

## オンライン Web ページに基づく付箋アノテーションシステムとその応用

佐野 博之<sup>†</sup> 近藤 圭佑<sup>††</sup> 浅見 昌平<sup>††</sup> 大園 忠親<sup>††</sup> 新谷 虎松<sup>††</sup>  
名古屋工業大学 工学部 情報工学科<sup>†</sup> 名古屋工業大学 大学院 工学研究科 情報工学専攻<sup>††</sup>

### 1 はじめに

WWW 上の情報を参照する手段として、Web ブラウザのブックマーク機能というものがある。Web ページは複数のコンテンツから構成される傾向があり、一つの Web ページの中には、多様な内容が含まれる場合が大半である。本研究におけるコンテンツとは、Web ページ内で提供されるテキストや画像、企業広告などの、各種情報のことである。目的のコンテンツを閲覧するためには、ブックマークから Web ページを開いた後に、ユーザが手作業で目的のコンテンツを探す必要があり、煩雑である。したがって、Web ページの一部を参照するためのしくみが求められている。

本稿では、Web ページ内の一部を指定するための機能として、Web コンテンツへの付箋アノテーションシステムを提案する。付箋アノテーションの活用として、ユーザが貼付けた付箋は、そのユーザの嗜好の学習に用いられる。関連が深いと思われるコンテンツに対して貼付けられた付箋同士、システムが自動的に双向リンク [1] を作成することで、ユーザの Web 閲覧支援を行う。また、嗜好の似通ったユーザと付箋アノテーションを共有することにより、Web ページにおけるコンテンツの推薦を行うことが可能となる。

### 2 Web コンテンツへの付箋アノテーション

#### 2.1 付箋を用いた Web コンテンツへのアノテーション

人が書物を参照する際、何度も閲覧するであろう頁に対して、付箋を貼付けるという動作を行う。WWW 上の情報を参照する際には、Web ブラウザのブックマーク機能を用いて、何度も閲覧するであろう Web ページをブックマークに登録する。書物に対して貼付ける付箋は、貼付ける場所によって頁内の特定の場所を指定することが可能であるのに対し、既存のブックマーク機能は、Web ページのタイトルと URL のみを保存しておくだけのものであり、Web ページ内の特定の場所を指定することを目的としていない。

Web ページ内のコンテンツを指定する方法として、HTML にはアンカー要素の name 属性が存在する。一般的に、アンカーは Web ページの作成者によって指定されるものであるが、本研究では、Web ページの閲覧者が、Web ページ内のコンテンツに対して自由にアンカー要素を指定することが可能なシステムを提案する。ユーザ自身がアンカー要素を指定することにより、Web ページのタイトルと URL のみならず、ユーザが Web ページ上のどのコンテンツに着目しているかを保存することを可能とする。本システムでは、付箋を貼付けるというインターフェイスを用いて、ユーザがアンカー要素を指定することを可能にする。

書物に対して付箋を貼付ける場合、その付箋に対してメモ書きを行うことがある。それと同様に、本システムでは、貼付けた付箋に対してユーザが自由にコメントを付けられる機能を用意する。これにより、ユーザが Web ページ上のコンテンツに対して付箋を貼付け、その付箋を用いて、コンテンツに対してアノテーションを与えることを可能にする。付箋

という形でアノテーションを可視化することを、本稿では付箋アノテーションと呼ぶ。ユーザは、Web ページ内に含まれるテキストデータや画像などに対して、付箋によるアノテーションを行うことができる。

既存のアノテーションシステムとして、Annotea[2] がある。ユーザのアノテーションを活用し新たなサービスに応用しようとする点で、Annotea と本システムの目的は一致している。アノテーションを新たなサービスに応用するためには、ユーザから多くのアノテーションを獲得する必要がある。Annotea を利用するためには、ユーザは、専用の拡張機能をインストールした Firefox<sup>1</sup> や、W3C が開発したオープンソースの Web ブラウザである Amaya<sup>2</sup> を使用する必要がある。これらのブラウザを利用していないユーザは、Annotea を利用するために新たにブラウザを導入する必要があるため、多くのユーザからアノテーションを獲得することが困難である。本システムは、ユーザが使い慣れた既存の Web ブラウザに対して、プロキシを指定するだけで使用可能である。また、付箋という見慣れた形でアノテーションを行なうことが可能であるため、ユーザからの積極的なアノテーションが期待できる。

