ウェブ閲覧履歴に基づく ウェブサイトの類似性判別

―SVDとICAによる 次元削減効果の比較―

青木友花、松田晃一（大妻女子大）、竹内彰一（ソネット）

発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

背景と課題

- 日本のインターネットの世帯普及率：2013年 86.2%（ITU）
- ウェブ閲覧は常にネット利用目的の上位2位以内

- ウェブ閲覧履歴
  - ビッグデータ
  - 様々な利活用

本発表では

- オンライン広告での活用を想定
  - ウェブサイトの類似性

背景と課題

- 従来ウェブ広告では、ウェブサイトを仲介として広告内容とユーザの興味・関心のマッチングを実現し、高い収益を実現
- この方法では、ゴルフに関心のあるユーザがゴルフ場予約サイトを閲覧している時に新しいゴルフクラブの広告は提示できても、不動産投資サイトを閲覧している時にはゴルフクラブ広告の提示はできない

A

B

背景と課題

- 従来ウェブ広告では、ウェブサイトを仲介として広告内容とユーザの興味・関心のマッチングを実現し、高い収益を実現
- この方法では、ゴルフに関心のあるユーザがゴルフ場予約サイトを閲覧している時に新しいゴルフクラブの広告は提示できても、不動産投資サイトを閲覧している時にはゴルフクラブ広告の提示はできない

A

Bを閲覧する80%のユーザーがAも見ているならば

B

サイトの類似性

- 類似性
  - 意味の類似性
    - 内容の意味的関連の深さ
    - テキストの意味解析
  - サイトのターゲットユーザの共通性
    - ユーザの重複の大きさ
    - ウェブ閲覧履歴解析

発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

関連研究

- ウェブ閲覧履歴のマイニングは多数研究されてきた
  - 大塚・喜連川2006
- ウェブサイトのクラスタリングは大きなトピック
  - 様々な応用：オンライン広告、ECサイト　等
  - クラスタリング手法：古くはSVD、新しくはICA
    - 小規模ユーザのウェブ閲覧履歴にICAを適用する研究（鶴原）
    - 特定サイト内の顧客のページ遷移履歴にk-meansやSVDを適用する研究
- 本研究
  - オンライン広告ログ（複数サイトを含む大規模閲覧履歴）にSVDとICAを適用し、サイト間の類似度を求め、比較した

発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

実験データ

- ウェブ閲覧履歴データ：ソネットメディアネットワークス株式会社（以下SMN、ネット広告会社）提供
- ウェブ広告の2013年の一ヶ月間(5月)の掲載ログを以下の形で使用。

★ユーザ・・・特定IDで判別

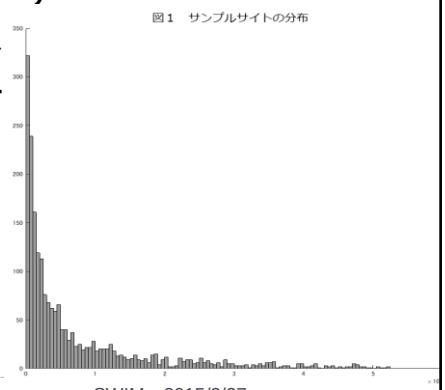
★ウェブサイト・・・SMNと契約しているURL

★何回訪れたか（閲覧回数）

ユーザID	サイトURL	閲覧回数
xxx	aaaaa	10
yyy	bbbbb	23
zzz	ccccc	2

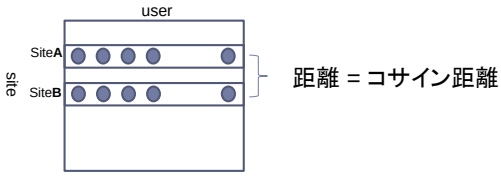
データの絞り込み

- 5万人をランダムに抽出
- 1ヶ月の広告リクエスト回数10,000回以上でかつ最低1度は広告をクリックしたユーザ
- 1ヶ月の閲覧数の多いサイトから少ないサイトまでカバーするようなサイト群
- 2084サイト
- 19871人

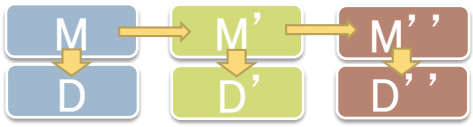


行動履歴行列

- 2084サイト×19871人
  - 類似度計算の基礎となるデータ
- 2084の各サイトは19871次元のベクトルで表現される



- サイト間距離をコサイン距離で定義する



発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. **評価指標**
5. 分析結果
6. まとめ

平均相互訪問確率

- 定義　・・・ペアとなるAとBのサイトがある時、

平均相互訪問確率　$\frac{1}{2}\left(\frac{Nab}{Na}+\frac{Nab}{Nb}\right)$

Na ・・・ Aを訪問した人がBを訪問する人数(A→B)

