オフライン Web 技術に基づく付箋アノテーションシステム

高崎 隼 † 佐野 博之 †† 大囿 忠親 †† 新谷 虎松 ††
名古屋工業大学 情報工学科 † 名古屋工業大学大学院 情報工学専攻 ††

1 はじめに

本論文では，オフライン時の Web 閲覧を支援するための付箋アノテーションの応用について述べる．近年，Web ページは多様なコンテンツを含むようになった．それに伴い整理や確認のため Web ページ上に注釈などの付加情報を残す機能が必要とされている．今まで Web ページに対する付箋アノテーションシステムの研究は数多くなされてきた．しかしその多くは Web アプリケーションとして実装されている．付箋をつけた Web ページの活用はオンライン時に限らない場合もある．

ネットワークにつながっていない状態において，Web サービスおよび，そのサービスを支える技術を利用することをオフライン Web という．関連して Web サービスを提供するサーバサイドをリモート，対してサービスを受けるユーザサイドをローカルと呼ぶ．本システムにおいて付箋サービス提供するサーバをリモート，サービスを受けるユーザの Web ブラウザをローカルとする．付箋アノテーションシステムにオフライン Web 技術を取り入れることで，ネットワークが不安定な環境に対応し，また本来リモートのサーバが行うべき機能の一部をローカルに委譲することによるサーバの負荷軽減なども期待できる．オフライン環境における付箋システムを実装する上で，ローカルとリモートのデータベースの同期を考えなければならない．同期させる量を考慮すると完全にデータベース同期させることは困難である．

そこで本研究では，同期に注目したオフライン Web 技術に基づく付箋アノテーションシステムを提案する．本システムではオフラインを想定しているため実装は CGI プロキシによるスクリプト付加ではなく，ブックマークレットによる実装を行った．

2 オフライン時の Web 閲覧を考慮した付箋アノテーションシステム

2.1 オフライン時の Web 閲覧支援

オフライン時を想定した付箋アノテーションシステムを実現するにあたって，付箋をつけた Web ページの検索などの管理機能はローカルで行う必要がある．任意の時点でのオフラインへの切替に対応するため，ローカルとリモートのデータベースの同期が課題の一つとなる．そこで本研究では，ユーザの意思表示に基づく同期の手法を用いた．また，付箋の位置の指定について 2 通り考えられる．1 つは HTML の DOM ツリーの解析に基づき Web ページにアンカータグを埋め込む方法，もう 1 つは座標による付箋の位置を指定する方法である．アンカータグを用いる方法は，フォントやウィンドウサイズが変更された場合でも表示位置がずれないという長所をもつ．ただし，この長所を活かした同期を図るためには，対象が DOM 構造および含まれるコンテンツが同一な Web ページでなくてはならない．本システムにおける同期は，計算量の関係から同期の対象を URL のみで判別する．したがって本システムでは，付箋の位置は座標により指定する．

なお今回は本システムを実装するうえで Web ブラウザの拡張機能である GoogleGears [1]を使用した．GoogleGears は主にローカルサーバ，ローカルデータベース，ワーカプールという 3 つの機能を提供している．ローカルサーバは取得した Web リソースをおくサーバである．Web ブラウザが Web から情報を取得しようとした時，代わりに取得しておいた情報を読ませる機能をもつ．ローカルデータベースは SQLite という軽量のライブラリを利用したデータベース機能である．ワーカプールは JavaScript においてマルチスレッド化することができる機能である，この機能利用して作成されたスレッドを本稿ではワーカと呼ぶ．本システムでは，前提としてこの GoogleGears がインストールされた Web ブラウザを必要とする．

