テキストマイニング，データマイニングと社会活動のトレース

山田篤（京都高度技術研究所）
yamada@astem.or.jp

１．はじめに

データマイニングとは，大量のデータを対象にして，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。このうち，特にテキストデータを対象にしたものをテキストマイニングという。ここでいうテキストデータとは，非定型的な自然言語表現のことである。たとえば，アンケートなどで，いくつかの選択肢の中から選択された結果の集合は，アンケート自体は自然言語で記述されていたとしても，離散的な値の集合であり，テキストマイニングの対象とはならない。これに対して，同じアンケートでも回答者による自由記述の部分はテキストマイニングの対象となる。

テキストマイニングにおいては，対象となる自然言語記述を計算機上で処理する必要があるため，計算機で扱えるように電子化する必要がある。近年，ウェブ環境などで電子テキストが増大し，マイニングの対象にすることができるようになったことで，テキストマイニングが盛んに行われるようになった。

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技術

電子化テキストにおける最小の構成単位は文字である。文字毎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。与えられたテキスト群に対して，最も原始的には，この文字を単位として，各文字の出現頻度や，文字の連鎖をｎ字組で表現し，ある文字の次に生起する文字の確率といった統計量を計算することができる。ただし，文字を単位とすると，そこからなんらかの意味を抽出することが困難であるため，通常のテキストマイニングにおいて文字を単位とすることはほとんどない。

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，言語学的には単語とは何かを規定することは必ずしも簡単なことではない。一方，技術的には，使用する電子化辞書に登録されている単位ということになる。これは単に，電子化辞書に単語の定義を押しつけただけであるが，世の中で実際に用いられている電子化辞書の中には，ある語は一語として登録されているのに，別の語は語頭と接尾辞に分割されて登録されているといった，語の斉一性の面で問題があるものもある。斉一性が保証されていないと，統計的手法で解析する際に問題が生じることがある。たとえば，ある接尾辞の用法を取り出そうとすると，それが電子化辞書で単語の一部として登録されてしまっているものについては取り出せないことになる。

ともあれ，テキストマイニングでは，対象となるテキストを，形態素解析と呼ばれる技術を用いて単語に分割することで，単語を単位とした解析を行う。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（たとえば，活用語の原形等）も付与することができる。形態素解析では，このために先に述べた電子化辞書を用いる。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 や UniDic がある。

形態素解析を行うことで，文字単位ではなく，単語単位の全文検索を行うことができるようになる。

さらに，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，2 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までは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。

日本語の場合，助詞は文章中での格役割を示すマーカとして働くため，助詞を伴った動詞との共起を分析することで格フレームといったこともできる。

いずれにしても，マイニング技術で取得できるのは統計量であり，それに基づく仮説である。実際にそれをどのように評価するかは，それを読み取る側の責任である。

また，自然言語記述には，指示語等を用いた照応や，省略，言い換え等の問題もあり，それらを含めて意味の解析を行うためには，より高度な自然言語処理技術を用いる必要がある。

３．社会活動のトレース

テキストマイニングの結果から得られることは，あくまでも対象となるテキスト群に字面として書かれていたことがもとになる。ただし，それを当該テキストにかかるメタデータと組み合わせることで，より詳細な情報が得られる。言い換えれば，それがどのようなテキストであるかといった，字面には現れていない情報（メタデータ）が重要である。

たとえば，ある特定の人物の書いたテキストばかりを集めて解析を行うことで，その人物の文体や用語法の特徴を捉えることができるかもしれない。

ある商品に関する自由記述のアンケート結果を集めて解析をすることで，その商品に関する評価を取り出せる可能性もある。

さらに，テキストにメタデータとして時間情報が付与されていれば，そこから時系列による変化を読み取ることもできる。もちろんテキスト内に時間に関する記述があれば，その情報を抽出することは可能であるが，一般にテキストには，そのテキストが書かれた時間と，テキストの中に書かれている時間が存在する。

すなわち，テキストマイニングを適切に行うためには，

・マイニング対象となる母集団

・そこから何を読み取りたいか

の設定が重要になる。

一般には，マイニング対象となるテキストは，匿名性を持ち，誰のものかわからない大量のテキストを対象とするが故に，テキストを取得した母集団となった社会の大域的なトレースができるにとどまる。それでも，その時々における社会のトレンドといったものは確実にテキストに反映される傾向がある。たとえば，新聞記事を対象にマイニングを行うと，その時々のトピックが得られ，その時系列的な変遷を知ることができる。研究論文を対象にしたトレンド分析も可能である。

逆に，母集団を特定の集団や個人といった単位で絞り込むことができれば，作家の文体研究のように，その母集団の特性を抽出することができるかもしれない。たとえば，ショッピングサイトにおいて，ある特定の商品を購入した人は，別の特定の商品を購入する傾向にあるといった情報は，購買履歴のデータマイニングから得ることができる。同様のことをテキストを対象にして行うためには，テキストに付与されたメタデータないしテキスト内の記述をもとに集団間の相関を計算することになる。たとえば，アンケート結果からのマイニングや，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等が実用的な応用としてある。最近では，ブログを対象にしたマイニングという報告もある。少なくとも，ウェブ上に公開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．おわりに

本稿では，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，手元のパソコンを使って個人でも簡単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，たと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。

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/