Vol.2016-CH-112 No.1
2016/10/29

絵入源氏物語のテキストデータに対する統計解析 webアプリケーションの設計

土山 玄†

近年、人文系オープンデータの活用の重要性は様々な場で議論されており、活用の事例も広く報告されている。そこで、本研究では国立情報学研究所より公開されている絵入源氏物語のテキストデータを用いたwebアプリケーション開発した。一般に、テキストデータの計量分析では統計解析のためのプログラミング言語であるRが用いられる。しかし、テキストデータについて基礎的な統計処理を行う場合においても、Rなどについてある程度習熟する必要があり、統計学の初学者にとっては敷居が高いと言える。そこで、本研究ではこのテキストデータを対象とし、統計学の初学者であっても利用可能なwebアプリケーションの設計について報告する。

Web Application Development for the Quantitative Analysis of *The Illustrated Tale of Genji*

GEN TSUCHIYAMA†

In recent years, many humanities researchers have discussed the importance of using open data. Even in Japan, many universities and institutes have released various data sets and applications for using the data. The National Institute of Informatics has released the text data of *The Illustrated Tale of Genji* into the public domain. This tale is a famous work of Japanese classical literature. To analyze text data, researchers commonly use R, which is a well-known programming language, but users must be acquainted with R to use it. Therefore, it is difficult for scholars of the tale to analyze its text data. Then, we developed a web application for the statistical analysis of the text data in the tale. In this study, we report on the development of the web application for users who are beginners in statistics.

1. はじめに

オープンデータの学術利用についてはすでに議論されており、多くの分野において活用されている。このようなオープンデータの活用は人文学分野においても例外ではなく、近年はオープンデータの様々な活用事例が報告されている[1]。そもそもオープンデータとはあらゆる制限から自由であり、誰もが再利用及び再配布が可能なデータである[2]。このようなオープンデータには画像データやテキストデータなど様々なデータが公開されているが、本研究ではテキストデータについて採り上げる。

日本において、公開されており自由に利用できるテキストデータについては、青空文庫というwebサイトが有名である[1]。青空文庫において公開されているテキストデータは著作権の消滅した、あるいは著者が許可した近現代の小説や随筆などの文学作品のテキストであり、このようなデータは広く学術利用されている。例えば、計量文献学と称されるような文章を対象に統計的な研究を行う分野においては、文学作品の著者識別[3]や執筆時期の推定[4]など多くの研究成果が報告されている。近現代の文章を対象とした計量的な研究において、このようなオープンデータは非常に重要な役割を果たしていると言える。その一方で、古典文学を対象とした計量的な研究は十分に展開されているとは言い難い。これは近現代の文章に比べて、古典文のテキストデータを作成するためには国語国文学の専門的知識を必要とするということ、及び校訂本文の取扱いに注意を払う必要が大きいからであると考えられる。しかし、現在は国立情報学研究所において「国文研古典籍データセット（第0.1版）」[2]が公開されており、古典文学作品を取り巻くオープンデータの環境も整ってきたと言える。この「国文研古典籍データセット（第0.1版）」は国文学研究資料館が所蔵する約30万点の古典籍を画像化したデータベースの構築を目指す「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」において作成されたデータセットである。

本研究では「国文研古典籍データセット（第0.1版）」において公開されているデータの1つである『源氏物語』のオープンデータを用いたアプリケーションを作成した。この『源氏物語』のオープンデータにはテキストデータが含まれており、本研究において報告するアプリケーションはこのテキストデータを用いて語の出現傾向を可視化するwebアプリケーションである。

† 同志社大学 研究開発推進機構
Organization for Research Initiatives and Development, Doshisha University

1 http://www.aozora.gr.jp/

2 http://www.nii.ac.jp/dsc/idr/nijl/nijl_list.html

2. 源氏物語のデータセット

2.1 絵入源氏物語とそのテキストデータ

先にふれた国立情報学研究所より公開されている『源氏物語』のオープンデータは江戸時代に出版された刊本の『絵入源氏物語』をデジタル化したデータである。この『絵入源氏物語』は江戸時代に何度か刊行されているが、本研究において用いたデータの底本は承応三年に刊行された刊本である。