#### 2.2 システムの構成

本システムは、Web ブラウザ上で動作する付箋クライアントと、Web ブラウザのプロキシとして動作する付箋サーバから構成される。図 1 は、本システムの構成図である。付箋クライアントは、付箋を貼付けるためのインターフェイスをユーザに対して提供するシステムであり、JavaScript で実装されている。付箋サーバは、ユーザが貼付けた付箋を保存しておいたためのシステムであり、プロキシとデータベース、及びユーザの貼付けた付箋を監視するエージェントから構成される。

ユーザが本システムに Web ページ取得の要求を行うと、Web ブラウザはプロキシを通じて Web ページを取得する。サーバには、過去に付箋を貼付けた Web ページの URL と HTML が保存されたデータベースがある。プロキシはデータベースにアクセスし、ユーザがアクセスしようとしている Web ページの HTML が、データベースに保存してあるかどうかを問い合わせる。データベースに HTML が保存されている場合には、その HTML を Web ブラウザに送信する。保存されていない場合には、Web サーバから新たに HTML を取得し、取得した Web ページに対して付箋クライアントを付加し、Web ブラウザに送信する。

ユーザが新たに付箋の貼付けを行ったり、既存の付箋の更新を行った場合には、ユーザがページを切り替えた際に、付箋クライアントがデータベースの更新を行う。

#### 2.3 DOM ツリー解析に基づく付箋表示位置の決定

本システムでは、DOM ツリーを解析し、HTML 内の要素に対して直接付箋のタグを附加することにより、Web ページ内に存在するコンテンツに対して付箋を貼付けることができる。要素に対して付箋のタグを附加することにより、テキストや画像など、Web ページ内に存在する全てのコンテンツに対して付箋を貼付けることができる。ただし、テキストに対して付箋を貼付ける場合、文字列の長さによっては、ユーザが意図した場所から、付箋の表示が大きくずれてしまう可

<sup>1</sup><http://www.mozilla.com/firefox/>

<sup>2</sup><http://www.w3c.org/Amaya/>

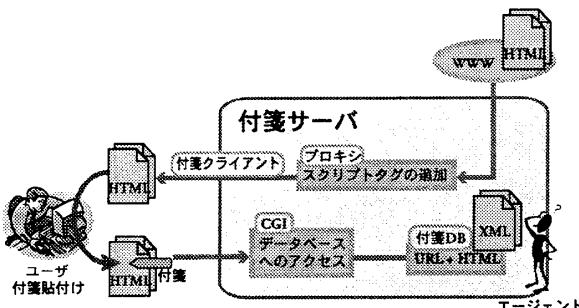


図1: システム構成図

能性がある。このため、テキストに対して付箋を貼付ける場合には、そのテキストをさらに細かい要素に分割し、付箋を貼付ける要素を決定する。

このような要素への貼付けを行わなくても、Webページ上でユーザがクリックした絶対座標を取得し、その絶対座標に付箋を表示するだけで解決できそうである。しかし、絶対座標を用いて付箋を表示すると、Webブラウザのウィンドウサイズやフォントサイズが変更された際に、コンテンツの表示位置が移動するが、ユーザが貼付けた付箋の画像の表示位置は元の場所から移動しない。そこで、付箋がもともと指示していたコンテンツからずれてしまうという問題が発生する。本システムでは、HTML内の要素に対して直接付箋を貼付けることにより、Webブラウザのレンダリング結果が変化しても、適切な場所に付箋を表示することが可能である。

#### 2.4 双方向リンクに基づく付箋アノテーションの共有

本システムでは、エージェントが付箋サーバ内部に常駐しており、ユーザの付箋貼付けを監視している。ユーザが付箋を貼付けると、エージェントは付箋が貼付けられた前後の文章を解析し、その解析結果から付箋の分類を行う。そして、関連が高いと思われる付箋同士に、自動的に双方向リンクを持たせる。

エージェントが付箋の分類を行う手法について述べる。付箋の分類を行う際の文書の類似度は、ベクトル空間モデルに基づき、2つのベクトルのコサイン尺度から求める。ベクトルは、各次元に索引語を割当て、各成分に索引語の評価値を割り当てるものとする。エージェントは、MeCab<sup>3</sup>を用いて付箋前後の文章を形態素解析し、文書の索引語を決定する。索引語の評価値はtf·idfの値とする。2つのベクトルを $d_1$ ,  $d_2$ とした場合、コサイン尺度は次の式で表せる。

