A-012

Webサイトのアドレス圧縮によるクラスタリング

Web Site Clustering by the Addresses Compression

佐藤 哲 †
Tetsu R. Satoh†

1. はじめに

インターネットを利用した情報検索は，日常生活の一部として既に定着している．しかしインターネットを用いた情報検索に対し，少なくとも次のような二つの問題が認識されている．

(1) ユーザ情報が Web サイトに取得・蓄積されることによる個人プライバシー侵害の可能性

(2) “情報の洪水” による，的確な情報入手の困難性

そこで本研究では，個人を特定する情報を用いずにユーザが望む情報を提供するための，個人の嗜好に基づき Web サイト群のクラスタを自動的に作成し，情報提供に繋げる一手法を提案する．本手法は各ユーザの Web サイト閲覧履歴を初期クラスタとし，クラスタ間の類似度を判定し的確にクラスタを統合し，大きなクラスタを作成していく．クラスタ群が作成されると，クラスタに含まれる Web サイトとユーザの初期クラスタの差分を調べることで Web サイトの推薦システムが構築できる．また，初期クラスタ群自体によりユーザを認識するので，ユーザを特定する ID のようなものは必要とせず，プライバシーに関する問題は生じない．従って，上記の 2 つの問題を同時に解決することができる可能性がある．

2. モデルを仮定しない類似度判定

本研究では図 1 のような，ユーザが閲覧した Web サイトのアドレスを時系列で結合した文字列を Web サイトアドレスのクラスタと呼ぶ．従ってクラスタを分類するためには，クラスタ間の類似度を判定する必要がある．文字列や画像など，類似性を定義することが難しいメディアに対する類似度の計算方法は莫大な研究があるが，本研究では情報論的な理論背景があり，かつモデルを必要としない圧縮アルゴリズムを用いた非類似度計算法 (Compression based Dissimilarity Measure:CDM)[1] により，クラスタ間の類似性を判定する．

一般に記号列 x，y に対し，CDM は次のように定義される：

$$CDM(x,y) = \frac{C(xy)}{C(x)+C(y)}. \qquad (1)$$

ここで，$C(x)$ は圧縮プログラムにより記号列 x を圧縮した結果のサイズを表し，xy は記号列 x と y を単

† 株式会社オプトリンクス，Optlynx Co.,Ltd.

図 1: Web サイト閲覧履歴例

純に結合することを意味する．CDM は非類似度が計算されるため，類似度を判定する際は CDM の値が最も低いものが最も類似性が高いと判断する．

本研究では，圧縮プログラムとして bzip2 ライブラリ † を用いた．例として，次のような 3 つキーワードを順次 Google 社の検索エンジンを用いて 3 回検索した場合のアドレスを考える．

(A)パソコン，計算機，ソフトウェア
(B)コンピュータ，ソフトウェア，価格
(C)FIT，大会，〆切

例えば (A) の場合，アドレス列は，次のようになる．

```
http://www.google.co.jp/search?
sourceid=mozclient&num=20&ie=utf-8
&oe=utf-8&q=%E3%83%91%E3%82%BD%E3%82
%B3%E3%83%B3!http://www.google.co.jp/
search?sourceid=mozclient&num=20&
ie=utf-8&oe=utf-8&q=%E8%A8%88%E7%AE
%97%E6%A9%9F!http://www.google.co.jp/
search?sourceid=mozclient&num=20&ie=utf-8
&oe=utf-8&q=%E3%82%BD%E3%83%95%E3%83%88
%E3%82%A6%E3%82%A7%E3%82%A2
```

この 3 群のアドレス列に対し，CDM 値を計算した結果を表 1 に示す．圧縮アルゴリズムによっては必ずしも対称性 CDM(x,y)=CDM(y,x) が成り立つことは保証されないが，今回は小数点以下 6 桁までの計算では対称的になった．表 1 より，同じアドレス列同士の CDM 値は確実に最小になり，同一のキーワード「ソフトウェア」が入っている (A) と (B) の CDM 値が，(A) と (C)，(B) と (C) の CDM 値に比べて小

