# テキストマイニングと社会活動のトレース

山田　篤
(財)京都高度技術研究所

## 1. はじめに

データマイニングとは，大量のデータを対象とし，統計的手法を用いて，そこから意味のある情報を抽出することである．技術の発達により，様々なデータが電子的に取得可能となり，日々膨大なデータが蓄積されるようになった．それに伴い，従来は人手や目視で処理可能であったものが，データ量の増大により困難になった．そこで，大量のデータ群の中に埋もれている情報をうまく取り出し，それを活用することが考えられるようになった．データ量が増大することにより，処理は困難となるが，逆に，従来はサンプル数が少なくて，単なる仮説でしかなかったものが，大量のデータを用いて検証できるようにもなった．

本稿で取り上げるテキストマイニングとは，特にテキストデータを対象にしたデータマイニングのことである．新聞，雑誌，書籍といった従来のメディアの電子化はもとより，インターネットの出現により，我々の身の回りには膨大なテキストが存在するようになった．また，自由記述アンケートに対する回答やコールセンターに寄せられた発話の記録なども電子的に保存されている．これらが大量に蓄積されると，人手による処理の限界を超え，何らかの処理システムの手助けが必要となってくる．検索エンジンを用いた，いわゆるキーワード検索を利用した経験を持つ人も多いと思われるが，大量のテキスト群の中から欲しい情報を掘り起こすことは，そう簡単なことではない．

テキストマイニングにおいては，量の問題以外に，情報が自然言語で記述されているという問題がある．自然言語による表現は，その言語を解する人間にとっては理解が容易であるが，計算機が，テキストの内容を人間と同様に「理解」することは可能なのかというのは，よく聞かれる問いである．

例えば，2008年6月に起こった秋葉原における連続殺傷事件において，容疑者がインターネット掲示板に予告の書き込みをしていたことが知られている．膨大なテキスト群の中から，このような書き込みを自動的に即座に検知するためには，自然言語処理をはじめとしたテキストマイニングの技術が必要になる．これについては，総務省の「インターネット上の違法・有害情報への対応に関する検討会」に設置された技術検討WGにおいて，インターネット上の犯行予告や隠語を使った犯罪情報を検知する技術の開発に関する検討を始めていることが報道されている．

以下では，はじめにテキストマイニングの技術的背景について述べた後に，社会活動のトレースという側面から考察を加える．

## 2. テキストマイニングの基盤技術

テキストにおける最小の構成単位は文字である．電子化されたテキストでは文字ごとに文字コードが与えられ，プレインテキストは計算機上では文字コードの列によって表現される．歴史的な背景から，複数の文字コードの体系が存在するため，そのテキストがどの体系を用いているかを知ること，また複数の体系が混在した環境ではそれらの間の相互変換を行うことが必要になる．このとき，異なる体系の間でのマッピングで問題が生じることがある．また，改行コードや空白文字の取り扱いについても一定の取り決めが必要になる．

テキストを文字の集まりと見なした場合，計量可能な統計量としては，各文字の出現頻度(確率)や，文字の連鎖であるn字組(n-gram)の出現頻度(確率)，更には，ある文字の次に生起する文字の確率といったものが考えられる．日本語の場合には，ひらがな，かたかな，漢字といった字種の出現頻度(確率)も計算できる．

部分文字列検索では，転置索引[1]などを用いて検索文

"Text Mining and Tracing of Social Activities" by Atsushi YAMADA (ASTEM RI/Kyoto).

字列の出現位置を計算する．次に述べる形態素という単位を用いずに文字の情報のみを用いた場合，「用語」という検索語に対して「活用語」の後ろ2字がヒットするといった問題もあるが，形態素解析の処理精度が100%ではない状況のもとで，候補を完全に漏れなく抽出したいといった場合に，文字を単位とした処理は有用である．

例えば，「コンピューター」の中に「コンピュータ」は含まれているので，後者で文字列検索をかけることにより，前者も取得することは可能である．しかし類義である「計算機」は全く異なる文字列なので，「コンピュータ」で検索しても出てこない．このように，より深い「意味」を扱おうとすると，文字だけを対象としていたのでは不十分である．

