアノテーションした文脈を想起することが可能であろう．そのため，一回の文書作成の利用に限定することなく，長期にわたり複数回利用されていくことが予想される．つまり，本実験だけではコメントの有効性を検証することはできないと考えられる．

5.2 類似文書検索に関する実験

4章で提案した，引用支援システムを運用することで得られる引用アノテーションを用いた類似文書検索手法の有効性を検証するために実験を行った．以下，まず実験方法について述べ，さらに実験結果と考察について述べる．

5.2.1 実験方法

本実験では，4章で提案した引用アノテーションを利用した類似文書検索手法の有効性を示すために，文書間類似度を算出する際に考慮する以下の指標の有効性に関して検証した．

- 引用意図の関係
  引用に付与されている引用意図が同じか否かに応じて共引用関係にある文書間の類似度を重み付けする．引用意図が同じ場合の共引用は，引用意図が異なる場合の共引用に比べ類似度を高く重み付けする．

- 引用箇所間の距離
  引用先文書における引用箇所間の距離の遠近に応じて共引用関係にある文書間の類似度を重み付けする．引用箇所間の距離が近い順に類似度が高くなるように重み付けする．

本実験では，上記した指標によって共引用を分類し，共引用によって関連付けられる文書間の類似性を，従来より提案されている類似度指標に基づき検証する．実データに基づき各重み付け手法の有効性を示すことが本実験の目的である．

検証するデータには，前節で述べた引用支援に関する実験によって得られた引用情報を利用する．引用支援に関する実験によって行われた引用回数は76回であり，引用された文書数は47本であった[3]．また，共引用の関係にある文書の組は52組であった．

検証に利用する文書間の類似度指標には，従来より一定の評価がなされているベクトル空間型モデルに基づく文書間類似度を利用する．以下に，ベクトル空間型モデルによる類似度指標の算出方法を示す [23]．

まず，各文書に含まれる語を形態素解析エンジンを利用して抽出し，各語を $tfidf(w_n, d_n)$ により重み付け，$tfidf(w_n, d_n)$ を要素に持つ文書ベクトルを生成する．重み付けに利用する $tfidf(w_n, d_n)$ の算出方法を5.1式に示す．また，抽出する語の品詞には名詞と未知語を利用しており[4]，不要語処理を行っている．

[3]引用支援に関する実験では，同一の文書において複数回同一の文書の異なる箇所を引用する行為が見られた．そのため，行われた引用数と引用された文書数は異なる．また，共引用を分類する際には，適切である引用のデータを利用した．例えば，引用箇所の距離であれば近い方を利用した．

[4]連続する名詞と未知語は連結して一つの語として扱っている．

$$tfidf(w_n, d_n) = \frac{tf(w_n, d_n)}{NoT(d_n)}(log\frac{NoD}{df(w_n)} + 1) \tag{5.1}$$

5.1 式の $tf(w_n, d_n)$ は文書 d_n における語 w_n の出現頻度，$NoT(d_n)$ は文書 d_n に出現する語の延べ出現回数であり，NoD は総文書数，$df(w_n)$ は語 w_n の出現する文書数である．

次に，文書間の類似度 $Sim(D_1, D_2)$ を，算出された文書ベクトル $D_1 = (x_i, x_2, \cdots, x_n)$ と文書ベクトル $D_2 = (y_i, y_2, \cdots, y_n)$ の内積として以下の 5.2 式により算出した．

$$Sim(D_1, D_2) = \frac{D_1 \cdot D_2}{||D_1||||D_2||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5.2}$$

本実験では，この $Sim(D_1, D_2)$ を文書間の類似度の指標値として利用する．以後，本節において単に類似度と述べられている場合，この値を指し示すものとする．

5.2.2 実験結果と考察

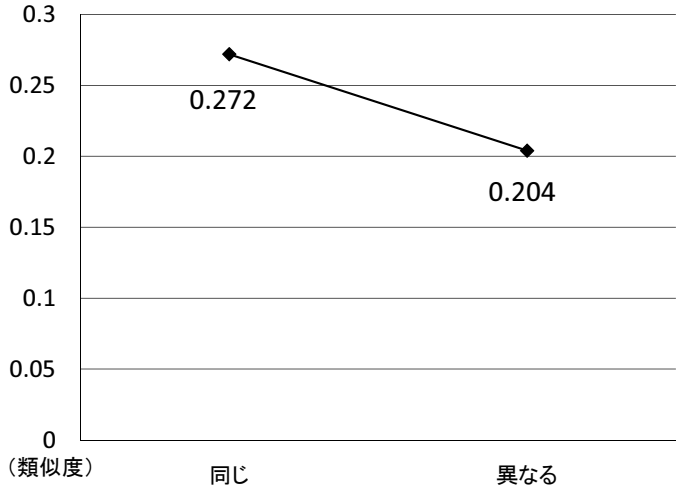


図 5.6: 引用意図の関係別の類似度比較

まず，引用意図が同じか否かで共引用を分類し，分類された共引用により関連付けられる文書同士の類似度 $Sim(D_1, D_2)$ の平均値を算出した．その結果を図 5.6 に示す．引用意図が「同じ」に分類された共引用による文書対は 35 組であり，「異なる」に分類された共引用による文書対は 17 組であった．図 5.6 からの分かるように，引用意図が「同じ」共引用のほうが引用意図が「異なる」共引用に比べて平均類似度が高かった．つまり，引用意図が「同じ」共引用によって関連付けられ

