社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (17)
2008－ＮＬ－183 (17)
2008／1／22

辞書に依存しない文章間類似度の比較評価手法

村上智哉[†1] 中谷直司[†1]
厚井裕司[†1] 後沢忍[†2]

高速なネットワークで世界中が相互に接続される現代社会は便利である反面，それを用いた犯罪や情報の流出などが常態化し，社会問題となっている．そのため，イントラネットを監視するための一技術であるネットワーク・フォレンジックが発達してきた．しかし，情報流出に繋がった可能性のあるメールなどを探索する際に用いられるのは，特定のキーワードがその本文に含まれているか否かによって判定する，というような極めて単純な手法が多い．しかし，このような素朴なアプローチでは重要な情報を取りこぼすか，あるいは誤検出するであろうことが容易に想像できる．そこで，ネットワークを通過したテキストデータと，既存の重要文書の関連性を推測する必要があるが，そのためには何をもって二つの文章が類似しているとみなすか，という一定の評価基準が必要となる．本稿では，文法や辞書などの事前知識に極力依存せず，文章の表層的な特徴に基づいて特徴点となる単語を抽出し，それを用いて文章間における類似箇所を発見する手法について提案する．

The dictionary independent evaluation technique between the two different types of documents

Tomoya Murakami,[†1] Naoshi Nakaya,[†1] Yuuji Koui[†1]
and Shinobu Ushirozawa[†2]

The modern society where the world is mutually connected by the high-speed network makes our lives more convenient. On the other hand, network crimes and information outflows occur frequently and those have become social problems. Therefore, the network forensic technique has been progressing day by day in order to supervise intranet. However, in case of electronic communication (e.g., email) the existing technique is normally used to specify the confidential data contained in the text or not. But this existing technique isn't sufficient to take the countermeasure against the information outflow. By using this existing method, a possibility of overlooking the important information or incorrect-detection of those confidential data is so high. So, it is necessary to compare and evaluate the similarities between the relevance of existing important document inside the server computer and the text data which passed through the network. In this approach, we would like to propose the new technique of extracting the feature point from a text data and existing document based on the surface-feature and compare between these data. Finally, this technique detected the similarities between the texts which are independent on prior knowledge such as grammar and the dictionary etc..

1. まえがき

今日のコンピュータ・ネットワーク社会において，多量のデータを蓄積し，適切に保存・運用することで得られる効果や利益は計りしれないものとなっている．反面，何万何十万といった件数の個人情報や機密書類のような重要なデータも，ネットワークを通じて送信する際は一瞬であり，また小型の記憶媒体に収まるようになってしまったために，それが納められた媒体を紛失したり，ネットワークを通じて流出してしまうような事件が増えている．

起こされる訴訟に対して適切に対応するため，あるいは自社ネットワークを監視するために，自社ネットワークの内部にフォレンジック・ソリューションを導入する企業や団体が増えてきた．フォレンジックとは，「法的な」という意味で，本来は法医学などの用語である．主にコンピュータに対する調査のことをコンピュータ・フォレンジックと呼ぶ．コンピュータ・フォレンジックは主にホスト型フォレンジックとネットワーク型フォレンジックに大別できる．前者は攻撃者に侵入されたと思われるホストに記録されている情報を破壊せずに，法的根拠のある形で取り出し保存，解析するものである．後者はイントラネットのゲート

†1 岩手大学工学部
Faculty of Engineering, Iwate University
†2 三菱電機株式会社
Mitsubishi Electric Corporation

ウェイ直前にパケットを取得するソリューションを設置し，送受信されたパケットを記録することで証拠とし，後からどのような通信が行われたかを確認するためのものである．あらゆる種類のパケットを保存するタイプのものを特にフルフォレンジック型，HTTP や SMTP/POP3 といった特定のプロトコルだけに対応するものを部分フォレンジック型と呼称する．[1)]

