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

林 亮介 † 土田 貴裕 † 大平 茂輝 †† 長尾 確 ‡

† 名古屋大学 情報科学研究科 †† 名古屋大学 エコトピア科学研究所 ‡ 名古屋大学 情報メディア教育センター

1 はじめに

引用は人間が意味を考慮して情報間を関連付ける作業であり，引用行為から得られる情報の有用性は高い．しかし，現在は引用されている文章の文脈情報や引用者の引用意図といった引用に関する意味的な情報は考慮されていない．これは引用情報が文書中のみに記述されていることに起因している．引用箇所に対して自然言語処理を行い，引用に関する意味的な情報を抽出するアプローチも考えられるが，引用箇所の記述形式も様々であり抽出は困難である．

本研究では，Web ドキュメントの新しい引用の仕組みを提案している [1]．提案する引用の仕組みは，引用・被引用文書の関係を内部要素レベル（セクションやパラグラフ）で扱うことが可能であり，引用者の意図を明確に記述できる仕組みである．また，引用情報を引用・被引用文書の両文書のアノテーション情報として定義することで，文書の内部要素間を引用意図に関する属性により双方向に関連付けることが可能である．

本稿では，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，閲覧時アノテーションと呼ばれる人間が文書閲覧中に付与するメタデータを利用した引用支援システムを構築し，実験を行った結果を報告する．さらに，得られる引用情報を利用した応用例として類似文書検索手法を提案する．

2 閲覧時アノテーションを利用した引用支援

2.1 閲覧時アノテーションと引用文書検索

我々人間は，文書を閲覧する際に文書の部分要素に対してマーキングやコメント付与といった様々なアノテーションを行う．本稿におけるアノテーションとは，原著作者を含むすべてのアノテーション行為を行う人間がデジタルコンテンツの階層化 [2] などを目的とし，人為的に作成する二次情報を指し，機械的に生成される類の二次情報とは区別して考える．本稿では，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．文書を閲覧する目的が明確な場合には，閲覧時アノテーションは人間の自然な行為である．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 [3]．

Quotation of Web Documents Using Reading Annotation and its Applications

† HAYASHI,Ryosuke(hayashi@nagao.nuie.nagoya-u.ac.jp)

† TSUCHIDA,Takahiro(tsuchida@nagao.nuie.nagoya-u.ac.jp)

†† OHIRA,Shigeki(ohira@nagoya-u.jp)

‡ NAGAO,Katashi(nagao@nuie.nagoya-u.ac.jp)

Graduate School of Information Science, Nagoya University (†)

EcoTopia Science Institute, Nagoya University (††)

Center for Information Media Studies, Nagoya University (‡)

Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

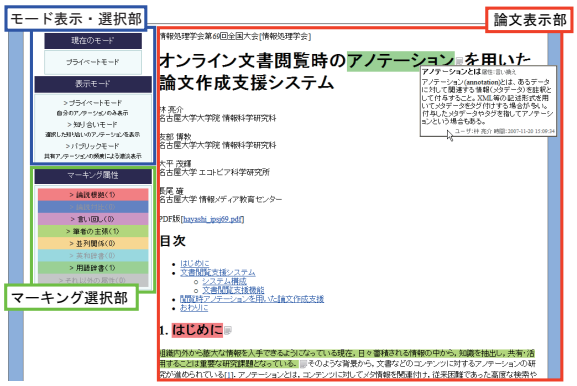

図 1: 文書閲覧インタフェース

そこで本稿では，ユーザの閲覧時アノテーションを記録し，アノテーションをトリガーとした引用文書検索手法を提案する．本手法では，従来のように引用する文書を発見した後，文書中の引用箇所を検索するという流れではなく，まず閲覧時アノテーションを検索し，引用箇所を直接的に探すことが可能である．また，過去に付与した閲覧時アノテーションを文章と共に参照することで，過去の文脈が想起され，文書の内容を容易に再理解することができる．

2.2 引用支援システム

本システムは，Web アプリケーションとして構築されており，ユーザは Web ブラウザ以外の特殊なクライアントソフトを必要としない．

まず，本システムの文書閲覧インタフェース (図 1) について述べる．本インタフェースでは，ユーザが Web ドキュメントに対して閲覧時アノテーションを付与することができる．可能な閲覧時アノテーションはマーキングとコメント付与である．まず，Web ドキュメントをユーザに提示する前処理として形態素解析を行い，テキスト部分を形態素単位に分割し，言語構造に関するタグを付与する．形態素単位にタグを付与することで，Web ドキュメントの詳細な内部要素に対するアクセス手段を確保し，アノテーション範囲を XPointer により指し示すことが可能になる．本インタフェースでは Ajax 技術を適用しており，ユーザは簡単なマウス操作によりアノテーションを付与することができる．また，閲覧時アノテーションには様々な属性情報を付与することが可能である．