Nb ・・・ Bを訪問した人がAを訪問する人数(B→A)

Nab ・・・ 両サイトを訪問する人数

平均相互訪問確率が高い → 片方のサイトを訪問したユーザがもう一方のサイトも訪問する確率が高い

距離集合の評価指標

- 評価指標
  - サイト間距離の短い上位6000のペアサイトに関して、サイト間の平均相互訪問確率とサイト間距離の相関とする
  - －1に近いほど良い性能であることを示す
- モデル性能
  - 19871人から計算した相互訪問確率との相関
- 予測性能
  - 新規19620人から計算した相互訪問確率との相関
- 19871人でのモデル性能と予測性能　（ベースライン性能）
  - モデル性能　＝　−0．13
  - 予測性能　＝　−0．05

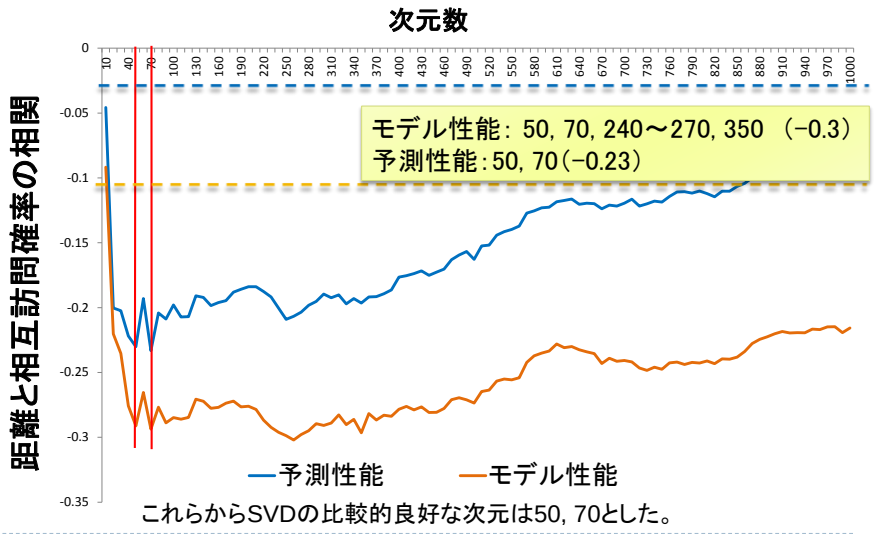
発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

次元削減

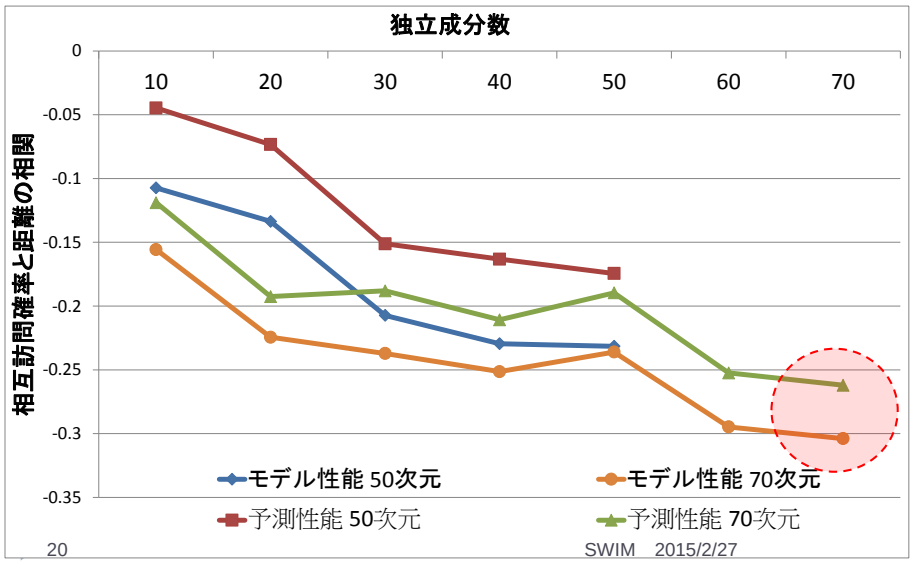
1. SVD
   - 19871次元を更に低次元に圧縮する
   - 最適な次元を調べる
2. ICA
   - SVDで求めた最適次元で最適な独立成分数を求める