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|}$$

コサイン尺度は、2つのベクトルの角度が小さい場合に、大きな値を取る。また、付箋を分類したクラスタについては、クラスタに含まれる付箋の文章のベクトルの平均ベクトルを、クラスタのベクトルとする。

エージェントは、ユーザが初めて貼付けた付箋の前後の文章を解析し、文書ベクトルを求め、それを1つのクラスタとする。以降、新しく貼付けられた付箋の前後の文章を解析し、文書ベクトルを求め、その文書ベクトルと、既存のクラスタの文書ベクトルとのコサイン尺度を計算する。このように、付箋が貼付けられた前後の文章にウェイトをかけて計算したWebページ間のコサイン尺度を、ここでは「付箋間の類似度」と呼ぶ。付箋間の類似度を計算する際に使用する文

書ベクトルは、次の式で表せる。ここで、 $w_{ij}$  は  $i$  番目の文書における語  $t_j$  の重み ( $j = 1, 2, \dots, M$ )、 $v_{ij}$  は  $i$  番目の文書の、付箋を貼った前後の文章における語  $t_j$  の重み、 $M$  は文書集合に含まれる異なる語の数、 $\alpha$  はウェイトである。

$$di = (w_{i1}, w_{i2}, \dots, w_{iM})^T + \alpha(v_{i1}, v_{i2}, \dots, v_{iM})^T$$

付箋間の類似度が閾値よりも大きければ、付箋をそのクラスタに追加し、クラスタのベクトル更新をする。付箋間の類似度が閾値を下回っていれば、新しくクラスタを作成し、付箋をそのクラスタに分類する。

ユーザが本システムの使用を開始した時点では、付箋の数は0であるため、クラスタ数も0である。ユーザが付箋を1つ貼付けた時に、初めてクラスタが発生する。その後付箋が増えるたびに、エージェントはクラスタリングを行い、同じクラスタに属する付箋が発生した場合には、双方向リンクの作成を行う。双方向リンクを作成する際には、付箋をノードとしてグラフで表現した場合、完全グラフとなるように双方リンクで結ぶ。

#### 3 考察

本システムを用いてコンテンツに対して付箋を貼付けることにより、エージェントが、それらのコンテンツを付箋を通じてリンクで結んでゆく。ユーザが重要だと思った箇所に付箋を貼付けることにより、コンテンツ同士が結びつき、ユーザが次に同じページを閲覧する際の情報収集の効率を改善することが可能である。

他のユーザが貼付けた付箋と、自分が貼付けた付箋の位置を比較することで、人気のあるコンテンツの発見を容易にすると思われる。ユーザ自身がWebページの重要なコンテンツを指定することは、人手によるコンテンツの評価が行われるということである。多数のユーザによる付箋アノテーションを獲得するコンテンツは、質の高いコンテンツである可能性がある。WWWのユーザは、情報収集をする際に検索エンジンを使う。通常の検索エンジンでは、サーチエンジン独自の検索アルゴリズム [3] を用いて、ページをランクしている。ユーザは、検索結果の上位にランクされたWebページを重要なものとみなす。上位にランクされたWebページは積極的に開くが、下位にランクされたWebページを開くことは少ない。たとえ検索結果が下位にランクされるWebページであっても、本システムでは、ユーザ自身がWebページの重要なコンテンツに積極的にアノテーションを行うことにより、そのWebページはより一層価値を増す。貼付けた付箋を多数のユーザ間で共有することで、検索エンジンに頼らない情報収集を行うことが可能である。

#### 4 おわりに

本稿では、Webページに存在するコンテンツへの付箋アノテーションシステムを提案し、その試作を行った。付箋アノテーションにより、既存の検索エンジンによる機械的な評価ではなく、ユーザ自身によるWebページの再評価が行われると考え、それらがリンクで結ばれていく有用性を示した。

#### 参考文献

- [1] 伊藤正詩、大畠忠親、新谷虎松，“ハイパーリンクの多機能化を目的としたBAC-Linkシステムの試作,” 合同エージェントワークショップ＆シンポジウム2006(JAWS 2006), Oct, 2006.
- [2] Jos Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, and Ralph R. Swick, “Annotea: An Open RDF Infrastructure for Shared Web Annotations,” 10th International World Wide Web Conference (WWW 2001), May, 2001.
- [3] 兼宗進, “検索エンジンの検索アルゴリズム,” 情報の科学と技術 Vol.54, No.2, pp.78-83, 2004.

<sup>3</sup><http://mecab.sourceforge.net/>