る文書同士の方が類似性が高く，提案している引用意図別に共引用による類似度を重み付けることの有効性を示すことができた．また，引用意図が「同じ」場合の共引用を，さらに引用意図の種類別（「論拠の明示」や「問題点の指摘」など）に分類して平均類似度を見てみたが，現段階のデータではサンプル数が少なく意味のあるデータは得られなかった．

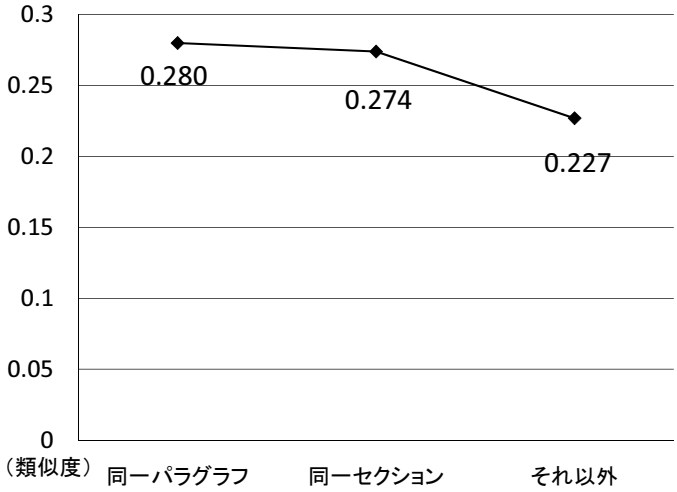


図 5.7: 引用箇所間の距離別の類似度比較

次に，引用箇所間の距離に応じて共引用を分類し，分類された共引用により関連付けられる文書同士の類似度 $Sim(D_1, D_2)$ の平均値を算出した．引用箇所間の距離は,「同一パラグラフ」,「同一セクション」,「それ以外」の 3 種類に分類した．つまり，共引用の引用箇所間の距離が最も近いのは「同一パラグラフ」であり，以下順に「同一セクション」,「それ以外」となる.「同一パラグラフ」に分類された共引用による文書対は 10 組,「同一セクション」に分類された共引用による文書対は 14 組,「それ以外」に分類された共引用による文書対は 28 組であった．

引用箇所間の距離別に平均類似度を算出した結果を図 5.7 に示す．図 5.7 から分かるように,「同一パラグラフ」が最も平均類似度が高く，次に「同一セクション」であり,「それ以外」が最も低かった．この結果は，引用箇所間の距離が近い順に類似性が高いということを示しており，提案している重み付け手法の有効性を示すことができた．しかし,「同一パラグラフ」と「同一セクション」にほとんど差が見られなかった．そこで「同一パラグラフ」に分類された共引用の内，類似度の低かったものを考察したところ，論文の「はじめに」のように大きく論理を展開しているパラグラフにおける共引用が確認された．また，大きく論理を展開しているパラグラフにおける共引用は引用意図が異なることが多かった．ある文書では「先行研究の例示」を示す引用の後，その先行研究の問題点を解決するための方法の「論拠を明示」を示す引用と引用意図の種類が異なっていた．

そこで，引用箇所間の距離別に分類した共引用を，さらに引用意図が「同じ」か「異なる」かによって分類を行い，平均類似度を算出した結果を図 5.8 に示す．図 5.8 は，図 5.7 と同様に横軸に引用箇所間の距離の違いをとっている．

