D-028

品詞情報を用いた個人情報フィルタリング手法の提案
Removing the Personal Information by Morphological Analysis

津田 侑 † 高田 秀志 †
Yu Tsuda Hideyuki Takada

1. はじめに

個人用携帯端末の普及により, 電子化された大容量の情報を個人で管理するようになった. ユーザが管理している情報は, ユーザが情報を送信するような能動的な操作をしない限り, 端末内の情報が第三者に流通することはなく, 情報が端末内に閉じ込められた状態となっている. 個人端末内の情報は, 第三者にとっても有益である可能性があり, これらの情報を流通させることで, 第三者は有益な情報を得ることができるかもしれない. しかし, 個人用端末内の情報を第三者に流通させるさい, 個人情報をフィルタリングしなければ, 個人情報が第三者に流通することになる.

本研究では個人端末内の予定表に焦点を当て, その予定表内の個人情報をフィルタリングし, 他者にとって有益な情報のみを抽出をする手法を提案する.

2. 個人用端末内の情報を共有するさいの個人情報の扱い

2.1 「街角メモリ」

我々は, 個人用端末を用いた情報共有を実現するための「街角メモリ」の構築を行っている [1]. 街角メモリは, 街中にあふれる機器, たとえば, 個人用携帯端末や IC カード, 駅の改札, キオスク端末などを「街中に存在するメモリ」ととらえ, その間で情報を流通させることを目的としている.

本研究では, 街角メモリの通信形態として, インターネットを用いたサーバ・クライアント型の情報の取得を目指している.

2.2 情報共有時における個人情報

個人端末内には第三者と共有してもよい情報もあるが, 多くの個人情報が含まれている. たとえば, ある日の予定として, 「花子と琵琶湖の花火大会」という情報が予定表に書き込まれていたとする. この場合, 「花子と」という語句から個人を特定することができる可能性がある. 一方で, 「琵琶湖の花火大会」という語句は, 第三者にとっても有益な情報となる可能性がある. 予定表内の情報で, 第三者にとって有益な情報となる可能性があるものを共有するには, 個人情報をフィルタリングする必要があると考える. 個人情報をフィルタリングしなければ, 第三者に個人情報が漏洩し, 迷惑メールや架空請求などの二次被害が発生する可能性がある.

特定の個人を識別できる情報として, 「住所」, 「人名」, 「生年月日」, 「電話番号」, 「メールアドレス」などの多くの情報が上げられるが, 本研究では予定表に登録される可能性が高い「人名」を対象とする.

†立命館大学 情報理工学部

3. 個人情報を除去した個人端末内情報の共有

3.1 サーバを用いた情報の共有

本研究では, 個人用端末内の予定表を第三者と共有することを目指している. 本システムの利用モデルを図 1 に示す.

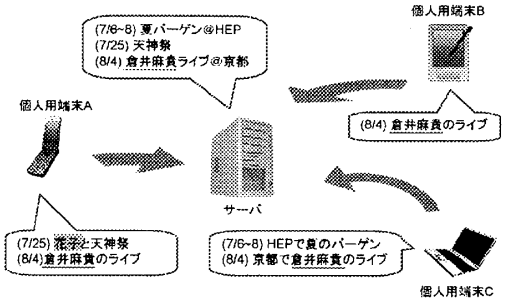

図 1: 予定表共有システムのモデル

システムはサーバ上に実装され, HTTP 通信によって個人用端末から予定表の情報がサーバに集積される. 集積された情報は, サーバ内で個人情報が除去され, サーバ内のデータベースに格納される. 本システムのユーザは, サーバにアクセスすることでデータベースに格納された情報を取得することができる.

3.2 品詞情報を用いた個人情報の除去

本研究では, 予定表内の文字列を形態素解析し, 語句に分解する. 形態素解析を行うことで, 文字列を単語に分解するだけでなく, 品詞情報も得ることができる. この品詞情報を用いて個人情報のフィルタリングを行う.

本研究では, 形態素解析エンジンとして MeCab[2] を用い, これによる形態素解析で得た品詞情報を用いて, 以下の手順で情報の抽出を行う.

1. 形態素解析の結果より, 「人名」と判別された情報を除去する
2. 助詞は単独では意味をなさないので, 除去する
3. 同じ品詞の単語が続くと, 繋げて 1 つの語句にする

これにより残った単語を個人情報を含まない情報とする.

たとえば, 先にあげた「花子と琵琶湖の花火大会」という文字列が予定表に登録されていると仮定する. この文字列を MeCab を用いて形態素解析を行うと, 図 2 のような結果が得られる.

```
花子　名詞,固有名詞,人名,名,*,*,花子,ハナコ,ハナコ
と　　助詞,並立助詞,*,*,*,*,と,ト,ト
琵琶湖名詞,固有名詞,地域,一般,*,*,琵琶湖,ビワコ,ビワコ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
花火　名詞,一般,*,*,*,*,花火,ハナビ,ハナビ
大会　名詞,一般,*,*,*,*,大会,タイカイ,タイカイ
```

図 2:「花子と琵琶湖の花火大会」を形態素解析した結果

図 2 の例から上記の手順で情報の抽出を行うと,「花子と琵琶湖の花火大会」という情報から,「琵琶湖」,「花火大会」という語句だけ抽出される.「琵琶湖」という語句の品詞情報には,「地域」という情報が含まれているので, 予定表に登録された予定は「琵琶湖」で行われることがわかる. また,「花火」,「大会」はともに「名詞」で連続しているので, 1 つに繋げて「花火大会」という 1 つの語句にする. このように情報を加工することによって,『「琵琶湖」という場所で「花火大会」というイベントがある』という情報が生成される. これは第三者にとっても有益となる情報であると考えられる.