現在主流となっているネットワーク・フォレンジック・ソリューションにおいて，記録したパケットから復元したファイルなどに含まれるプレーンテキストに“禁止語句”が含まれている場合に，その通信が危険なものであるとみなし通信停止，あるいは管理者に警告する，といったフィルタリング機能をもつものがある．[2)3)] しかし，このような手法では“禁止語句リスト”のようなものを事前に用意する必要があり，大抵の場合それは膨大な量となるだろう．また誤検出の可能性もある．極端な例を上げるならば，“重要”という単語が含まれている文章が必ずしも重要とは限らない．逆に，禁止語句に設定された言葉を一切含まないような文章が社外秘に相当する可能性もあるだろう．このように，禁止語句を用いた方法はポピュラーではあるが，確実性には欠ける可能性がある．そこで，流出を阻止したい重要文書から事前に何らかの特徴点を抽出しておき，それがネットワークを流れるテキストの特徴点とどの程度類似しているかを算定し，それがある一定の閾値を超えた場合にその文章を流出させてはならない“重要文書”とみなすという，より柔軟なフィルタリングへ応用可能な文章間類似度の算定法を提案する．

2. 文章間類似度の算定法

2.1 手法の位置づけ

まず前提として，蓄積したパケットから文書ファイルを，ひいてはプレーンテキストを抽出できる状況であることを確認する．元の文章に暗号化や圧縮などがかけられたままでは，提案する手法は適用できない．また，文面を詳しく吟味する方式は前述の禁止語句方式と比較して複雑であるため，リアルタイムな処理は難しい可能性がある．そのため，文面の危険度算定を二つの段階に分割する．

(1) 従来どおりの“禁止語句”方式で判定を行い，それを通過したパケットは外部へ送信する．
(2) パケットを蓄積し送信しようとしている文章を復元，機密文章との“類似度判定”を改めて行う．

このような方式をとることで，より柔軟かつ強力なフィルタリングを実現したいと考えている．

2.2 類似の定義

問題となるのは実際にどこまでを“類似する文章”とすべきか，何をもって“類似”していると判断するか，という点である．[4)] 例えば，“同じ事実を同じ論調で別の新聞記者が記事にしたもの”までも同一のものだとしてしまうと，偶然の一致による誤検出が多くなることは想像に難くないし，かといって一字一句が完全一致する場合とすれば，句読点や改行の有無などの微細な変化で検出は不可能になってしまう．そこで，語尾の変化したものや，句読点の挿入及び削除，センテンスの順序の入れ替えを行うなどの改変を加えたものは，オリジナルと同一とみなすことにする．たとえば，

- 文の順番を入れ替える
- 語尾の言い換えをする
- 略語は本来の言葉に直す
- 長い熟語はひらがな交じりに崩す
- 記号や句読点の追加・削除

これらの改変を加える前後の文書は同等のものとみなして考える．このように定義した“同一”に近い文章間において相対的に高い類似度のスコアが発生するようなアルゴリズムを考案すればよい．様々な文章の観察を行いわかったことは，文章の言い換えを行う際に，まず書き換えられるのは接続詞や助詞，記号であり，文章から何らかの特徴点を抽出する際に重きを置かないようにする配慮が必要だろうということだ．言い換えれば，日本語の文章は，漢字[5)] やカタカナで構成された固有名詞などをひらがな列で“糊付け”したものだというように表層的にとらえることができる．無論，ひらがなだけで構成された名詞も多々あるが，それを判別するためには形態素解析のような手法と辞書が必要となるため，ここではあえて深く論じない．

2.3 提案手法

はじめに，文章からその特徴点となる“単語”を抽出する．英数字，あるいはカタカナの連続文字列はそのまま抜き出し，最短単語長よりも短いものは取り除く．漢字の連続文字列は n-gram 方式で分割し，そのそれぞれを単語とみなす．なお n-gram 方式とは，対象の文字列を n 文字ごとに分割する手法のことで，主にデータベースの全文検索などの分野で用いられている．例えば，“あいうえお”という文字列を 2-gram で分割すると，“あい”，“いう”，“うえ”，“えお”というように 4 つの文字列となる．同様に 3-gram の場合は，“あいう”，“いうえ”，“うえお”のようになる．

与えられた文章から“単語”を抽出する手順について，例文を用いて説明する（図 1 (a)）．まずは，文中からひらがなや記号を取り除き，その箇所を仮に単語の区切りとする（図 1 (b)）．その単語の中から，最短単語長より短い単語を取り除く（図 1 (c)）．ここでは仮に，長さが 1 以下の単語，すなわち“物”が取り除かれる．残った単語のうち，漢字列は n-gram で分割し，(なおここでは，n=2 とする．) そのそれぞれを単語とする．(図 1 (d))．