次に，引用文書検索インタフェースについて述べる．本インタフェースでは文書を作成する際に閲覧時アノテーションをトリガーとした引用文書検索が可能である．図 2 のように閲覧時アノテーションはリスト形式で表示され，キーワードと閲覧時アノテーションにおいて付与した属性情報を用いて検索することが可能である．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書きたい場面がある．その場合，あらかじめ文書閲覧中に対比したい文書の該当箇所をマーキングし，自己の主張との差分をコメントとし

図 2: 閲覧時アノテーションのリスト表示

て記述し，問題点属性を付与しておく．その後，実際に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アノテーションを検索することが可能である．

2.3 実験と考察

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．10 名の被験者に 2 つのグループに分かれてもらい，本システムを用いて文書を作成してもらった．A グループには文書閲覧時にアノテーションしてもらい，B グループにはアノテーションしてもらわず，文書を作成時に適切な文書を引用してもらった．引用対象とした 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日本語論文†であり，その総数は 57 本であった．

本実験では，2 つのグループにおける引用コストを比較した．引用コストとして，引用 1 回あたりの引用文書検索に費やした時間を採用した．

表 1: 引用コストの比較

	平均引用検索時間 (s)
A(アノテーションする)	123.0
B(アノテーションしない)	191.5

実験結果を表 1 に示す．ユーザが引用箇所を検索した平均回数は A グループが 3.86 回であり，B グループは 5.33 回であった．閲覧時アノテーションを記録・検索可能にすることで，引用 1 回あたり 68.5 秒の検索時間を短縮でき，有効性を示すことができた．

3 引用アノテーションに基づく類似文書検索

本システムを運用することで，引用アノテーションと呼ばれる情報がシステムに蓄積される．通常，引用情報とは引用文書にテキスト形式で含まれるものであるが，本研究では引用情報を引用文書と被引用文書の両文書のアノテーション情報であると捉え，文書に埋め込まれない XML 形式で管理する．引用アノテーションは，主に引用文書の引用箇所，被引用文書の被引用箇所，引用意図に関する属性の 3 つの情報を保持する．従来の引用情報のように単純に文書間を関連付けるだけでなく，文書の部分要素間を引用意図で関連付けることから，文書間の関係をより意味的に表すことができる．

† http://www.nagao.nuie.nagoya-u.ac.jp/papers/papers.xml

従来から共引用 (Co-citation) と呼ばれる関係を利用した文書間の類似尺度が提案され，類似文書検索に利用されている [4]．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被引用文書同士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．

そこで本研究では，引用アノテーションを利用して共引用の関係をより詳細に扱うことを試みる．具体的には，引用文書の引用箇所間の関係と引用意図に関する属性情報の関係を利用した類似文書検索手法を提案する．以下に，類似文書検索に利用する文書 D_1 と D_2 の類似度 $Sim(D_1, D_2)$ の算出式を示す．

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

$$CocitedW_i = DistanceW_i * IntentW_i \quad (2)$$

式 (1) における N は文書 D_1 と D_2 の共引用回数であり，$CocitedW_i$ は共引用の関係に基づく値で式 (2) で示す．式 (2) における $DistantceW_i$ は，共引用の関係にある引用箇所を含む要素間の XPath のステップ数の逆数であり，引用箇所間の距離が近いほど大きくなる値である．XPath のステップ数とは，例えば XPath が”/papers[1]/docbody[1]/section[1]/para[1]”と”/papers[1]/docbody[1]/section[1]/para[2]”なら 2 となり，”/papers[1]/docbody[1]/section[1]/para[1]”と”/papers[1]/docbody[1]/section[2]/para[1]”なら 4 となる値である．また，$IntentW_i$ は共引用の関係にある引用同士の引用意図が同じか否かに応じて与えられる 0 以上 1 以下の定数であり，引用意図が同じ場合に大きくなる．つまり，類似度 $Sim(D_1, D_2)$ は，引用箇所間の距離が近く，引用意図が同じ共引用の回数が多いほど大きくなる値である．本手法では，類似度 $Sim(D_1, D_2)$ の大きい順に文書を提示する．

4 おわりに

本稿では，閲覧時アノテーションと呼ばれる情報を引用文書検索する際のトリガーに利用することを提案し，実験・評価した．また，本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を類似文書検索を例に示した．引用アノテーションに基づく文書間のネットワークは従来のハイパーリンクに基づくネットワークに比べて，より深い意味的関係を反映したものになることが予測され，類似文書検索以外にも様々な応用に適用できるであろう．

今後の課題としては，類似文書検索手法の評価などが挙げられる．

参考文献

[1] 林亮介，友部博教，大平茂輝，長尾確，オンライン文書閲覧時のアノテーションを用いた論文作成支援システム，情報処理学会第 69 回全国大会，2007.

[2] Katashi Nagao, Digital Content Annotation and Transcoding, Artech House Publishers, 2003.

[3] Marshall C. C., Annotation: from paper books to the digital library, Proceedings of Digital Libraries ’97, 1997.

[4] Small H., 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.