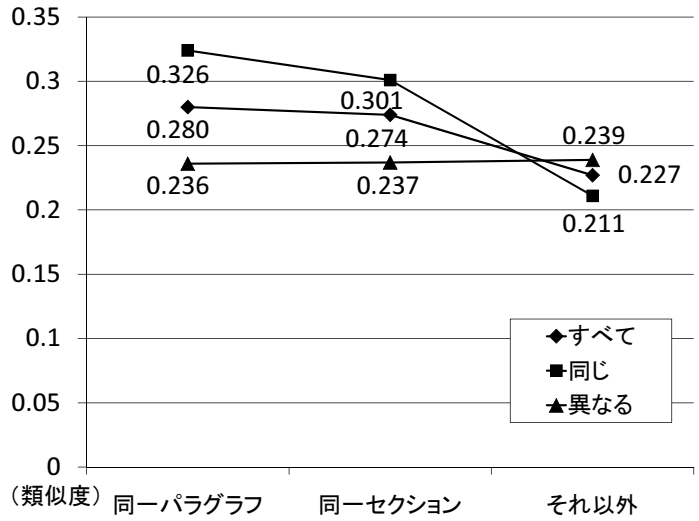


図 5.8: 引用意図の関係と引用箇所間の距離別の類似度の比較

最も平均類似度が高かったのは「同一パラグラフ」で引用意図が「同じ」共引用であり，両指標を組み合わせることでさらに高精度に共引用を分類できることが分かった．また,「同一パラグラフ」と「同一セクション」では，引用意図が「異なる」場合の共引用の平均類似度に比べて「同じ」場合の共引用の平均類似度が高いが,「それ以外」ではほとんど違いがなかった．つまり，引用意図による分類は引用箇所間の距離が近い場合に効果があるが，距離が遠い場合には効果がないと言える．さらに，引用意図が「異なる」場合には，引用箇所間の距離の関係による類似度の違いが見られなかった．今後は，この実験結果を踏まえて各共引用の重みの値を改良していく必要があるだろう．

5.3 まとめ

本章では，まず，3 章で提案した引用支援システムにおける閲覧時アノテーションを用いた Web ドキュメントの引用支援手法の有効性に関して行った実験について述べた．アノテーションするグループとアノテーションしないグループに分けて文書を引用してもらったところ，アノテーションするグループの方が引用文書検索に費やす時間が短いことが分かった．また，閲覧時アノテーションを付与するのに費やした時間と短縮できた時間の関係を費用対効果の関係とし，その値を

算出したところ，コストに対して約3倍のパフォーマンスが得られることが分かった．その結果，本研究の引用支援手法の有効性を示すことができた．しかし，引用箇所の決定フェーズの割合の変化やコメントの効果，属性情報を用いた検索の仕組みなどを検証するために，今後は長期的な実験を行う必要があることが分かった．また，検索インタフェースを改良していく必要があることが分かった．

次に，引用支援手法に関する実験において得られた引用情報を用いて4章で提案した共引用に基づく類似論文検索手法の有効性に関して行った実験について述べた．具体的には，引用意図の関係と引用箇所間の距離に基づく共引用の有効性を検証するために，各指標別に共引用を分類し，分類された共引用によって関連付けられる文書同士の平均類似度を比較した．その結果，引用意図の関係別の比較においては，引用意図が「同じ」共引用が類似している傾向が得られた．また，引用箇所間の距離別の比較の場合は，最も近い距離である「同一パラグラフ」における共引用が最も類似している傾向が得られた．さらに，両指標を組み合わせた場合,「同一パラグラフ」の「同じ」意図の共引用文書が最も類似している傾向が得られた．その結果，本類似文書検索手法の有効性を示すことができたと考えられる．

第6章　関連研究

本章では，本研究に関連する研究をいくつか紹介する．関連する分野として，アノテーションシステムに関する研究と類似文書検索に関する研究について述べる．また，関連研究と比較することで本研究の位置づけを明確にする．

6.1　電子文書に対するアノテーションシステム

これまでに，Web ドキュメントのような文書に対するアノテーションからメタデータを作成・保存し，そのメタデータを様々な応用に適用するシステムは多く開発されている．その中から，SemCode[25]，SmartCourier[26]，イロノミー [27]，XLibris[3]，Annotea[28]，ComMentor[29] の 6 つのシステムに紹介し，本システムと比較する．

6.1.1　SemCode

SemCode[25] は，テキストやビデオを含む Web コンテンツ一般に対するアノテーションと、その応用としてのトランスコーディング（コンテンツの要約や翻訳などの変換）を実現した先駆的なシステムである．さらに，SemCode ではアノテーションの二次利用を念頭において，コンテンツに対する直接のアノテーション（第一層アノテーション）からオントロジーに相当する部分を抽出する仕組みを提案している．一般にアノテーションはコンテンツと直接関連付けられてしまっており，他のコンテンツに適用することは困難であるが，SemCode ではアノテーションからオントロジーを抽出して適用範囲を広げることを提案している．オントロジーは，RDF（Resource Description Framework）によって第一層アノテーションと関連付けて記述されている．

SemCode におけるアノテーションは，専門的な知識を持つアノテータと呼ばれる専門家がオントロジーエディタと呼ばれる専用ツールを用いて作成するメタデータであり，本研究におけるアノテーションに比べて質の高いアノテーションであると言える．また，SemCode では構築されたオントロジーを用いたトランスコーディングの例として，専門用語の言い換えを行う仕組みを実現している．本研究

において，ユーザの閲覧時アノテーションを利用して専門用語辞書を構築する仕組みを提案しているが，これは SemCode の考えに基づいている．

6.1.2 SmartCourier

SmartCourier[26] は，ユーザの研究学習活動を対象に，興味関心や研究目的といったコンテキストをアノテーション行為から抽出することで，適応的な関連文献推薦や，近い興味を持つ同僚研究者のマッチメイキング，アノテーションの共有サービスを提供している．SmartCourier では，本研究のようにマウス操作によるアノテーションではなく，手書きペン入力による自由なアノテーションを対象としている．