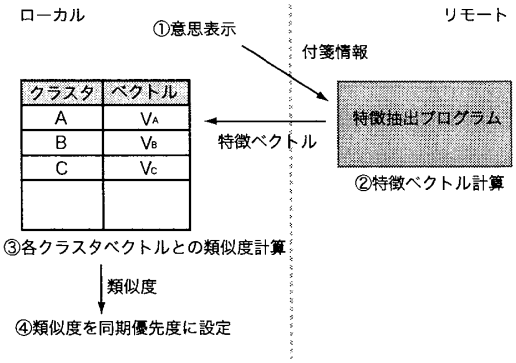

図 1: 同期優先度の決定

既存の付箋システムとして，FireFox アドオンである Internote[2]は，付箋情報を Web ブラウザに保存しておき，再訪問すると付箋及びその位置を再現する．Annotea[1] はユーザから付箋を収集し Web サービスに活かそうという研究である．また本研究室で試作した付箋アノテーションシステム [2] は付箋を共有し，関連ある付箋同士を双方向リンクで結び，コンテンツ推薦などに役立てることができる．差分として，本システムはオフライン環境においての付箋システムの使用を想定している．

2.2 付箋に基づく同期優先度の決定

システムの理想的な形として，ローカルとリモートのデータベースが完全に同期していることが望ましい．本システムにおいて同期させる対象となるのは，付箋情報および Web ページ情報である．付箋情報は ID，付箋ソース，更新日時，付箋を貼った Web ページの情報から構成される．Web ページ情報は，貼られた付箋数，更新日時，特徴ベクトルから構成される．これらの情報の完全同期を考えると，付箋情報が増えるに連れローカルとリモートの差分は大きくなり，ブックマークレット起動中の限られた時間の中では計算量の問題から完全な同期が難しくなる．そこで同期する頻度，範囲を適切に指定する必要がある．以下にユーザの意思表示に基づく同期優先度の決定過程を説明する．

本システムでは，ユーザが過去に付箋を貼った Web ページについて分類をする．分類は付箋を貼り付け時に求める Web ページの特徴ベクトルを利用する．特徴ベクトルとは，Web ページを形態素解析して得られた索引語と tf-idf 値からなるベクトルに，付箋を貼った要素内に含まれる索引語に重みを加えたものである．過去に付箋を貼ったことがある Web ページの場合は，過去の特徴ベクトルとの平均の値を新しい

Jun TAKASAKI, Hiroyuki SANO, Tadachika OZONO, and Toramatsu SHINTANI
Dept. of Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

[1] http://code.google.com/p/gears/

[2] https://addons.mozilla.org/ja/firefox/addon/2011

```
w : ユーザが興味をもった Web ページの特徴ベクトル
C={C0, C1, C2, ,,, ,Cm}  クラスタ集合
D={d0, d1, d2, ,,, ,dm}  クラスタのベクトル集合
R={r0, r1, r2, ,,, ,rm}  クラスタの類似度集合
wp : Web ページ w0 の同期優先度

procedure GetPriority(w,C,D)

begin
  n =  対象のWebページ数

  for (i=0 ; i < m ; i++)
        類似度をもとめる
      ri  = cos(w,di)

  クラスタを類似度 R で降べきの順にソート
  SortDown(C,R)

  for(j=0 ; j < m ; j++) {
  クラスタ Cj の Web ページを更新時間 wt で昇べきの順にソート
    SortUp(Cj,wt)
    for(k=0 ; k < |Cj| ; k++){
      wjp = n
       n=n-1
    }
}

end
```

図 2: 同期優先度アルゴリズム

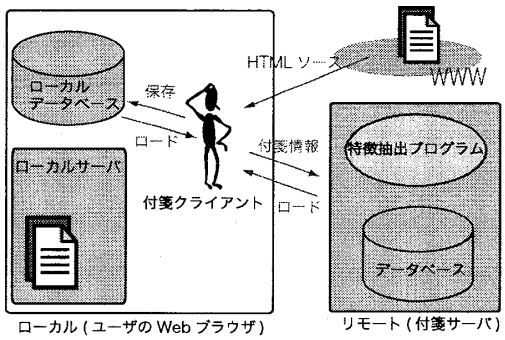

図 3: システム構成図

特徴ベクトルとする．この特徴ベクトルを元に Web ページの分類をする．Web ページの特徴ベクトル同士のコサイン尺度を求めて，類似度としてクラスタリングを行う．はじめはクラスタは存在せず，ユーザが付箋を貼った Web ページが 1 つならクラスタは 1 つとなる．2 つめ以降は閾値を用いる．既存のクラスタのベクトルと比較して閾値以上ならクラスタに追加，未満ならば新規クラスタを作成する．以降ユーザが Web ページに付箋を貼付ける度にクラスタリングが行われる．