公開されているテキストデータには『絵入源氏物語』の本文の他に巻名や頁数、行数が含まれている。以下にテキストデータの一例として第１巻「桐壺」の冒頭を示した。

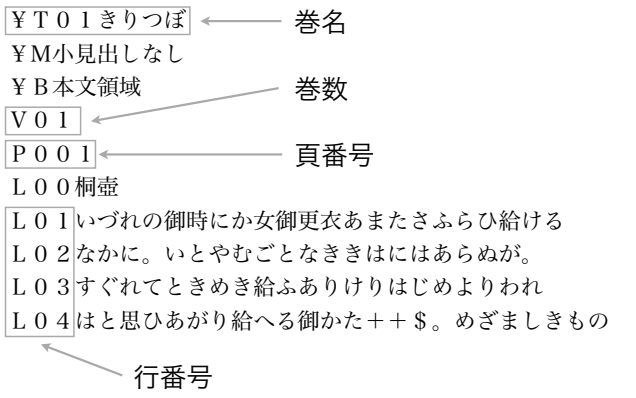

図１　テキストデータの一例

また、このテキストデータは踊り字についても原則としてそのまま表記されている。踊り字とは日本語における特殊記号の１つであり、上掲の例には含まれていないが、「々(同の字点)」や「ゝ (一の字点)」などがこれに該当し、これらは直前の１文字の繰り返しを意味する記号である。しかし、踊り字には同の字点や一の字点の他に図２に示すような「くの字点」と称される踊り字があり、これは直前の２文字以上の繰り返しを意味する。

図２　くの字点の例

くの字点をあらわす記号が存在しないため、『絵入源氏物語』では「＋＋」という記号が使用されており、上掲の図１における「御かた＋＋＄」がこれに該当する。なお、「＋＋＄」は２文字以上の繰り返しに加え、繰り返し部分の文字列にあるいずれかの文字が濁音になることを含意している。

上述したように、本研究では語の使用傾向を目的としたwebアプリケーションの開発を目的としている。そのため形態素解析を行う際に、この「くの字点」が「＋＋」であるとノイズとなり、適切に単語を認識することができない。従って、何らかの処理を加えなくてはならない。これについては後述する。

2.2 活用事例

また、『絵入源氏物語』のテキストデータはすでに活用事例があり、国文学研究資料館より「絵入源氏物語データベース」[3]が公開されている。この「絵入源氏物語データベース」は文字列検索を目的としたwebアプリケーションであり、ユーザが検索したい文字列を入力することで入力された文字列が含まれる前後の文章及び検索文字列のあらわれる冊数名や頁などが返される。

3. webアプリケーションの設計

3.1 webアプリケーションの概要

すでに述べたように、近代や現代の小説を対象とした計量的な研究は広く展開されており、その研究成果もまた数多く報告されている[3][5]。そのようなテキストデータを対象とした計量的な研究では、統計解析のためのプログラミング言語であるRが用いられることが多い。しかし、テキストデータについて基礎的な統計処理を行う場合においても、Rなどについてある程度習熟する必要があり、統計学の初学者にとっては敷居が高いと言える。近現代の文章については、MTMiner [6]やKHcorder [7][8]と言ったような高度なプログラミング技術を必ずしも要求しないフリーウェアがあり、統計学の初学者にとっても文章の計量分析は十分に可能である。

その一方で、やはり古典文を対象とした統計解析を支援するアプリケーションは少ない。そこで、本研究では『絵入源氏物語』の本文を対象に、統計学の初学者が基礎的な統計処理を行えるwebアプリケーションの設計を試みた。本研究では『絵入源氏物語』の54巻における任意の単語の出現傾向を可視化するwebアプリケーションについて報告する。