SmartCourier では，アノテーションされたテキストからキーワード（名詞）の抽出を行いそのキーワードベクトルに基づくユーザモデルを構築している．そのため，文章の論理構造を考慮してアノテーションされたテキストの位置を推定する必要がある．しかし，本研究と異なり PDF フォーマットによる論文を対象としているため，ヒューリスティックスにより一部の論文誌の書式を想定するに留まっている．また，囲み文字やアンダーラインといったマーキングの種類による意味の違いを 5 段階の重要度と自由記述による検索フレーズで表している．本研究では，マーキングの意味をあらかじめ用意した属性と対応付けてもらうことで，柔軟な検索を可能にしているが，SmartCourier ではユーザによる自由記述によって実現している．そのため，アノテーションの共有を円滑に行うためには，一般的なソーシャルブックマークサービス[1]のタグと同様に検索フレーズの語の表記ゆれ等の問題に対処する必要がある．また，マーキングの共有を前提として，あらかじめマーキングの種類に属性を対応付けることが可能である点は本研究と類似しているが，本研究ではマーキングの属性だけでなく，辞書引きなどのオペレーションを対応付けることが可能である点が異なる．

6.1.3 イロノミー

イロノミー [27] は，多くの人手で膨大な情報の分類をおこなうという意味を持つ Folksonomy の考えに基づき，Web ドキュメントの分類を目的としたアノテーションシステムである．自由入力によるタグではなく，下線が引かれた箇所の文字列を特徴的なキーワード，つまりタグとして捉えている．ユーザは斎藤の提案している三色ボールペン読書法 [30] と呼ばれる方法で，下線を引きながら文章を

[1] ソーシャルブックマークサービス（SBS）とは，よく使うサイトのアドレスを登録しておく「ブックマーク」をネットワーク上に保存し，他のユーザと共有するサービス．分類に従来の Web ディレクトリのような固定的・階層的な分類法ではなく，ユーザが自由に設定できる「タグ」と呼ばれる単語やフレーズを利用しているものが一般的である．

読み進めると自動的に保存されるようになっている．また，保存されたアノテーションはキーワードにより検索可能であり，アノテーションの協調フィルタリングに基づいて未読文書の推薦も行っている．

下線とハイライトマーカという違いはあるものの，同じマーキング情報を検索に利用するという観点から本システムと類似性は高い．三色ボールペン読書法に基づいているため，イロノミーで保存されるアノテーションの属性は赤色は「客観的にとても重要」，青色は「客観的にまあ重要」，緑色は「主観的に重要」である．しかし，検索においては，それら曖昧な属性を指定しても，目的に合致した結果を得ることが困難であるという問題点がある．また，本システムでは，ユーザが利用したい属性を自由に色に対応付けることが可能であるが，イロノミーにおける属性と色の対応関係は固定である．そのため，それら属性の意味を理解できないままアノテーションを行っていたユーザも存在していたという調査結果も出ている．これは，ユーザによって下線の色に対応する情報の種類が異なり，用意された下線引きの属性がユーザの直感に合わなかったことを示している．

6.1.4 XLibris

XLibris[3] は，ユーザの「Active Reading」を支援するシステムである。「Active Reading」とは，単純に文章を読むだけではなく，文章に対して下線を引いたり，コメントを付与したりしながら，考え，学ぶような読み方を言う．XLibris は，ペンタブレットを用いて，電子文書に手書きペン入力によって書き込みをしながら文書を読むことができるシステムであり，紙の文書に対して行うような自由な書き込みを可能にしている．また，Reader's Notebook と呼ばれるビューで，付与されたアノテーションをリスト形式で表示することができる．Reader's Notebook では，アノテーションのインクの色を指定して検索・表示することも可能であり，本システムの機能と類似している．また，一つの文書中で複数ユーザのアノテーションを表示する際に，ユーザ別に色を対応付けて表示することが可能である．

XLibris は，主に紙のメリットである簡便性や携帯性に焦点を当てているため，計算機的な問題点が存在する．例えば，電子文書に対してペンタブレットを用いて自由なアノテーションを可能にしているため，計算機側で文書のどの個所に対するアノテーションであるかを特定する必要があるだろう．しかし，人間のアノテーション行為から完全に特定することは困難であろう．つまり，人間には全く負担をかけず，計算機が支援可能な範囲で人間の行動を支援するというアプローチをとっている．それに対して本システムは，属性を付与してもらうなど人間に対して負担かけることで計算機に理解しやすい形で情報を取得する代わりに，その分計算機が人間の行動をより高度に支援するというアプローチをとっている．

6.1.5　Annotea

Annotea[28] は Web ブラウザを通して特定の箇所にコメントを付与することができ，コメントは RDF フォーマットのメタデータとして保存される．これは Semantic Web[31] の実現に向けて，メタデータの作成を支援する目的があるためで，将来的にはこのメタデータを使って新たなサービスを提供するものと思われる．

アノテーション行為により得られたメタデータを再利用して新しい応用を提供するという考え方は本研究と類似している．しかし，Annotea は一般的な応用を対象とした汎用的なメタデータを作成することに着眼点を置いているが，本システムでは実現する特定の応用を絞り込み，その応用に適した形でメタデータを作成している点が異なる．一方で，アノテーション箇所のポインタ情報を XPointer により管理していることも本システムと類似している．しかし，本システムのように前処理としてテキスト部分を形態素レベルまで分割し，タグの付与をしていないため，指し示すことのできる範囲はあくまでも HTML で定義される既存のタグのレベルであり，本システムのように柔軟なテキスト範囲に対するアノテーションは記述できない．