ローカルに存在する付箋を貼った Web ページのクラスタに基づき同期の優先度を決定する．その流れを図 1 に示す．本システムの付箋インターフェースにはボタンが設置されており，ユーザは興味の強いコンテンツに貼られた付箋でこのボタンを押し意思表示を行う．意思表示を行った Web ページの特徴ベクトルと既存のクラスタの特徴ベクトルとの類似度を求めてクラスタ優先度とする．クラスタ優先度の高いクラスタから，クラスタ内の優先度を求める．Web ページの優先度は更新時刻の古い順に高く設定する．

3 システム構成

本システムは，ローカルの付箋クライアント，ローカルサーバ，ローカルデータベース，そしてリモートの付箋サーバから構成される．図 3 は，本システムの構成図である．付箋クライアントは，付箋のインターフェースの表示，およびサーバとの通信を行うプログラムである．ローカルサーバは，付箋を貼った Web ページ，および付箋管理ページを保存する．ローカルデータベースは付箋情報および，付箋を貼った Web ページの情報を格納する．付箋サーバはデータベースと特徴抽出プログラムで構成される．

ユーザは付箋を貼りたい Web ページにおいて，ブックマークレットを起動することでシステムを使用することができる．ブックマークレットによって起動した付箋クライアントはローカルおよびリモートのデータベースを，Web ページの URL をキーとして検索し付箋情報をロードする．同時にローカルとリモートのデータベースを同期させる．

ユーザが Web ページに付箋を貼るとローカルのデータベースに付箋情報および Web ページ情報を保存，更新する．オンライン状態では付箋が貼られるときリモートの付箋サーバにも付箋情報が送信される．付箋サーバでは，特徴抽出プログラムが送信された付箋情報をもとに Web ページの特徴ベクトルを求め，付箋情報とともにデータベースに保存，更新する．Web リソースの取得はブックマークレット起動時に閲覧中の Web ページに対して行われる．

付箋を管理するシステムはローカルサーバにおかれる．この管理システムはローカルおよびリモートのデータベースを検索し結果を表示する．表示された表からは該当する Web ページにリンクが貼られ参照することができる．

4 考察

付箋情報について，付箋が貼られた Web ページを特徴ベクトルから分類し，同期の優先度を決めることでブックマークレットの起動中の短さを考慮した同期を可能にした．他の同期優先度の決定法として，更新時刻が古い順に高く設定する方法が考えられる．この方法の利点は，計算がないため，限られた時間を全て同期処理に使うことができる点である．付箋を貼った Web ページ数が少なければ，ローカルの付箋情報は一様な新しさ，かつ，どの Web ページにおいても最新に近い情報を保持できる．しかし付箋を貼った Web ページが増えるに連れ，一様な新しさが次第にばらつき，望んだ Web ページの同期がなされていない場合が増える．したがって付箋を貼った Web ページが増えるにつれ，本システムで用いた手法の方がユーザにとって利便性が高いと考えられる．

本稿では，オフライン Web 技術に基づいた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．ローカルとリモートのデータベースの完全な同期が難しいシステムにおいて，ユーザの意思を反映させる同期を行うことで，単に更新時刻を元に同期を行う場合と比較して，Web ページ数が増えた場合において有効であることを示した．

参考文献

[1] 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.

[2] 佐野博之，浅見昌平，大囿忠親，新谷虎松，"Web エージェントを用いた Web コンテンツへの付箋アノテーションの実現"，合同エージェントワークショップ＆シンポジウム 2007，Oct，2007.