意味の最小の構成単位として，単語が考えられる．英語の場合は，ホワイトスペース[††]によって区切られた単位で単語を認定することができるが，日本語の場合，このような分かち書きがなされないため，何らかの処理を施す必要がある．対象となるテキストを，辞書に登録してある一定の単位に分割することを形態素解析と呼ぶ．形態素解析における主要なタスクは，単語分割(タギング)と品詞付与である．その副作用として，品詞以外の様々な辞書格納情報(例えば，活用語の原形等)も付与することができる．そのアルゴリズムについては解説を省略するが，簡単に言えば，辞書に登録されているエントリの組み合わせで，入力テキストをもっともよく被覆するものを計算する．どうしても辞書中のエントリでは被覆できない箇所は未知語と解析される．

形態素解析では，解析エンジンと辞書を用いる．オープンソースで入手可能な日本語の形態素解析エンジンとして，JUMAN，ChaSen，MeCab等がある．また，ChaSen，MeCabで利用可能な形態素解析用の電子化辞書としてIPADic，NAIST-jdicやUniDicがある．

形態素解析用電子化辞書において，テキストマイニングにおける利用を考えると，単位の斉一性[†††]を保証する必要がある．ある辞書で「幾何学」は1語，「心理｜学」は2語と解析されたとしよう．すると，接尾辞の「学」の用法をマイニングしようとした場合に「心理学」は取り出せても，「幾何学」は全体で1語と解析されているため，取り出すことができない．既存の辞書では，内部処理の都合で比較的長い単位で辞書のエントリが構成されていることが多々あるが，これに対する一つの解決法は，より細かく分割した場合の情報を辞書の内部に持っておき，解析処理の後で，必要に応じてそれを出力する(複合語の処理)というものである．

また，意味にかかわる別の問題として，表記上は異なっていても，同一の語と見なせるものの存在がある．一般には異表記の問題として，検索時に問題となることが多いが，マイニングにおいても同様の問題が生じる．先の「コンピュータ」と「コンピューター」の扱いや，「表す」と「表わす」の扱いなどである．

これら同一性の問題に対して，伝らは以下のような細分化を行い，語彙素，語形，書字形，発音形という四つのレベルで見出しを定義することによって，これらの同一性を表現している[1]．

**細分化：**

(1)　語形の変異
- (1a)　活用語の語尾変化
- (1b)　語の複合に伴う語頭音の変化
- (1c)　語の複合に伴う語末音の変化
- (1d)　口語活用と文語活用の違い
- (1e)　サ行変格活用の五段化・上一段化
- (1f)　外来語の語形の違い
- (1g)　慣用読みによる変化
- (1h)　その他の音の転化

(2)　表記の変異
- (2a)　送り仮名の違い
- (2b)　新旧字体の違い
- (2c)　漢字と仮名の違い
- (2d)　漢字の違い
- (2e)　外来語の表記の違い

(3)　発音の変異
- (3a)　外来語の発音の違い

**語彙素**：変異を考慮せず，元来同一と見なしうる語に対して同一の見出しを与えたもの

**語形**：同じ語彙素に所属するものに対して，活用や音変化などによる形態の変異を区別したもの

**書字形**：同じ語形に所属するものに対して，表記の変異を区別したもの

**発音形**：同じ語形に所属するものに対して，発音の変異を区別したもの

上記のうち，(1a)から(1c)までは一つの語形から変化形を派生させることにより表現する．(1d)から(1h)までは同じ語彙素に所属する異なる語形として登録する．更に(2a)から(2e)は同じ語形に所属する異なる書字形として，(3a)は同じ語形に所属する異なる発音形

---

[†] 索引語が含まれる文書や，その位置などを記録したインデックス

[††] スペース，タブ文字，改行文字などの空白文字

[†††] 一様，等質であること

として登録する．例えば，UniDic を用いると，(1 f)については，図1のように「アイディア」と「アイデア」が同じ語彙素(lemma)「アイディア」に所属するという解析結果が得られ，(2 a)については，図2のように同じ語形(form)の異なる書字形として解析される．

アラワス【表わす】という語彙素の語形としては「アラワス」が，書字形としては「表わす」「表す」「あらわす」が登録されている．なお，「計算機」と「コンピュータ」は異なる語彙素となるため，このレベルでは同一性を表現することはできない．シソーラスや類義語辞書といった別の仕組みを用いる必要がある．