6.1.6　ComMentor

ComMentor[29] は，Web 上の文書に対するアノテーションを共有することでグループのコミュニケーションを促進させることを目的とした協調作業支援を行うアノテーションシステムのためのアーキテクチャである．ComMentor におけるコメントはそれ自身がドキュメントであるとし，ユーザはコメントに対してコメントを追加することができる．これにより，オリジナルの Web ページに対してスレッド形式でディスカッションすることが可能である．また，ComMentor は Web ブラウザを通じたアノテーションを想定しており，Annotea と同様に応用全般を対象とした一般的なメタデータを作成することに着眼点を置いている．

ComMentor では，プライベート，グループ，パブリックの 3 段階でアノテーションに対するアクセス権限を管理することが可能であり，本システムにおける「公開」・「非公開」の設計と類似している．また，アクセス権限の管理方法からも分かるようにグループに特化した仕組みになっている．共著で Web ドキュメントを作成することを念頭に置き，本システムにおいても今後グループという概念を取り入れることを検討していきたい．

6.2　引用情報に基づく類似文書検索

本節では，引用情報を類似文書検索に応用した研究について述べる．

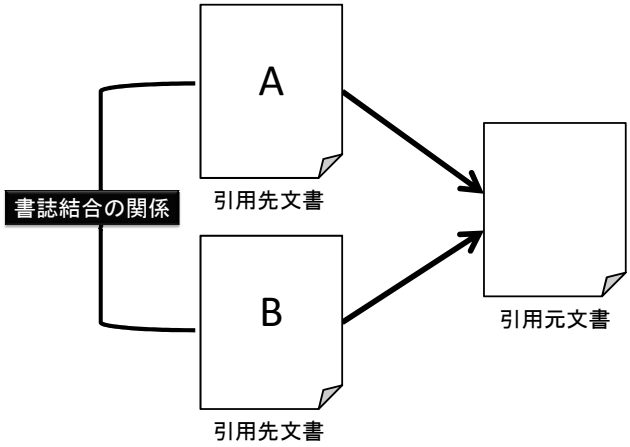


図 6.1: 書誌結合の関係

難波ら [7] は，手掛かり語[2]を用いて抽出される参照の理由[3]を考慮して書誌結合 [8]（bibliographic coupling）の類似度を算出している．書誌結合の関係とは，共引用より古くより提案されている文書間の類似度指標で，ある同一の文書を引用している文書同士の関係を表しており，図 6.1 のような関係にある文書 A と文書 B の関係を書誌結合の関係という．難波らは，参照の理由を「論説根拠型（type B）」「問題点指摘型（type C）」「その他型（type O）」の 3 種類に分類し，その参照の理由の一つの，typeC を考慮して書誌結合のような尺度を算出することで，文書間の類似関係をより高精度に捉えることが可能になると述べている．typeC は参照論文の問題提起や研究動機を示す参照と考えることができ，論文間で多くの typeC の参照が一致すれば，これらの論文の著者は共通の問題意識を持っていると考えられるため，難波らは typeC に着目している．本研究では，現時点ではデータ数が少なかったため引用意図が同じか否かの違いによって共引用を重み付けしたが，引用意図の種類の違いによる類似度の違いを検討する必要があるだろう．

また，江藤 [9] は，論文は体系づけられて構成されるため意味的に近い内容のものはまとまって述べられるという仮説を立て，引用箇所間の間隔が，共引用における文書間の類似性に影響を与えると述べている．そして，引用文書の文脈情報を利用することで共引用による文書間類似度は精度の高いものになることを実験・検証している．例えば，引用論文の冒頭の「はじめに」で引用された論文と末尾の「おわりに」で引用された論文の類似度と,「関連研究」について述べられている部分の同一パラグラフ内で引用された論文同士の類似度は異なるだろう．引用箇所間の関係により共引用を「列挙」「同一文」「同一段落」「非同一段落」の 4 種

[2]手掛かり語とは，文書の意味的な構造を示す特徴的な語.
[3]参照の理由とは，本研究で言う引用意図に対応している概念.

類に分類し，最も距離の近い「列挙」が類似度が高いことを実験により示している．本研究においては，引用箇所間の距離を「同一パラグラフ」「同一セクション」「それ以外」に分類しており,「列挙」と「同一文」という分類が存在しない.「列挙」に関しては，孫引きが多く行われる引用形式であるため本システムではできないようになっている．また,「同一文」に関しては，現時点のシステムの運用においては，たまたま確認できなかったため考慮していない.

