卒業論文2016年度

嘘データの用途と需要その応用例に関する研究

慶應義塾大学総合政策学部４年
野村隼斗

嘘データの用途と需要その応用例に関する研究

概要

嘘データは様々な場面で生成され、利用されている。
本論文は、嘘データの用途と需要について調査し、既存の嘘データ作成法を紹介したものである。
さらに、その応用例について考察した。

目次

第一章 序論

本章では研究の背景と目的について説明する。

1.1背景

嘘データは様々な場面において求められる。小説を書くときの設定、信頼の置けないサービスにユーザー登録するために入力する個人情報、会員制webサービスを作成時のテスト用大量の個人情報、強力なパスワード生成に用いるため、現代において嘘データは多様な場面で必要とされていると考えられる。

1.2目的

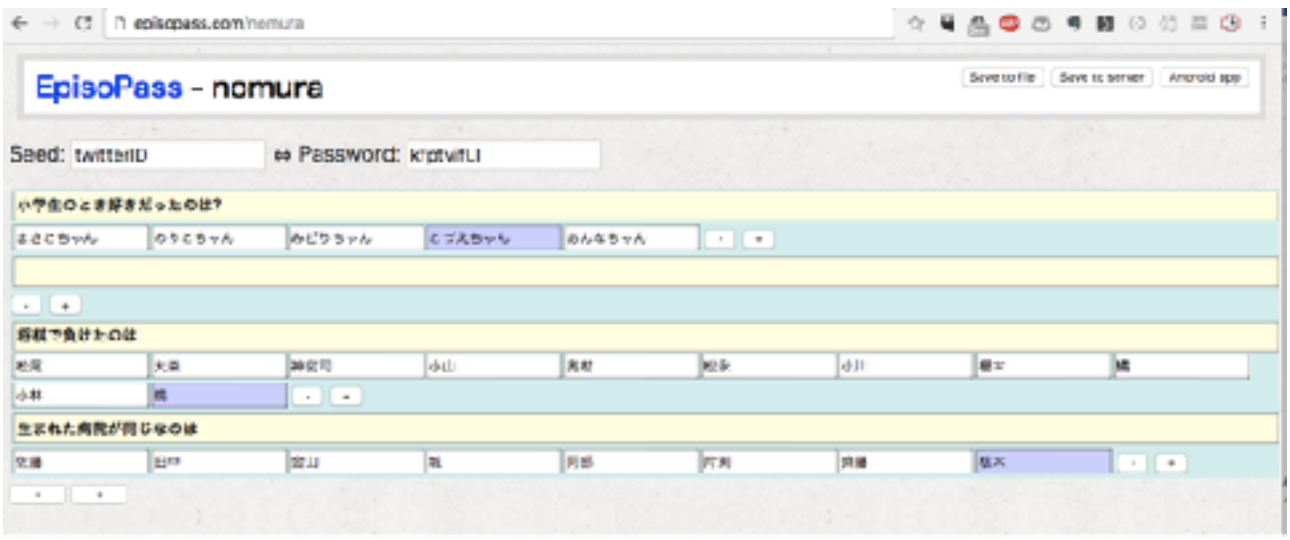
嘘データはしばしば生成が煩雑である。
小説に用いるための登場人物の氏名や住所、webサービスのテストに用いるための氏名、アカウント名、電話番号、メールアドレスなどを手作業で作るのは苦行であると考えられる。

偽データの需要と既存の偽データ生成サービスの調査、そしてその応用例の考察が本論文の目的である。

1.3用途

1.3.1 EpisoPass[1] [2]

　EpisoPassとは慶應義塾大学環境情報学部 増井俊之教授が開発し、運用しているパスワードを自動生成するシステムだ。
人間は新しく覚えた情報を必ず忘れるものであるならば、新しいパスワードを考えたり覚えたりする努力は不毛である。一方、忘れることがないエピソード記憶を誰もが持っているのであれば、そのようなものを利用してパスワードを生成する方が妥当であると思われる。そういったアイディアに基づいて増井教授は「EpisoPass」を開発された。



上図は私がtwitterのパスワード生成にEpisoPassを用いたものである。
ここでは、シード文字列として「twitterID」を指定しており、３つの秘密の質問に対する回答選択に応じて「kfptvifLI」のようなパスワード候補が生成されます。異なる答えを生成したり、異なるシードを指定するとまったく異なる文字列が生成されます。
質問は私の古いエピソード記憶を利用しているため、私が忘れる可能性は極めて低いですが、他人にはわかりづらいものとなっている。
EpisoPassにおいて重要な要素の一つが認証の際の偽データによる誤答選択肢の生成である。
現行のEpisoPassにおいては誤答選択肢のデータは自らで手入力しなくてはならない。
誤答選択肢が存在しないことにはこのサービスは成立しないためこの工程を省くことは出来ない。
同じような重みの情報が上手に自動で生成されればさらに便利なサービスと成ると考えられる。

1.3.2テスト(会員制サービス)