このような手順によって単語の抽出が可能になった．問題となるのは，この単語群をどのように用い，2 つ

語尾の変化した物や, 句読点の挿入及び削除, センテンスの順序

(a) 元の文章

語尾, 変化, 物, 句読点, 挿入及, 削除, センテンス, 順序

(b) ひらがなや記号を取り除く

語尾, 変化, 句読点, 挿入及, 削除, センテンス, 順序

(c) 最短単語長未満の要素を取り除く

語尾, 変化, 句読, 読点, 挿入, 入及, 削除, センテンス, 順序

(d) 単語の抽出完了

図 1 単語の抽出手順

の文章間における類似点を発見するかということである.

前述の "同一の文章" の定義より, 2つの文章中に共通して出現する単語は, 文章間における類似性を表していると考えられる. そこで, 比較元となる文章と比較対象となる文章からそれぞれ単語を抽出し, 共通して出現する単語の割合を求めればよいと思われるが, 単純に文章全体で一対一の対応をとるだけでは, 文章中の局所的な一致を薄めることになってしまうことは想像に難くない. また, フォレンジック・ソリューションへの搭載を前提としている以上, ネットワークを通過するパケットから元のプレーンテキストの完全な復元を待つのは時間が掛かりすぎる. さらに言えば, 蓄積した数パケットから抽出した短文が使用できるような手法であれば理想的である. つまり, 比較対象となる文章が不完全かつサイズの小さなものでも適用可能な手法である必要がある. そこで, 単語リストを一定の単語数ごとに "区切り", そのそれぞれについて注目することでそれらの問題に対応する.

比較対象となる文章から抽出した単語リストにおける, ある一区切りに含まれる単語の全てについて, それが比較元から抽出した単語リストに含まれるか否を調べ, その一区切りにおける "単語の一致率" を求める. 同様に, 全ての一区切りについて同じ操作を行う. それぞれの一区切りについての "単語の一致率" は 0 %から 100 %の範囲で表され, それが高く算定される場合, その一区切りには他方のリストとの類似性があるとみなす. この手法によって, どのような結果が得られるか実際に実験して試す.

3. 実　　験

3.1 日本語文章の場合

全く関係のない二つの市が発行した広報に, ほぼうりふたつと言ってよい市町村合併についての説明文が掲載された事例[6] を発見したため, それらを対象とすることにした. 都市名などの固有名詞や言い回し, 記号などの細部が変更されてはいるものの, 全体の構成はほぼ同一であるように思われる. ここから, 類似箇所を検出することができるか実験する. 比較対象として与える一節は図 2 のようになる. この一節に類似している箇所が類似元の文章に含まれているかを調べた結果を, 図 3 に図示する. なお, 漢字の n-gram 幅 = 2, 最短単語長 = 2 であり, また折れ線グラフの縦軸は一致率, 横軸は区切り数を表す. グラフからも分かる通り, ある一部分だけ一致率が高く算定されている. そこで, 一致率が 40 %を超える部分を比較元の文章から切り出すと, 図 4 のようになる. 予想通り, ある程度までの分割幅の短縮はグラフを先鋭化させる効果が生まれ, 閾値を選択する上でよい影響が出るが, それを過ぎるとノイズが急激に増加する.

このように, 区切り幅による前後の "ごみ" などを除けば, おおよそ納得できる結果であるといえるだろう. 上の実験で浮かび上がってきた問題は, 最適な単語の区切り幅をいかに決定するか, というものである. 上の例では暫定的に, 比較対象となった短文に含まれる単語数である 36 としたが, 視覚的にもグラフはなだらかであり, 閾値を設定するのは難しい. より鮮明なグラフを得るためには単語の分割幅を狭めることが不可欠であろうが, それが過ぎれば偶然の一致による

広域行政にも限界がある
○ 広域行政には, 一部事務組合や広域連合, 協議会などがあり, 守口市では, 効率的に事務を行うため, 古くからこれらを積極的に活用してきました.
○ しかし, 理屈の上では, 広域行政で行う方が合理的と考えられる分野でも, それぞれの市の事情が異なるため調整が難しいなど, 現実にはなかなか進まないのが実情です.