3.3 語句の出現頻度による公的情報の抽出

形態素解析をして人名のみを除去するという単純な手法のみでは, 有益な情報も除去してしまう可能性がある. たとえば, ある日の予定表に「京都で倉井麻貴のライブ」という予定を登録しているとする. ここで,「倉井麻貴」という歌手が存在すると仮定する. この文字列に対して形態素解析を行うと, 図 3 となる.

```
京都　名詞,固有名詞,地域,一般,*,*,京都,キョウト,キョート
で　　助詞,格助詞,一般,*,*,*,で,デ,デ
倉井　名詞,固有名詞,人名,姓,*,*,倉井,クライ,クライ
麻貴　名詞,固有名詞,人名,名,*,*,麻貴,マキ,マキ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
ライブ 名詞,一般,*,*,*,*,ライブ,ライブ,ライブ
```

図 3:「京都で倉井麻貴のライブ」を形態素解析した結果

形態素解析の結果より「人名」,「助詞」を除去する方法では,「京都」「ライブ」という名詞のみしか残らず,「誰の」ライブなのかがわからない. これでは有益な情報とは言うことができない.

ここで, 歌手によるライブは公的なイベントであるので, 複数のユーザが「京都で倉井麻貴のライブ」と類似した情報を予定表に登録している可能性がある. 語句の出現頻度によってランキングを作成し, 上位にランクインした人名は公人のものとし, 除去しないようにする.

3.4 検索ヒット数や PageRank を用いた公的情報の判別

別の例として,「倉井麻貴さんと食事会」という予定があるとする. この「倉井麻貴」は歌手ではなく, 一般人であると仮定する. 語句の出現頻度だけで判別すると, 歌手の「倉井麻貴」と同姓同名なので, ランク上位の人名となってしまう. これでは, 人名が除去されずに, サーバ上のデータベースに登録されることになる.

これを解決するために, Google の検索ヒット数を用いる. まず, 形態素解析を行い, 語句を生成する. そして, その語句をキーワードとし, Google に検索クエリを送る. ライブなどの公的なイベントなどは検索ヒット数が多いが,「倉井麻貴」と「食事会」を含む検索キーワードは公的な情報ではなく, ヒット数が少ないと考えられる. ヒット数が少ないものについては, 公的な情報でないと考え, 人名を除去する.

また, ヒット数が多いものは, 公的な情報と判別し, 検索結果の最上位の URL をユーザに提示するようにする. Google の検索結果は, PageRank[3] に基づき順位が決定されるので, 最上位の Web サイトはもっとも予定に関連した情報を持つ Web サイトであると考えられる. ユーザにこの URL を提示することによって, ユーザの情報に対する信頼度が向上すると考えられる.

4. 既存のソフトウェアとの比較

本研究と類似した機能を持つソフトウェアとして, 野村総合研究所の TRUE TELLER 個人情報フィルタ [4] がある. このソフトウェアでは, 人名, 電話番号などの個人情報を自動的に抽出し, マスク処理を行う. また, ユーザが手動で辞書を作成することで, 利用状況に合わせて個人情報として扱うかどうかを定義することができる. 本研究では, Google での検索ヒット数や PageRank を用いることによって, 個人情報か公的な情報かの判別を自動的に行うことができる. また, 本研究では, 個人情報をフィルタリングした情報を他のユーザと共有する機能を持つ.

5. おわりに

本稿では情報共有時における個人情報のフィルタリング手法を提案した. 形態素解析で得た品詞情報によって個人情報をフィルタリングするだけでなく, 語句の出現頻度と Google での検索ヒット数や PageRank を用いることで, 公的な情報の抽出も可能である. 今後は, 本システムを実装し, 評価を行う予定である.

謝辞

本研究を進めるにあたり, 有益なご助言を頂きました立命館大学情報理工学部島川教授および研究室の方々に感謝いたします.

参考文献

[1] 高田 秀志, 伊東 寛修, 大西 雅宏, 玉井 祐輔, 津田 侑, 野口 尚吾, "「街角メモリ」: 個人情報端末間の能動的情報交換による日常的コミュニケーション支援," インタラクション 2007 ポスター発表, 2007 年 3 月.

[2] http://mecab.sourceforge.net/

[3] Page Lawrence, Brin Sergey, Motwani Rajeev, Winograd Terry, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies, Working Paper SIDL-WP-1999-0120, 1998.

[4] http://www.trueteller.net/filter/index.shtml