～個人情報の安全性と利用性を両立させるために～

個人情報のSEM(検索エンジン広告)価格に基づいたk-匿名化手法の提案

小栗 秀暢　曽根原　登　　国立情報学研究所・総合研究大学院大学

どんな研究？

ビッグデータを安全に利用するためには、匿名化の技術が不可欠です。

一般的な匿名化技術である「k-匿名化処理」は、個人を特定するための識別子の数が(k-1)個以上存在するまで、選択肢を抽象化します。

ですが、実際に匿名化を行うと、膨大な抽象化候補と選択肢の組み合わせが発生し、安全性と情報損失量のみで判定された、利用性の低いデータを作成してしまう場合があります。

本研究は、匿名化の抽象化候補に対して「SEM価格」を用いて重み付けし、安全性と広告価値の両立した匿名化処理を実現するアルゴリズムを考案することで、計算量の削減やデータ量の圧縮などを実現します。

何がわかる？

検索エンジンには、インターネットユーザが必要とする言葉が入力され、その入力した語に対応する企業広告が紐づけられています。

つまり、ある言葉に対する検索エンジン広告の価値は、インターネットマーケティングの分野でどれだけ必要とされているかの指標になると言えます。

多くの企業や政府が、個人情報を匿名化して公開する場合、ある「お仕着せのパターン」による匿名処理を実施しがちで、結果的に誰からも利用されない情報が公開されることとなります。

社会のニーズに合わせた、適応的な匿名化処理によって、マーケティング的価値が高い情報のみを、安全に提供・流通させることができます。

状況設定

ある個人情報を含むテーブルTに対して、k-匿名化処理を実施するという状況を設定します。

例えば、年齢と住所を匿名化するため、年齢の抽象化候補の辞書 P1 と 住所の抽象化候補の辞書 P2 を用意します。

各抽象化候補となる語のSEM価格を求め、対応する人数に合わせて金額を合計し、抽象化の優先度を設定します。

この匿名化アルゴリズムは、情報損失量で判定樹を分岐させるのではなく、SEM価格総額の高い順番に処理を行い、安全性とマーケティング価値の両方が高い匿名化処理を優先的に実施します。

T：個人情報が含まれるテーブル

ID	年齢	住所
1	12	東京都千代田区
2	14	大阪府大阪市
3	16	神奈川県川崎市
4	12	千葉県浦安市
：	：	：

P1：「年齢」を抽象化するパターン辞書

Type1	10代、20代‥	1
Type2	小学生、中学生‥	2
Type3	未成年、成年	3
：	：	：

語のSEM価格による処理の優先順位

P2：「住所」を抽象化するパターン辞書

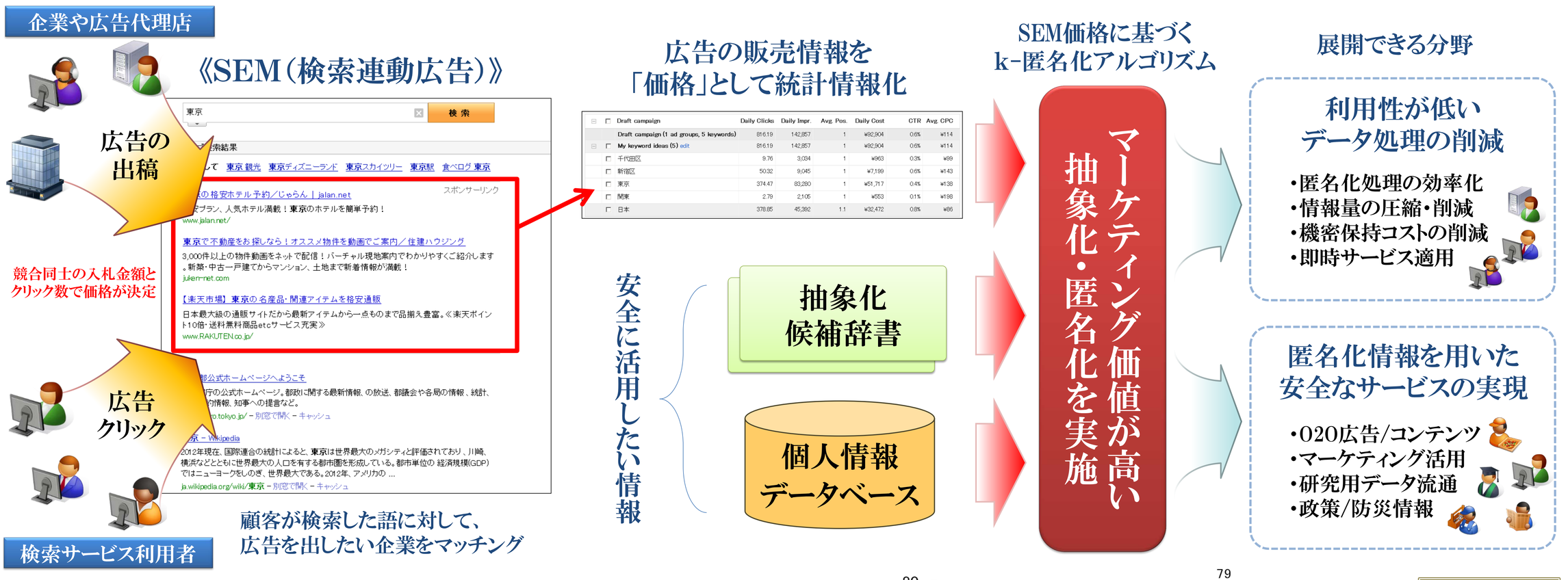
Type1	東京都、大阪府‥	1
Type2	関東、関西‥	2
Type3	東日本、西日本	3
：	：	：