図 2 比較対象となる文章

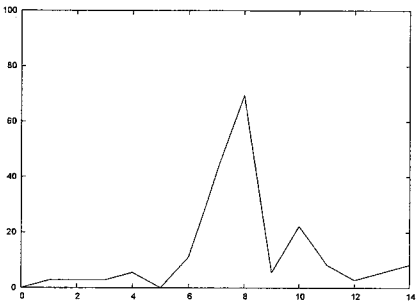


図 3 二種類の公報における類似箇所

財政運営の健全化など行財政改善に積極的に取り組んでいます．しかし，1市単独での努力では限界があります．さらに，国・地方を問わない厳しい財政状況を考えれば，より抜本的な方策を講じる必要性がますます強くなってきています
■ 広域行政にも限界がある
・広域行政には，一部事務組合や広域連合，協議会があり，門真市では，効率的に事務を行うため，古くからこれらの広域行政を積極的に活用してきました
・広域行政で行う方が理屈の上では，合理的と考えられる分野でも，それぞれの市の事情が異なるため調整が難しく，1市でも反対すれば実現しないなど，現実にはなかなか広がらないのが実情です
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝
門真市広報　　平成14年2月1日付3

図 4　実際に検出された箇所

ノイズが増加し，かえって一致箇所の抽出が難しくなる可能性がある．そこで，分割幅を元の半分の18から，ほぼ最短の3まで変化させ，結果にどのような影響が現れるか試した結果が図5である．

3.2 英語文章の場合

同様に，英文についても実験を行った．英文は単語間が空白で区切られているので，前述のように単語の抽出には苦労しない．しかし，英文は日本語の文と比較して，前置詞や代名詞のように文脈に関係なく出現する単語が多く，また文中の殆どの語句が単語として抽出される．これらが望まない偶然の一致をし，結果に悪影響を及ぼす可能性がある．上の実験と同じ手法を用いた場合，日本語が大部分を占める文章の結果とどのような差異が生まれるか，調べる．実験に用いるのは，ボットウェアネットワークについての論文である.[7] そこから図6のような一文を抜粋し，元の文章との類似性を求める．まず，分割幅を比較対象の文章に含まれる単語数と等しい数で固定し，最短単語長を1から6まで変化させ，その差異を図7に示す．単語長が長くなればなるほど，ノイズは減少するが，それ以上に正当な一致も減少する．3文字以下の単語に重要なものは多くないであろうことから，英文における最短単語長を4とおくことが可能かもしれない．次に，最短単語長を4で固定し，分割幅を変化させた場合の差異を調べる．例として用意した論文に含まれる単語を数え上げ，出現回数の多いものから順に並べると，上位20番は図8のようになる．

"malnet" は，malicious（悪意ある）と network を合わせた，この論文の筆者によって作られた造語であり，意味合いとしては "ボットウェアネットワーク" に近い．その他の単語も，ネットワークについて論ずる際にはありふれている単語が多い．しかし，"that" や "with" といった単語を特徴点とみなすのは苦しい．そこで，抽出された単語を参考に意味の薄い単語のリストを作成し，そこに含まれる単語を考慮しないことで，実験結果にどのような影響があるか実験する．なお，考慮しない単語のリストは図9のように作成した．考慮しない単語のリストを使用する場合としない場合のそれぞれについて，分割幅を変化させた場合の結果の差異を調べたものが図10である．特に，分割幅が広い場合について，ノイズレベルはそのままに，おおよそ一致箇所を先鋭化させることができた．

Despite the lack of directed organization, a random malnet can have good connectivity. For different values of node degree r, we tested twenty 10,000-node mal-nets; for each network, we tested different values of x; and ten different random cuts were measured for each value of x.