† http://www.bzip.org/

表 1: CDM 計算例

	(A)	(B)	(C)
(A)	0.588384	0.615776	0.654891
(B)		0.584615	0.649315
(C)			0.617647

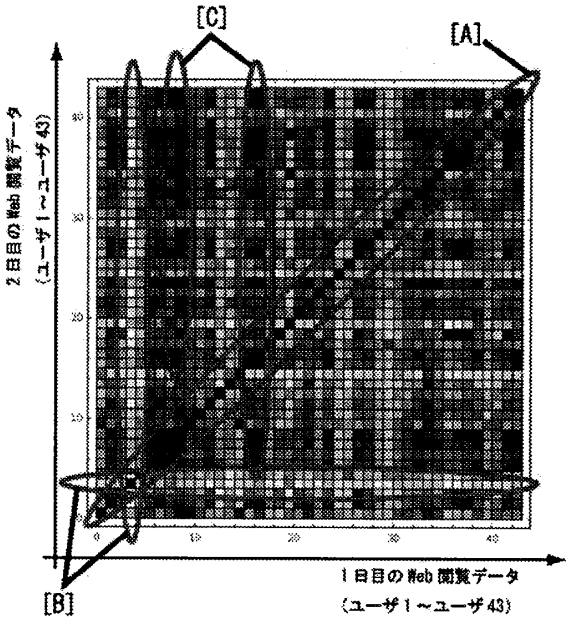

図 2: 類似度マップ作成例

さくなっていることが確認できる．従って，文字列間の類似度が適切に計算されていることが分かる．

3. クラスタリングアルゴリズムと実験例

類似度判定が適切であれば，クラスタリングは類似度判定とクラスタ統合の繰り返しで実行できる．例えば Web サイトアドレスが次々と入力されてくるオンライン・クラスタリングの場合，次の処理を繰り返すことになる．

(1) 既に存在する全てのクラスタと，入力されてきたアドレスとの類似度を計算する
(2) 最も類似するクラスタに，入力アドレスを追加する

ただし Web サイト分類に対するオンライン・クラスタリングは結果の評価が難しいため，ここでは時系列でクラスタ間の類似度を計算し，類似度マップを作成した上でどのような分析が可能かを説明する．図 2 は，43 名の被験者から 1 日の Web 閲覧履歴を提出してもらい，さらにその 3 日後に 1 日の Web 閲覧履歴を提出してもらって，そのデータ間の類似度マップを作成した例である．類似度マップは各被験者の 1 日分の Web 閲覧履歴を 1 クラスタとし，類似度が高いほど濃い色で表示されるよう可視化した．横軸が 1 日目のユーザ 1 からユーザ 43 に対応するデータであり，縦軸が 2 日目に対応する．ユーザ番号は類似度判定・クラスタリングには必要ないが，正確性を判定するためにこの例ではユーザに対応する番号を使用している．その結果，例えば以下のようなことが確認できる．

- (A) に示されるように，データ取得日が違い，Web 閲覧履歴が同一では無いにも関わらず，同一ユーザ同士の類似度は高くなるよう計算されている
- (B) に示されるように，特定の被験者は 2 日間とも他のユーザと類似度の低い Web サイトを利用している
- (C) に示されるように，特定のユーザが 1 日目に閲覧したサイトと類似度が高いサイトを，他のユーザが 2 日目に閲覧している

異なる日のデータ間の類似度計算による同一ユーザの判定に失敗してる計算例は 43 名中 3 例あり，正解率は約 93 %となった．従ってこの正解率の範囲なら，ユーザを識別する情報を使用せずに特定ユーザのクラスタ時系列追跡が可能である．

4. おわりに

本研究では，Web サイトのアドレス群に対し圧縮アルゴリズムを用いた類似度判定を利用し，ユーザが閲覧するサイト群間の類似度を判定したりサイト群をクラスタリングする手法を提案した．ただし実用化するためには次のような課題が残されている．

- 圧縮アルゴリズムを利用しているため，リアルタイム処理が難しい．圧縮・解凍を毎回繰り返すことを避けるアルゴリズムの開発が必要である．
- クラスタの個数を動的に変更したり上限を定めることが難しい．例えば既存のどのクラスタとも類似度が高くない入力情報があれば新たなクラスタとして登録したり，クラスタ数が一定の個数まで増大したら新規クラスタの作成を抑制する仕組みが必要である．

前者の問題に対しては情報のキャッシングを使い，後者に対しては新規クラスタを作成するかどうかを MDL[2] 等の情報量基準を用いて判定する研究を進める予定である．

参考文献

[1] Keogh, E., Lonardi, S. and Ratanamahatana, C. A.: Towards Parameter-Free Data Mining, *Proc. 10th ACM Int. Conf. Knowledge Discovery and Data Mining*, pp. 206–215 (2004).

[2] Rissanen, J.: Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, Vol. 30, No. 4, pp. 629–636 (1984).