難波らや江藤の研究は，参照の理由や文脈情報のような引用に関する意味的な情報を考慮して共引用や書誌結合のような文書間の関係を扱っている点で本研究と類似している．しかし，引用情報を抽出する方法が引用文書のテキスト処理による自動抽出であり，引用するシステムから提案している本研究とは異なる．例えば，難波らは参照の理由を手掛かり語を用いて文書テキストから自動抽出している．そのため，自動抽出の精度が類似文書検索の精度に影響を及ぼすという問題点が存在する．しかし，本システムは引用者に明示的に引用意図を付与してもらい，文書のアノテーション情報として文書テキストと分割して管理する仕組みであるため，自動抽出における問題点を考慮する必要がない．実際に，難波らは自動抽出の精度を高めることで，類似文書検索精度をさらに高めることができると述べている．また，本研究では自動抽出の精度を考慮する必要がない反面，提案する仕組みをどのようにして多くのユーザに利用してもらうかを考える必要がある．この点に関しては今後の課題である.

第7章　おわりに

本章では，本論文をまとめ，今後の課題について述べる．

7.1　まとめ

本論文では，まず，既存の引用の記述形式であるサイテーションとクオテーションについて考察し，Web ドキュメントの適切な引用の仕組みを提案した．また，提案する引用の仕組みを実現するために文書の内部要素を扱う必要があることについて言及し，XML に基づく文書フォーマットを提案した．さらに，形態素解析を行い形態素単位に動的に言語構造に関するタグを付与することで，文書 XML で定義される内部要素より詳細な要素に対しても，既存のポインタ言語である XPointer を用いてアクセス可能であることを述べた．

次に，提案する引用の仕組みを実現するために構築した引用支援システムについて述べた．本システムは Web アプリケーションとして実装されており，現在の引用を行う際の問題点に対する解決策として，閲覧時アノテーションを利用した引用支援機能を有することを述べた．また，引用支援手法の有効性に関して実験を行った．具体的には，延べ人数 16 人の被験者に対して，文書閲覧時にアノテーションするグループとアノテーションしないグループに分かれて，文書作成時に引用文書を検索してもらった．その結果，特に引用文書検索に費やした平均時間に関して，本手法の優位性を示すことができた．また，閲覧時アノテーションに費やした時間と文書作成時に引用文書検索において短縮した時間を比較したところ，閲覧時アノテーションに費やした時間の 3 倍以上の時間を文書作成時に短縮できることで，本手法の有効性を示した．

さらに，本システムを運用することで得られる引用情報である引用アノテーションについて考察し，引用アノテーションを用いた応用例を示した．具体的には，引用先文書の引用箇所間の距離と引用意図の関係を利用した共引用の詳細な重み付けに基づく類似文書検索手法を提案した．また，類似文書検索に関する実験において，提案する重み付け手法の有効性を示し，類似文書検索手法の可能性を示した．具体的には，引用意図の関係と引用箇所間の距離に基づきに共引用を分類し，各共引用によって関連付けられる文書同士の類似度を比較した．その結果，引用意図の関係と引用箇所間の距離のどちら指標に基づき分類した場合においても，良

好な結果が得られた．また，両指標を組み合わせることで，共引用の関係にある文書同士をさらに高精度に分類できることを示した．この結果により，従来の引用情報による応用に比べて，引用アノテーションは意味的な応用を実現できることを示した．

7.2 今後の課題

今後の課題としては，以下のことが挙げられる．

被引用箇所情報を利用した共引用の重み付け

本論文では，引用アノテーションを用いた応用例として，共引用に基づく類似文書検索手法を提案した．その際に，文書は体系づけられて構成されるため意味的に近い内容のものはまとまって述べられることを考慮して，引用先文書の文脈情報を利用して共引用における文書間類似度を重み付けした．しかし，引用先文書において意味的に近いとしても，その引用箇所間の関係がそのまま引用元文書同士の関係になるとは限らない．例えば，被引用箇所が引用元文書において主となる主張ではなく，本論から枝分かれした瑣末な主張である場合，引用箇所の内容がそのまま引用元文書の内容を表してはいないだろう．そのため，類似度を算出する際には，引用元文書の文脈情報を利用して被引用箇所がその文書において主となる主張なのかどうかを判断して考慮する必要がある．

そのためには，引用元文書の特徴的な箇所を抽出し，被引用箇所との関係を明確にする必要があるだろう．特徴的な箇所の抽出方法には，ユーザの多くにマーキングされている箇所は該当文書において重要な箇所として抽出する方法が考えられる．また，被引用箇所に含まれる文字列に，引用元文書全体を特徴付ける語が多く含まれる場合には，その箇所はその文書において特徴的な箇所として抽出するような方法も考えられる．特徴的な箇所と被引用箇所の情報を利用すると，共引用の類似度の信頼性を考慮することができると思われる．例えば，引用元文書の特徴的な箇所を被引用箇所とする引用の場合には，引用元文書の主となる主張を引用していると考えられ，その引用を用いた共引用により算出された類似度の信頼性は高いと判断できる．逆に，もし，マーキングのほとんどない箇所であった場合，その箇所は該当文書において重要な箇所ではなく特徴的な箇所でないと推定できる．そのため，算出された類似度の信頼性は低くなると考えられる．つまり，被引用箇所情報は共引用の重み付けに利用できる可能性があり，強いては類似文書検索に応用できると考えている．

