

形態素解析を用いた個人情報マスキングシステムの研究開発

5 T - 0 1

田中成典* 古田均* 中山浩太郎**

*関西大学総合情報学部 **関西大学大学院

1. はじめに

近年, e-Japan 戦略を始めとする政策で情報化が進む中, 自治体における情報公開へのニーズが急速に高まりつつある. 各都道府県・政令指定都市・中核都市などの地方公共団体では, 情報公開条例が制定・施行され, 更に「行政機関の保有する情報の公開に関する法律(情報公開法)」が施行されるなど, 情報公開の重要性は高まる一方である[1]. しかし, 開示要求される情報の中には, 個人・法人の権利・利益および公共の利益に損害を与える情報が含まれており, 情報公開の際には, これらの情報を保護しつつ公開を進めなければならない.

一方, 情報公開の現場における個人情報の漏洩を防ぐ手段は, 担当者がマジックで塗りつぶす方法しか存在しない. この作業は, 人海戦術で行われており, 公開要求される膨大な数の文書の中から個人情報のみを塗りつぶすためには, 莫大な時間・労力・コストがかかる.

このような状況の下, 本研究では, 自然言語処理技術により, 膨大な数の文書中から, 個人情報を高精度で抽出し, マスキング(隠蔽)処理を自動的に施すシステムの開発を目指す. 本研究の実現によって, 情報の公開が容易に, 迅速に, かつ低コストで行なわれることが期待される.

2. システムの概要

本システムは, ①ファイル解析モジュール, ②マスキングエンジンモジュール, ③ユーザインタフェースモジュールの3つのモジュールで構成(図1)する. 本システムは, まず電子文書を入力し, 次に自治体の検査員がシステムを操作し, 内部で形態素解析を行う. 形態素解析を行う際には, 認識精度を向上させるための処置を施し, 最後にマスキング文書を出力するという順序で処理を行う. 形態素解析エンジンには茶笥[2]を採用

した.

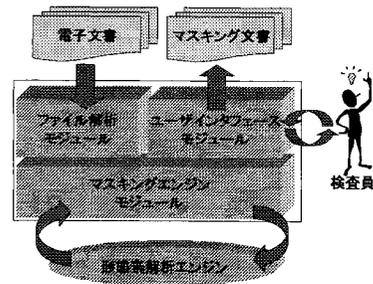


図1 システム構成

2.1 ファイル解析モジュール

ファイル解析モジュールでは, 書式情報などの不要な情報を除去し, テキストのみを抽出し, マスキングエンジンモジュールへ渡す. また, ファイル解析モジュールでは, 電子政府の実現を見越し, Web 上での電子文書交換基盤となるファイル形式へ対応する. 具体的には, PDF, HTML, XML, リッチテキスト, そして Word の5ファイル形式へのマスキングを可能にした. PDF へのアクセスは「AcrobatSDK」, HTML へのアクセスは Microsoft 社の HTML パーサである「MSHTML テクノロジー」, XML へのアクセスは Microsoft 社の DOM (Document Object Model) である「MSXML」, リッチテキストへのアクセスは, ActiveX コントロールである「リッチテキストコントロール」, Word へのアクセスは「OLE」をそれぞれ利用した.

2.2 マスキングエンジンモジュール

マスキングエンジンモジュールでは, ファイル解析モジュールから渡されたテキスト情報を形態素解析を施し, 個人情報(人名)を検出し, ユーザインタフェースモジュールへ渡す.

現在の形態素解析の精度は 100%でなく, 誤認識を含むことは周知の事実である. 既存の形態素解析技術の精度では, 本システムに要求される精度に達していないため, 本研究では, 形態素解析エンジンの精度向上に最も注力し, システムの開発を行った. 精度向上のために施した処

置は、①希少人名辞書の追加、②茶笥のカスタマイズ、③学習機能の追加の3点である。

①希少人名辞書の追加とは、「海老沢」、「続」、「元重」など、珍しい名前を辞書に追加する処理である。②茶笥のカスタマイズとは、茶笥のLisp ファイルを修正し、形態素解析の品詞コストと連結品詞を設定[3]することである。③学習機能の追加とは、認識漏れした人名を辞書に追加することにより、認識漏れを防ぐ機能である。

2.3 ユーザインタフェースモジュール

ユーザインタフェースモジュールでは、膨大な文書を一括でマスキング処理し、人間の労力を最小化することや、検査員がチェックしやすいようなユーザインタフェースを提供する。これは、現在の人海戦術による個人情報保護方式を鑑みると、膨大なコストが浪費されているからである。このような状況を打破するために、電気代が安価で、従業員がいない夜中の間に膨大な量の文書に対してマスキングを行う一括処理機能等、作業効率を向上させる機能を付加する必要がある。

これらの機能は、情報公開の現場の声を取り入れる必要があったため、大阪府府民情報課および高槻市市民情報室からヒアリングを行った後、設計を行った。

2.4 マスキング処理

マスキング処理では、抽出された個人情報を電子的に塗りつぶす機能を提供する。マスキング処理が施されたファイルを保存する機能を実装することにより、インターネット上や紙面に印刷して公開することが可能になる。マスキング処理を施された結果画面を図2に示す。

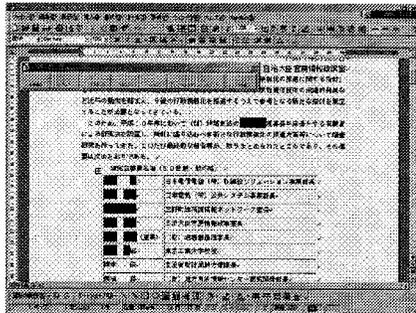


図2 マスキング処理

3. 評価実験

本研究では、マスキングモジュールで開発した3つの処理、①希少人名辞書の追加、②茶笥のカスタマイズ、③学習機能の追加に関して精度を検証した。

3.1 システムの検証方法

システムの評価には、人名認識漏れ率および人名過剰認識率の2つの指標を採用した。人名認識漏れとは、文章中の人名を人名以外の品詞として認識することである。人名認識漏れ率は、文書中に認識漏れの発生する割合である。

人名過剰認識とは、人名以外の品詞を人名として認識することである。過剰認識率は、文書中に過剰認識の発生する割合である。

3.2 システム評価

システムの実験結果を表1に示す。ヒアリングした自治体からは、9割を超える精度であれば、十分実用化に耐え得るとのコメントを頂いた。最終的には、認識漏れ率 3.6%、過剰認識率 2.5%を記録し、目標を大きく上回る結果となった。

表1 検証結果

	認識漏れ率	過剰認識率
希少人名辞書の追加	9.0%	45.7%
茶笥のカスタマイズ	6.1%	2.5%
学習機能の追加	3.6%	2.5%

4. おわりに

本研究では、形態素解析を用いて情報公開法に向け個人情報の抽出と塗りつぶしの自動化を実現した。それに伴い、情報の公開が容易に、迅速に、かつ低コストで行うことが可能となった。今後、更なる情報抽出精度向上のため、構文解析、意味解析などの高度な自然言語処理技術の適用を考慮していかなければならない。

参考文献

- [1] 松井茂記: 情報公開法入門, 岩波新書, 2000.11.
- [2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸: 日本語形態素解析システム「茶笥」Version 2.0 使用説明書 第二版, NAIST Technical Report, 1999.12.
- [3] 長田靖, 吉田敬一: 後続の品詞を考慮した形態素解析, 情報処理学会第60回全国大会, 2-89, 2000.3.