ＳＶＤのモデル性能と予測性能

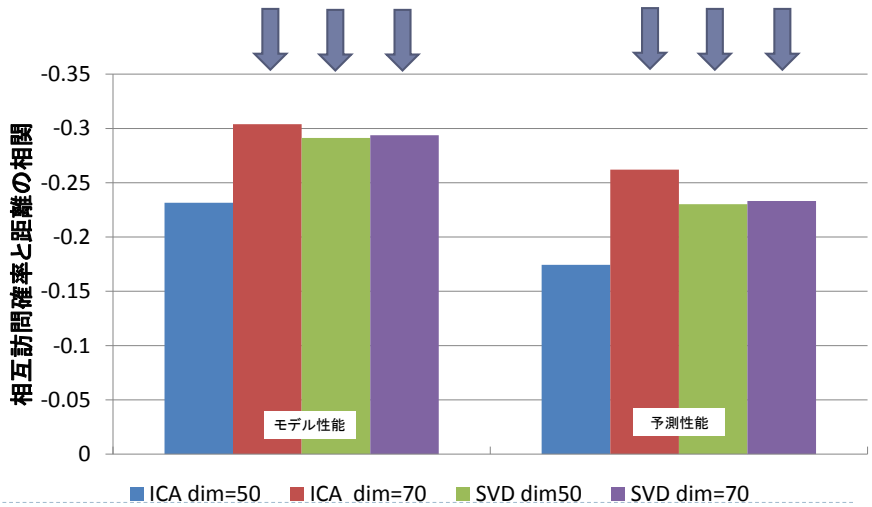


これらからSVDの比較的良好な次元は50, 70とした。

ICAのモデル性能と予測性能

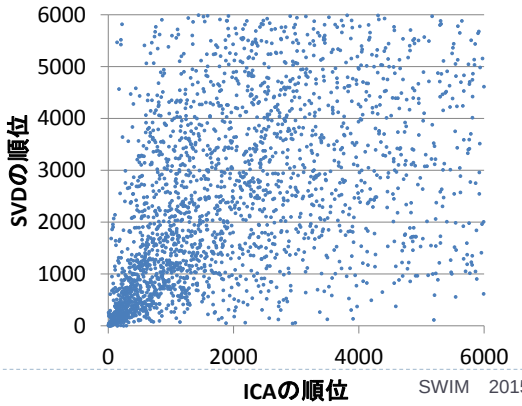


SVDとICAの性能比較



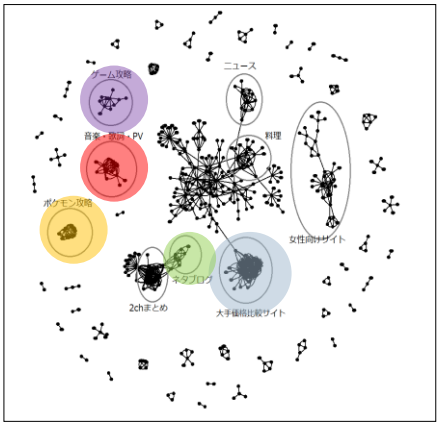
70次元における順位散布図

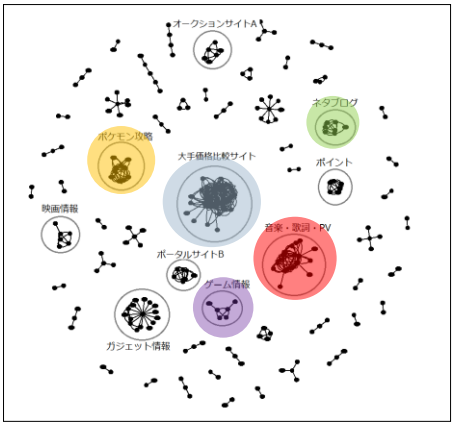
- SVD70次元とICA70次元（独立成分数70）の比較
  - それぞれの距離の近いサイトペア上位6000を比較
  - 両方で上位6000位にランクインしているサイトペアは2402組(40%)



類似サイトの可視化

▼SVD70次元　　▼ICA70次元





距離の近い上位6000サイトペアの近接関係をグラフ表現
個別の訪問者数が10人以下、ペアサイトの同時訪問者が9人以下のペアは含めていない

発表構成

1. 研究背景と課題
2. 関連研究
3. 実験データ
4. 評価指標
5. 分析結果
6. まとめ

まとめと今後の課題

- ウェブ行動履歴からサイトの類似性を計算する方法について、SVDとICAにおいてそれぞれ複数のパラメータ下で検討
  - 次元削減によるデータサイズ削減と性能向上
- SVDでは50、70次元で最良効果
  - 次元圧縮（1/200以下）によるノイズフィルター効果
- SVD,ICAそれぞれの類似サイト群
  - 抽出した類似サイト群に意味的類似性が見られるものが多かった
  - SVDとICAの差が確認できた
  - サイトのターゲットユーザの類似性に相当する類似性は確認できなかった
- 今後の課題
  - SVDとICAに見る共通部分と独自部分の相違点の解明
  - サイトのターゲットユーザの類似性に相当する部分の追求