人間の意味解釈の誤りによる引用の防止

引用の中には論理的に正しくない引用や，引用元文書を誤読した引用が存在する．共引用のような引用分析を行う場合には，そのような誤引用は排除する必要がある．

本論文では，引用箇所と被引用箇所をハイパーリンクで繋ぐことで，直接的に引用・被引用関係を参照可能な仕組みを提案した．本仕組みは，引用の構造を明確に示す仕組みではあるが，本論文では引用の構造から誤引用を判断する方法については深く考察していない．考えられる対処法としては，引用元文書の著者が引用に対して人為的に評価する方法が挙げられる．本システムは，引用元文書に被引用箇所を表示している．その箇所に対して引用元文書の著者が自己の主張を正しく解釈し，引用しているかを判断し，もし，誤っている場合には引用に対するアノテーションとして誤引用であるという情報を付与することを可能にする．そうすることで，引用分析を行う際には排除することが可能になる．この手法の場合，引用元文書の著者が誤ったアノテーションを行う可能性があるだろう．そこで，アノテーション情報を公開して表示することで，著者に対して慎重に精査しアノテーションすることを促す．また，蓄積されたアノテーション情報を解析することで，機械的に誤引用であるかどうか判断する方法を検討することも考えられる．

類似文書検索手法の評価

本論文では，引用アノテーションを用いた応用として，共引用に基づく類似文書検索手法を提案した．本論文では，提案する類似文書検索手法に利用する共引用における文書間の各類似指標の妥当性を検証したが，あくまでも提案手法の予備実験に過ぎず，提案手法自体を評価したわけではない．そのため，実際に提案手法自体を評価する方法を考える必要がある．

評価するためには，まず，提案手法による類似文書検索を実現するために必要なデータを収集する必要があるだろう．本手法は文書間の類似度を統計的に算出する手法であるため，多くの引用に関するデータを必要とする．そのためには，システムを一般に公開し，より多くのユーザに利用してもらう必要がある．次に，得られたデータを用いて被験者実験を行い，既存の類似文書検索手法と比較評価する．そして，各手法の検索精度を算出し，定量的に優位性を示す方法が考えられる．

閲覧時アノテーションの共有

本論文では，閲覧時アノテーションを文書の引用支援という個人に閉じた形でに利用した．しかし，閲覧時アノテーションを共有することで，実現可能になる

応用も存在するだろう．

その一つに，Folksonomy のような考え方に利用できる可能性がある．Folksonomy は Web ドキュメントに対して人々がタグを付与することで，Web ドキュメントの分類を共同で行うという考え方である．Web ドキュメントに付けられたタグや他人が付けたタグをたどることで，今まで知らなかった Web ドキュメントを発見できることが新しく，多くの人が Folksonomy のサービスを利用している．同様の考え方に基づき，付与されたマーキングを利用して Web ドキュメントを分類していく仕組みが考えられる．

また，共有された閲覧時アノテーションに対して統計的な処理を行い，意味のある情報をマイニングすることも考えられる．例えば，多くのユーザがマーキングを付与している箇所は，その文書にとって重要な箇所であり，その箇所を抽出することで要約を生成できる可能性があるだろう．

一般の Web ドキュメントへの適用

本論文では，Web ドキュメントの一例として，文書 XML に限定して引用の仕組みをシステムを実装した．しかし，提案した引用の仕組みは一般の Web ドキュメントに対して実現不可能な仕組みではない．そのため，今後の課題として，一般の Web ドキュメントを対象としたシステムに拡張していくことが挙げられる．その際には，Web ドキュメントの信頼性や保存性，公開性のような問題点に関して考慮する必要がある．また，Web ドキュメントの文章部分のみならず，石戸谷ら [32] のような画像やビデオなどのマルチメディアデータの部分に対する引用の仕組みについて考えていく必要があるだろう．

謝辞

本論文を執筆するに当たり，数多くの方々の御支援，御協力を賜りました．この場で皆様に感謝の言葉を申し上げたいと思います．

名古屋大学情報メディア教育センターの長尾確教授には，研究者としての心構えから，研究活動，及びプレゼンテーションに至るまで幅広い御指導を頂きました．また，学会発表を始め，数多くの貴重な体験をさせて頂きました．心より御礼申し上げます．

同大学エコトピア科学研究所の大平茂輝助教には，研究活動の中で数多くの御指導を頂きました．特に，プロジェクトゼミにおいて，研究に関する多くの有益な御意見，及び論文執筆に関する御指導を頂きました．

同大学情報科学研究科博士課程の山本大介さんには，プログラミングに関するアドバイスを始め，研究活動全般に関する助言を頂きました．同研究科博士課程の土田貴裕さんには，ディスカッションマイニングプロジェクトのリーダーとして，論文や発表資料の添削を始め，様々な形でお世話になりました．