図 6　比較対象となる英文

malnets = 75
nodes = 68
that = 58
malnet = 58
malnodes = 41
world = 40
small = 39
node = 36
with = 36
graph = 32

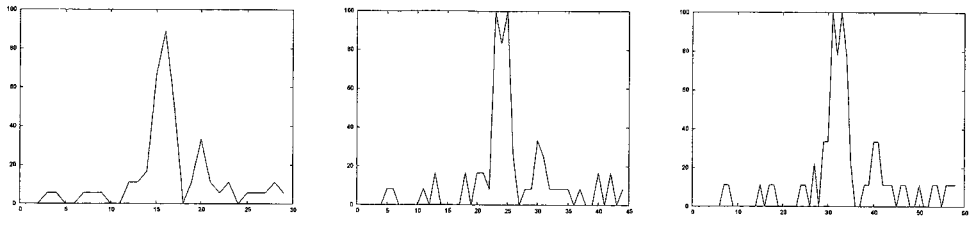
図 8　用意した論文に多く含まれる単語の種類とその数

also, been, can, could, each, even, from, had, has, have, how, like, many, may, more, much, must, only, other, over, some, such, than, that, their, them, then, there, these, they, this, very, what, when, where, which, who, will, with, would

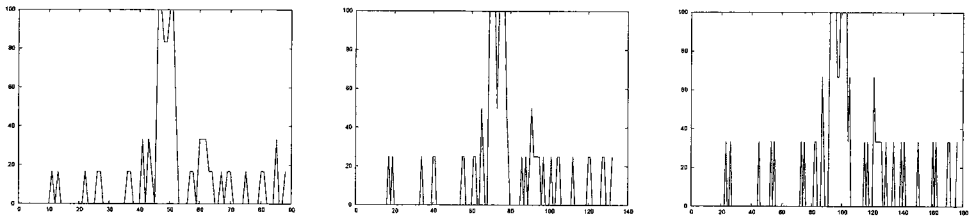
図 9　考慮しない単語のリスト

4. 考　　察

提案した手法を用いる場合，あまりに短すぎる文章からは，望むような結果を求めることは難しい．"単語" が充分に抽出できないほどに短い，あるいは過剰にひらがなが多いような文章においてその傾向は顕著