このように，形態素解析によって得られた単語列に対して統計的な処理を施すことで，テキストマイニングは行われる．主な手法としては，単語や単語連鎖(n-gram)の出現頻度，二つの単語間の相関や，近傍での共起確率を用いた解析が行われる．単語の出現頻度によって，どのような語がよく用いられているか，相関や共起確率によって，どのような語が関連して出現するかがわかる．ただし，相関がある，ないし共起するということはわかるが，それがどのような意味を持つかを自動的に取り出すことは難しい．多くの場合，高確率で共起するということは，何らかの強い関係があると推定する．

工藤らの Web 日本語 N グラム[2]は，Google がクロールした，一般に公開されている Web ページ(文数約200億文)から，出現頻度20回以上の1〜7グラムを構築しており，その際に

（1） 文字コード変換
（2） 正規化
（3） 文の分割
（4） 対象文の同定，選別
（5） 単語分割

といった前処理を施している．

形態素解析結果に対して，更に深い自然言語処理として，係り受け解析や同義表現の置き換えを行うことによって，より高度なマイニングを行うことが可能となる．

黒橋らのオープンサーチエンジン基盤 TSUBAKI[3]では日本語の Web ページ約1億件を対象に，深い言語解析に基づくインデキシングを行うことにより，より柔軟な検索を可能としている．例えば，「風邪薬を飲む時の留意点」は以下のように解析される．

・形態素解析
　風邪/薬/を/飲む/時/の/留意/点
・係り受け解析
　風邪→薬，薬→飲む，飲む→時，時→留意，留意→点
・同義表現
　風邪＝感冒，薬を飲む＝服用，留意＝注意

これにより，感冒→服用や注意→点といった係り受けをもつものまで広げてマイニングを行うことができる．

## 3. 社会活動のトレース

テキストマイニングの技術を用いて，社会活動のトレースを行う例として，奥村らの blogWatcher がある(現在は公開を終了している)[4]．ブログのテキスト群を対象にテキストマイニングを行っているのだが，中でもブログのテキストから月ごとのブログ著者たちの行動傾向の分析を行っている点は興味深い．ブログは一種の日記であるから，テキストの書かれた日時の情報が取得で

```
<cha:D xmlns:cha="http://www.unidic.org/chasen/ns/structure/1.0">
  <cha:S>
    <cha:W1 orth="アイディア" kana="アイディア" pron="アイディア" pos="名詞-普通名詞-一般" orthBase="アイディア" kanaBase="アイディア"
      pronBase="アイディア" lForm="アイディア" lemma="アイディア" form="アイディア" aType="1,3" aConType="C1"
      goshu="外">アイディア</cha:W1>
  </cha:S>
  <cha:S>
    <cha:W1 orth="アイデア" kana="アイデア" pron="アイデア" pos="名詞-普通名詞-一般" orthBase="アイデア" kanaBase="アイデア" pronBase="アイデア" lForm="アイディア" lemma="アイディア" form="アイデア" aType="1,3" aConType="C1" goshu="外">アイデア</cha:W1>
  </cha:S>
</cha:D>
```

図 1　UniDic による(1 f)の解析結果

```
<cha:D xmlns:cha="http://www.unidic.org/chasen/ns/structure/1.0">
  <cha:S>
    <cha:W1 orth="表す" kana="アラワス" pron="アラワス" pos="動詞-一般" cType="五段-サ行" cForm="終止形-一般" orthBase="表す"
      kanaBase="アラワス" pronBase="アラワス" lForm="アラワス" lemma="表わす" form="アラワス" aType="3" aConType="C1"
      goshu="和">表す</cha:W1>
  </cha:S>
  <cha:S>
    <cha:W1 orth="表わす" kana="アラワス" pron="アラワス" pos="動詞-一般" cType="五段-サ行" cForm="終止形-一般" orthBase="表わす"
      kanaBase="アラワス" pronBase="アラワス" lForm="アラワス" lemma="表わす" form="アラワス" aType="3" aConType="C1"
      goshu="和">表わす</cha:W1>
  </cha:S>
</cha:D>
```

