2006 年度卒業研究概要

参照元同一文書の検知手法に関する提案

大谷紀子研究室

0232176　西川泰史

1. 研究背景と目的

パソコンの急速な普及により文書の電子化が進み、文書の作成、編集は紙媒体で行うよりも簡便なものとなっている。しかし同時に他人が作成した文書を改変、あるいは複製して使用する人が増えている。他人の著作物を自分の著作であると偽って使用することは、著作物への諸権利を侵害する行為である。しかしながら、現状では人がすべての文書を確認する以外に複製や改変を発見することはできず、膨大な文書を人が確認した場合非常に多くの時間を費やすことになる。

本研究では人間が文書を確認する際の手間と時間の軽減を目的として、機械による検知を行う際の判断基準と手法を提案する。人間が確認する前にふるい分けを行い、人間が改めて確認すべき文書を提示するシステムを構築し、有効性を検証する。

2. 文書比較方法

本研究では、文書全体を 1 つのベクトルとして比較する方法、類似した文を含めた前後 3 つのベクトルを比較する方法、ユーザの指定する単語を含んだ文と前後 3 つのベクトルを比較する方法の 3 つの手法を提案する。

提案手法では、ある文書における出現頻度が高く、特定の文書に集中的に現れている単語が文書の特徴を示しているという考えに基づく TF-IDF 法[1]を利用する。形態素解析システム茶筌[2]を利用して抽出した名詞を TF-IDF 法で用いる単語とした。また文書の比較にはベクトル空間モデル[3]を用いる。

2－1. 文書全体を利用した類似性検知

すべての文書を形態素解析し、解析結果から名詞をすべて抽出したものを索引語集とする。索引語集を利用し、TF-IDF 法に基づいて各文書の有効語を抽出したものを語彙集合とする。語彙集合を利用して各文書をベクトル化し、類似度の高いものを出力する。

2－2. 文を利用した参照元同一可能性検知（1）

前項と同様の手法を用いて句点を区切りとした文単位の類似性比較を行う。類似性の高い文同士はそれぞれの文面上における前後 1 文同士の類似度も算出し、3 つの文ベクトル同士の平均類似度の中で高いものを出力する。

2－3. 文を利用した参照元同一可能性検知（2）

ユーザが入力した文字列を含んだ文を比較対象となる文の集合から検索し、２－２節の比較結果のうち、検索された文を使用した結果のみを抽出する。２－２節と同様に文面上における前後1文同士の、平均類似度の中で高いものを出力する。

3. 実験

ある授業で提出されたレポートから無作為に抜き出した 171 部に対して比較実験を行った。レポートは提出者の個人を示す情報以外のフォーマットの指定はなく、明確な模範解答が存在しない問題である。

3－1. 実験結果

類似性検知では、類似性の高いと判断された上位 100 個の類似度を出力した結果、最高が 0.904、最低が 0.720 となった。内容を確認したところ一見して複製した、あるいは改変したとわかるものが複数存在した。しかし類似性が 0.8 を超えるものであっても内容からでは複製したとは判断できないものもあり、適合率は 21%、再現率は 37%となった。文書群全体の類似性の分布は図 1 に示す。

参照元同一可能性検知（1）でも同様に上位 100 個の類似度を出力した。結果は最高が 1、最低が 0.272 であった。結果が 1 だったものは完全に同一の文書を含んでいたが、類似度が高い場合でも内容が同じとは判断できないものもあり、適合率は 13%、再現率は 14%となった。文全体の類似性の分布を図 2 に示す。

参照元同一可能性検知（2）ではレポート内に比較的多く使われた複数の単語を指定してそれぞれの結果を出力した。結果は使用した単語ごとに異なるが、最高で 1、最低で 0.218 であった。高い類似度を示したもののいくつかは問題文をレポート内に書き込んだ文章だったが、内容的に類似した文章の場合もあった。

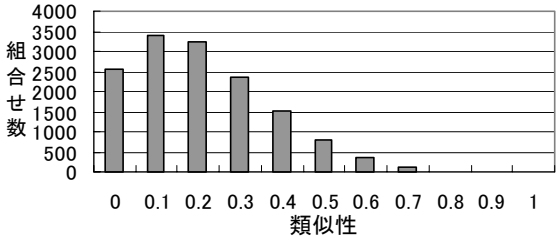

図 1：類似性分布

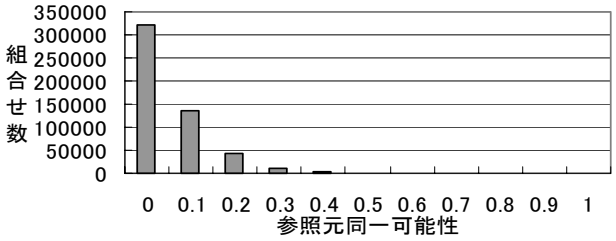

図 2：参照元同一可能性分布

4. まとめ

実験によって、提案手法は比較対象の類似性、および参照元同一可能性をある程度検知できるといえる。複数の手法を併用することで参照したと思われる部分の特定や、回答の傾向などを把握することも可能である。しかしながら類似度の高いものであっても必ずしも複製や改変されたものであるとはいえず、最終的には人間による確認作業が必要なことには変わりない。また今回の実験で使用したレポートは明確なフォーマットの統一がされておらず、単純な比較では内容の異なる文書の正確な類似度を示すことができなかった。今後の課題としては、文章の長さや文書のフォーマットに左右されない、より正確な類似度を計測する手法の研究と、単語抽出の際に多く使われた単語の自動抽出などの機能の追加が考えられる。

参考文献

[1]岸田和明，“情報検索の統計モデル”，人文学と情報処理，No.28，pp.6-15，2000.

[2]ChaSen’s Wiki，http://chasen.naist.jp/hiki/ChaSen/

[3]大谷紀子，“情報検索におけるベクトル空間モデルの応用”，武蔵工業大学環境情報学部紀要，第五号，pp.99-109，2004.