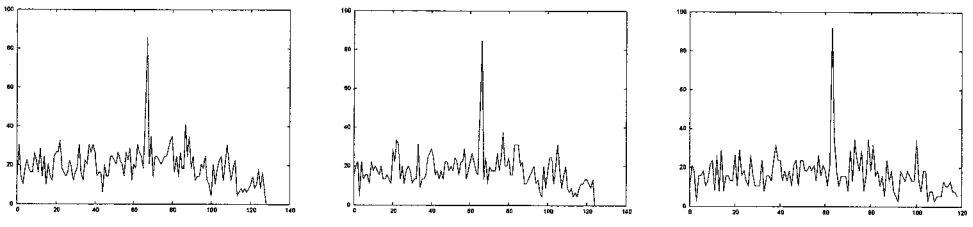


(a) 分割幅 = 18　(b) 分割幅 = 12　(c) 分割幅 = 9

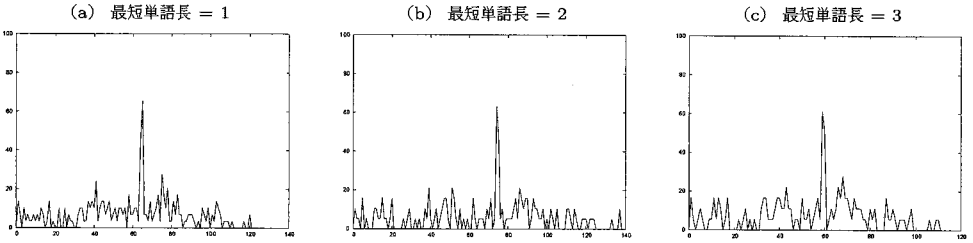


(d) 分割幅 = 6　(e) 分割幅 = 4　(f) 分割幅 = 3

図 5　単語リストの分割幅を変化させたときの結果



(a) 最短単語長 = 1　(b) 最短単語長 = 2　(c) 最短単語長 = 3



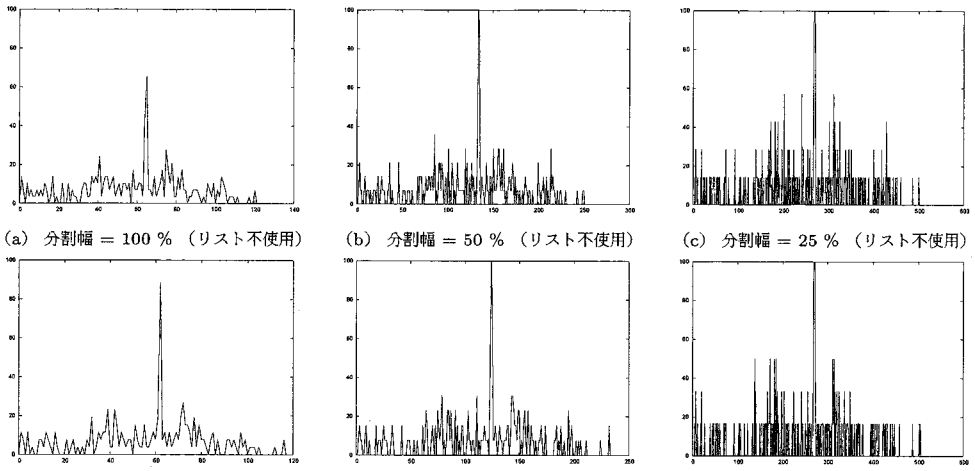
(d) 最短単語長 = 4　(e) 最短単語長 = 5　(f) 最短単語長 = 6

図 7　最短単語長を変化させた結果

である．類似度をある程度正しく算出するためには，分割幅を狭める必要からも，おおよそ 30〜40 単語が必要である．それ以下である場合，偶然の一致によるノイズが多くなりすぎるようだ．また，最適な分割幅を自動的に決定する方法についてはまだ考案できていない．英語の文章については，日本語のそれと比較してノイズが多い傾向にあるようだ．原因の一つはおそらく，前置詞や代名詞などを多用する英語の文法にある．それらを特例として考慮から外すことで，結果はある程度向上することがわかったが，決定的な対策とは言えない．

5. むすび

本稿では，文章の表層的な特徴から単語を抽出し，それらの出現パターンから文章間の類似性を発見する手法を提案した．この手法は，従来の“禁止語句”方式のネットワーク・フォレンジックと組み合わせることで，重要データの取りこぼしや誤検出などの少ない，より柔軟かつ強力なフィルタリングが可能になると考



(a) 分割幅 = 100 % (リスト不使用) (b) 分割幅 = 50 % (リスト不使用) (c) 分割幅 = 25 % (リスト不使用)

(d) 分割幅 = 100 % (リスト使用) (e) 分割幅 = 50 % (リスト使用) (f) 分割幅 = 25 % (リスト使用)

図 10 考慮しない単語のリストを使用する場合としない場合の結果

えている．提案手法は，文章間類似度の算定法も含めて未だ未完成であり，すぐに実用化が可能なわけではない．文章間類似度の算定法においては，場合によっては正当な一致とノイズによる偶然の一致を区別しにくいという問題があり，解決手段を模索しなければならない．単語リストの区切り幅がその最たる例であり，文書の表層的な特徴からそれを自動決定できれば理想的である．現状のネットワーク・フォレンジックの問題点の一つに，記録されたログやパケットから有為な情報を取り出す作業が煩雑になりやすいというものがある．提案手法は単純であるが直感的であり，そういった問題を多少なりとも改善できる可能性があるのではないだろうか．

参 考 文 献

1) 日本ネット技術研究所：Cust.FAI システム，http://www.netpub.tsuzuki.yokohama.jp/fai/index.html.
2) 日本電気株式会社：システムの設定不備や管理不備による文書漏洩を未然に防ぐ機密文書漏洩検査システムを開発，http://www.nec.co.jp/press/ja/0503/0402.html.
3) ネット・エージェント株式会社：ネットワーク監視装置 情報漏洩対策 PACKET BLACKHOLE，http://www.packetblackhole.jp/.
4) 飯箸泰宏：文字列類似度の汎用的尺度，http://www.sciencehouse.jp/etc/kenkyu050313.pdf.
5) 豪田まりぽ：日本語文書の文字・単語出現頻度解析ツールとデータ，http://www.madin.jp/docs/wordcount.html.
6) 戸田ひさよし：文書盗用？門真・守口の広報にうり二つの奇っ怪，http://www.hige-toda.com/_mado01/2002/index_2002_02_02.htm.
7) Jun Li, Toby Ehrenkranz, Geoff Kuenning, Peter Reiher: Simulation and Analysis on the Resiliency and Effciency of Malnets, *Proceedings of the 19th Workshop on Principles of Advanced and Distributed Simulation* (2005).