同研究科修士課程の伊藤周くん，成田一生くん，石戸谷顕太朗さん，増田智樹くん，尾崎宏樹くん，学部生の安田知加さんには，ゼミにおける貴重な御意見を始め，研究室の活動の中で大変お世話になりました．

長尾研究室の秘書である鈴木美苗さんには，不自由なく研究活動が行えるように，研究室生活全般に関するサポートをして頂きました．

長尾研究室 OB の友部博教さん，梶克彦さんには，研究室在籍中に大変お世話になりました．

以上の方々を始め，本論文を執筆するにあたり，御支援，御協力頂いたすべての方々に，深く感謝の意を表します．

最後に，日々の生活を支えてくれた父，母，また，陰ながら支えてくれた兄，祖母に心より感謝いたします．

参考文献

[1] W3C, XML Path Language, http://www.w3.org/TR/xpath.html.

[2] W3C, XML Pointer Language, http://www.w3.org/TR/WD-xptr.

[3] B. N. Schilit, G. Golovchinsky, M. N. Price, Beyond Paper: supporting active reading with freeform digital ink annotations, In Proceedings of CHI '98, pp.249-256, 1998.

[4] H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for lnformation Science, Vol.24, pp.265-269, 1973.

[5] CiteSeer, http://citeseer.ist.psu.edu/.

[6] M. Weinstock, Citation indexes, Encyclopedia of Library and Infomation Science, Vol.5, pp.16-41, 1971.

[7] 難波 英嗣, 神門 典子, 奥村 学, 論文間の参照情報を考慮した関連論文の組織化, 情報処理学会論文誌, Vol.42, No.11, p.2640-2649, 2001.

[8] M. M. Kessler, Bibliographic coupling between scientific papers, Ammerican Documentation, Vol.14, pp.10-25, 1963.

[9] 江藤 正己，引用箇所の間隔に基づいた共引用の検討，電子情報通信学会第 18 回データ工学ワークショップ，L1-1，2007.

[10] テッド・ネルソン, リテラリーマシン―ハイパーテキスト原論, アスキー, 1994.

[11] E. Garfield, The history and meaning of the journal impact factor, Journal of the American Medical Association, Vol.295, No.1, pp.90-93, 2006.

[12] EndNote, http://www.endnote.com/.

[13] C. C. Marshall, Annotation: from paper books to the digital library, In Proceedings of Digital Libraries '97, pp.131-140, 1997.

[14] 伊藤 清美, 柳沢 昌義, 赤堀 侃司，Web 教材への書き込みを共有する学習環境 WebMemo システム，電子情報通信学会技術研究報告，Vol.100，No.467，pp.35-40，2003.

[15] 松岡 有希, 坂本 竜基, 中田 豊久, 伊藤 禎宣, 武田 英明, 論文概要に対する色付きアンダーライン付与システムの運用・分析, 電子情報通信学会第 17 回データ工学ワークショップ，1A-i8，2006.

[16] 藤田 節子，電子文献の参照をめぐる問題点，情報と科学の技術，Vol.51，No.4，pp.239-244，2001.

[17] Sen, http://ultimania.org/sen/.

[18] MeCab, http://mecab.sourceforge.net/.

[19] goo 辞書，http://dictionary.goo.ne.jp/index.html.

[20] W3C, XMLHttpRequest, http://www.w3.org/TR/XMLHttpRequest/.

[21] CiNii, http://ci.nii.ac.jp/.

[22] GoogleScholar, http://scholar.google.com/.

[23] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing Management, Vol.24, No.5, pp.513-523, 1988.

[24] 増田 智樹, 山本 大介, 大平 茂輝, 長尾 確, オンラインアノテーションを利用したビデオシーン検索, 人工知能学会第 21 回全国大会, 1G1-6, 2007.

[25] K. Nagao, Digital Content Annotation and Transcoding, Artech House Publishers, 2003.

[26] 伊藤 禎宣，角 康之，間瀬 健二，國藤 進，SmartCourier:アノテーションを介した適応的情報共有環境，人工知能学会論文誌，Vol.17，No.3，pp.301-312，2002.

[27] 坂本 竜基，中田 豊久，伊藤 禎宣，松岡 有希，小暮 潔，イロノミー：色付き傍線による Web 文章を対象としたフォークソノミー，人工知能学会第 20 回全国大会，3D1-4，2006.

[28] J. Kahan, M. -R. Koivunen, E. Prud'Hommeaux, R. R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In Proceedings of the WWW 10th International Conference, pp.623-632, 2001.

[29] M. Roscheisen, C. Mogensen, T. Winograd, Shared Web Annotations as a Platform for Third-Party Value-Added, Information Providers : Architecture, Protocols, and Usage Examples, Technical Report CSDTR/DLTR, 1994.

[30] 齋藤 孝，三色ボールペン情報活用術，角川書店，2003.

[31] W3C, The Semantic Web Community Portal, http://www.semanticweb.org/.

[32] 石戸谷 顕太朗, 増田 智樹, 山本 大介, 長尾 確, 引用の構造化によるマルチメディアコンテンツの意味的統合支援システム, 情報処理学会第 70 回全国大会, 2008.