図 2　UniDic による(2 a)の解析結果

きるため，時系列に沿った分析が可能になる．

新聞記事についても同様のマイニングが可能である．テキストマイニングでその時々の話題を抽出することで，それらの変遷を見ることが可能となる(動向分析)．

一般に，テキストマイニングで取得することができる情報は，マイニングの対象として設定したテキスト群の大域的な傾向である．先の時間情報のようにテキストと組み合わせて利用できる情報(メタデータ)があれば，それを利用して，より細かな分析も可能になる．

例えば，ある製品に関する自由記述のアンケート結果に対してマイニングをかけると，その製品に関する全般的な評価の傾向が得られる．更に，回答者の様々な属性を分析に加えることができれば，どのようなユーザがどのような評価を下しているかといった分析が可能になる．このとき，統計的に有意な情報を得ようとすると，ある程度の量が必要となることはいうまでもない．

別の例として，1．にあげたWebからの犯罪情報のマイニングのような場合には，マイニング対象はWeb全体となり，その中から該当する情報をさがすというタスク設定になる．この場合には，取り出したい情報が決まっていて，膨大なテキスト群からそれをいかに抽出するかが問題となる．この場合，たどり着くのはテキストまでであって，そこから著者を特定する部分はプロバイダなどに任される．

現代社会においては，ある種の利便性と引き替えに，個人情報を提供する場面が少なくない．そのような情報は提供された側で適切に管理されていることが前提ではあるが，たとえ情報が匿名化されていても，それらの断片をかき集めることによって，トレースが可能となることもあり得る．一方で，データマイニングの分野でもPrivacy-preserving data mining(PPDM)というテーマでの研究もなされている．

アンケート結果やカスタマーセンターに寄せられた意見からのマイニングなどは，特定の当事者のみが取得できるテキストを対象とするのに対して，ブログなどWeb上で公表されているテキストを対象としたマイニングは，現状ではクローリング[†4]や複製の問題は残るものの，誰にでも可能なものである．少なくとも，ウェブ上で公表したテキストは検索やマイニングの対象になるということは，十分認識しておく必要がある．

[†4] インターネット上のWebページを巡回しながら，情報を収集すること

## 4．おわりに

本稿では，主に自然言語処理の観点からテキストマイニングの概要について述べ，社会活動のトレースという側面から考察を加えた．近年CPUや記憶媒体の性能があがり，個人でも比較的容易にマイニングを試すことができる状況になってきている．また，公開情報からのマイニングだけでなく，例えば企業内の非公開情報を対象にしたマイニングも重要性を増している．本稿が，技術と社会の関係を考える契機の一つとなれば幸いである．

## 参考文献


1) 伝　康晴，小木曽智信，小椋秀樹，山田　篤，峯松信明，内元清貴，小磯花絵："コーパス日本語学のための言語資源―形態素解析用電子化辞書の開発とその応用―"，日本語科学，22，pp. 101-123 (2007)．
2) 工藤　拓，賀沢秀人："Web日本語Nグラム第1版"，言語資源協会 (2007)．
3) 黒橋禎夫，新里圭司："TSUBAKI：深い言語処理を特長とするオープンサーチエンジン基盤"，情報処理，Vol. 49，No. 8，pp. 931-934 (2008)．
4) 奥村　学，南野朋之，藤木稔明，鈴木泰裕："blogページの自動収集と監視に基づくテキストマイニング"，人工知能学会，セマンティックウェブとオントロジー研究会，SIG-SWO-A 401-01 (2004)．


## URL

JUMANホームページ：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
ChaSenホームページ：http://chasen-legacy.sourceforge.jp/
MeCabホームページ：http://mecab.sourceforge.net/
UniDicホームページ：http://www.tokuteicorpus.jp/dist/
TSUBAKIホームページ：http://tsubaki.ixnlp.nii.ac.jp/index.cgi


山田（やまだ）　篤（あつし）

1986年京都大学工学部情報工学科卒．1991年同大学院博士後期課程研究指導認定退学．同年京都大学工学部助手．1996年より(財)京都高度技術研究所勤務．自然言語処理，音声言語処理，XML変換系などの研究に従事．博士(工学)．京都大学情報学研究科情報社会論分野客員准教授．情報処理学会，自然言語処理学会などの会員．