会員制Webサービスのテスト用に一定量のまとまった個人情報があると有用であると考えられる。

現状では、なんちゃって個人情報[3]や疑似個人情報データ生成サービス[4]を用いて大量の架空個人情報を生成し、これをテストに用いることが可能である。

1.3.3小説,漫画,記者のなりすまし等

　小説や漫画などのフィクション作品における登場人物の個人情報や住所等が必要になる場合がある。
これらは架空のものである場合が多くわざわざ生み出さなくてはならないものだ。
含意のある名詞を生成したい場合はここに当てはまることはないが、「少年A」に個人情報を付与したい場合に有用であると考えられる。
また、新聞記者の友人は危険人物に取材を行う際、しばしば市民等になりすまし取材を行うようである。このとき行う個人情報生成が煩雑なようなので、上手に架空の個人情報（氏名、職業等）を簡単に生み出すことができる個人情報生成サービスがあれば、その際に有用である。

第２章 生成方法

　本章では既存の嘘データ生成法を紹介する。

2.1既存サービス

2.1.1なんちゃって個人情報[3]

　ダミーの個人情報を大量に出力できるwebサービス。
名前、ふりがな、メールアドレス、性別、年齢、誕生日、婚姻の有無、血液型、都道府県、電話番号、携帯番号、カレーの食べ方の１３項目のうち任意の項目を最大5000人分生成し、HTML、XML、CSV、タブ区切りテキストの形式でダウンロードすることができる。これは、ユーザー登録を必要とするwebサービスやアプリケーションを開発する人間にとって、あれば有用な場面が生まれうるwebサービスである。

2.1.2 疑似個人情報データ生成サービス[4]

　なんちゃって個人情報[2]と同じようにwebサービスやアプリケーションを開発する人間にとってあれば有用な場面があるwebサービスであるが、こちらは生成項目が、連番、氏名、性別、電話番号（一般）、電話番号（FAX）、電話番号（携帯電話）、メールアドレス、住所、出身地、血液型、乱数、パスワードとなっており、なんちゃって個人情報とは異なっている。

2.2 既存の類語検索

2.2.1 @ymrl式[5]

　@ymrl氏が本名っぽい文字列を作るために使った手法

名字データベース等からクロールしてデータを加工、本名っぽい文字列を生成し、twitterやmixi上で使ったものである。

2.2.2 らいばるサーチ やーやら[6]

「らいばるサーチ：やーやら」は、京都大学情報学研究科の田中克己教授、大島裕明特定助教等のグループが開発したものであり、ユーザーが指定した特定の人物・地名・物の同位語（＝ライバルや仲間にあたる語）や話題語をウェブから検索してグラフに表示するものだ。例えば「織田信長」と入力すれば、その同位語や話題語が具体的にグラフ表示されるというものである。

第三章 考察

本論では今回の研究を踏まえての考察を述べる。

3.1応用例

3.1.1　EpisoPass改良

同じ頻度・重みの情報をより簡単に生成するサービスがあれば嘘データが必要な場面において非常に有用であると考えられる。
殊、エピソード記憶をもとに運用されるEpisoPassにおいては問題として名前を生成する場合に、姓であればDBから正解と同程度の世帯数である姓を抽出し、選択肢として生成する。
同様に住所・地名の嘘選択肢を生成する場合はapiを用いて、正解と同じような緯度経度から地名を抽出し、選択肢として生成する。

自動で同じ重みの不正解である肢を生成することができる機能が付加されれば、EpisoPassはよりその利便性を増す。

3.1.2 試験問題の嘘選択肢

現在ほとんど全ての問題の選択肢を人間が生成しているが、多肢選択式問題などはデータベースを用意することで、正答と同カテゴリ・頻度・重みの情報もった誤答をに生み出すことが可能であるはずである。 これが実現すれば、試験問題作成現場において利用したい人間は少なくはないはずである。
岩手県立大学大学院の菅原遼介氏、 高木正則氏は、記述式問題の誤回答を用いて、誤答選択肢を生成するシステムを開発したようである。[7]

3.2結論

本論文において、嘘データの需要・用途について調査し、嘘データ生成のための既存サービスを紹介した。 また応用例についても述べた。
EpisoPassにおける認証の選択肢という高尚なものから、webサービスにおけるテスト用、また小説、漫画や成りすまし用のデータというものまでその需要は多岐にわたると考えられる。

これからも嘘データがの需要はあり続けると考えられるし、よりよい嘘データを簡単に生成することができるサービスが生まれる必要があると推察する。

謝辞
本論文を執筆するにあたり、担当の増井俊之教授には多大なるご指導、ご支援を頂きました。
また、研究室の皆様も多くのアドバイスを下さいました。この場を借りて感謝の意を表します。

参考文献

[1]増井俊之 EpisoPass:エピソード記憶にもとづくパスワード管理 WISS2013
[2] EpisoPass http://episopass.com
[3]なんちゃって個人情報 http://kazina.com/dummy/index.html
[4]疑似個人情報データ生成サービスhttp://hogehoge.tk/personal/
[5]http://mkdir.g.hatena.ne.jp/ymrl/20110417/1303013060
[6]らいばるサーチ やーやら http://www.kyoiku-press.com/modules/smartsection/item.php?itemid=10767 (2017年1月閲覧)

[7]菅原遼介,高木正則 記述式問題の誤回答を用いた誤答選択死自動生成システムの開発 情報教育シンポジウム2013論文集 2013

(((((大島 裕明, 小山 聡, 田中 克己. Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見. 情報処理学会論文誌. データベース, 47(19):98–112, December 2006.))))