語のSEM価格による処理の優先順位

それぞれの選択肢の語のSEM価格で重み付け

k-匿名性が保たれマーケティング的に一番価値が高いデータに変換する

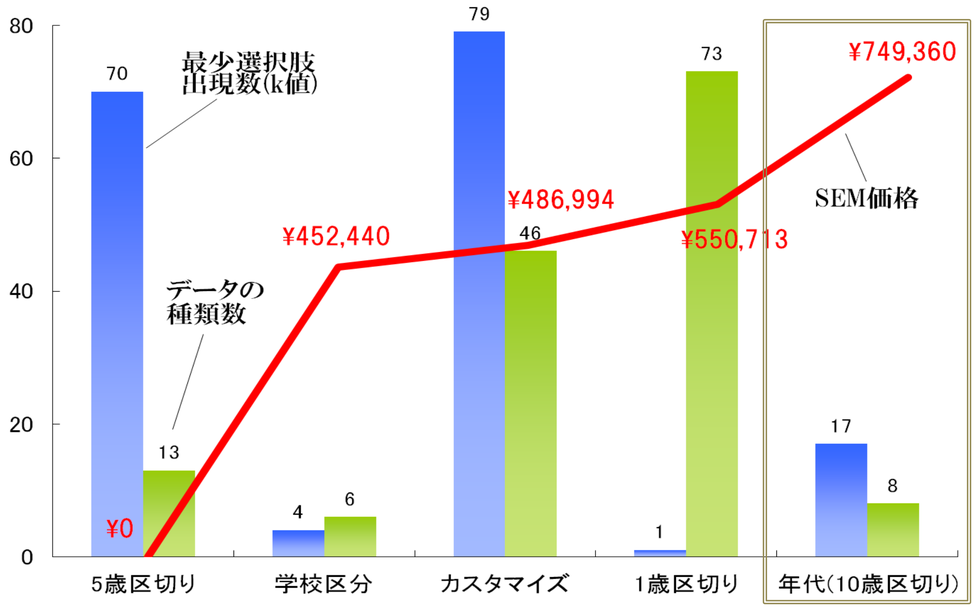
研究内容



個人情報にて頻繁に利用される、年齢/住所/職業　などの情報は、ある程度抽象化された方が、安全性が高く、マーケティング的な価値も高くなる場合があります。

例えば、「年齢」の項目は、1年区切りで情報を扱うよりも、10代、20代‥と区切った方が、広告的な価値が高い場合があることが判明しています。(右グラフ参照)

場所や時間、文化や流行などによって変化するユーザのニーズに対して、適応的に匿名化処理を行うことは、データ量の圧縮や管理コストの低減を実現するだけでなく、リアルタイムのWEBサービスへの活用や防災情報などへ展開することが期待できます。



個人情報を「年代」区切りに抽象化すると、匿名化レベルも高く、SEM価格も高いレベルとなる。

連絡先：小栗 秀暢　国立情報学研究所　情報学社会相関研究系　曽根原研究室　／ニフティ株式会社
TEL：03-4212-2731　FAX：03-4212-2120　Email：oguri_hidenobu@nii.ac.jp