3.2 テキストデータの加工

先にふれたように、国立情報学研究所より公開されている『絵入源氏物語』のテキストデータは踊り字のくの字点が「＋＋」あるいは「＋＋＄」と表記されている。加えて、くの字点を意味する「＋＋」の指示する文字列が直前の２文字を繰り返すのか、あるいは直前の３文字以上を繰り返すのか、自動で修正するための資料がない。そのため、「＋＋」の指示する文字列は文脈から判断するしかない。そこで本研究では原文にあたり、手作業でくの字点についての加工を行った。なお、表１はくの字点が３文字以上の文字列を指示した例である。

３ http://base1.nijl.ac.jp/infolib/meta_pub/G0001501genji

表 1　3 文字以上を繰り返す「くの字点」の例

本文	加工後の文字列
ころ＋＋$	（と）ころどころ
ゝろ＋＋$	（こ）ゝろごころ
返し＋＋	（うち）返しうち返し
はり＋＋$	（か）はりがはり
給へ＋＋	（なのり）給へなのり給へ
陀仏＋＋	（阿弥）陀仏阿弥陀仏
うし＋＋$	（ざ）うしざうし
へる＋＋$	（か）へるがへる
君$＋＋と	（あかしのあま）君あかしのあま君

3.3 形態素解析

本研究において加工した文字列は上述のくの字点に関わる箇所だけである。ただし、これに加えて、図 1 において示した記号などは除去した。このような処理を加え、『絵入源氏物語』のテキストデータについて形態素解析を行った。

用いた形態素解析ツールは国立国語研究所より公開されている「Web 茶まめ」[4]である。「Web 茶まめ」では形態素解析を行う際に UniDic、すなわち辞書を選択することで、上代から現代までの多様な文章を形態素解析することが可能である。本研究では中古和文の辞書を選択し、『絵入源氏物語』の本文の形態素解析を行った。

3.4 R 及び Shiny の利用

本研究において報告する web アプリケーションの統計的な処理については R を用いた。R は統計解析向けのプログラミング言語であり、様々な統計処理を容易に行うことが可能である。また、R には多用なパッケージがあり、これを R にインストールすることでより多くの機能が R に実装されることとなる。そのような R のパッケージの 1 つに Shiny があり、これは R 言語で記述されたプログラムを web アプリケーションにするパッケージである。本研究においてもこの Shiny を用い、『絵入源氏物語』における語の出現傾向を可視化する web アプリケーションを開発した。

図 3 は Shiny を用いて作成した『絵入源氏物語』の各巻における任意の語の出現率を可視化した際に web ブラウザに表示される画面をキャプチャーした画像である。図 3 は「あはれ」という語の出現率を可視化したグラフであるが、図 3 において示したテキスト入力用フィールドにユーザが任意の語を打ち込むことでその語の出現傾向が可視化される。また、ここで計算される出現率とは入力された語の各巻における頻度の延べ語数に対する割合である。なお、活用する語については終止形を入力することで全ての活用形を一括して集計する。

次いで、形態素解析の際に、すべての語は品詞タグが付与される。これを用いることによって、各巻における品詞の比率を求めることも可能である。図 4 は品詞の比率を可視化した際の web ブラウザをキャプチャーした画像である。図 4 において示したアプリケーションではセレクトボックスに『絵入源氏物語』の各巻がリストされており、ユーザが任意の巻を選択することで、その巻の品詞の比率が棒グラフとして表示される。

このように本研究において概観した web アプリケーションは、『絵入源氏物語』の文体に係わる定量的な要素を容易に可視化できる。また、可視化されたグラフを R について習熟していないユーザであっても、インタラクティブに操作できることから、このような web アプリケーションは古典文学作品を対象とした研究者を支援しうると考えられる。

4. まとめ

本研究において概観した『絵入源氏物語』のテキストデータを用いた web アプリケーションは、従来では統計学やプログラミング言語に習熟していなければ実行できなかった統計処理をより容易に行えると考えられる。しかし、現段階ではこのアプリケーションは極めて初歩的な統計処理

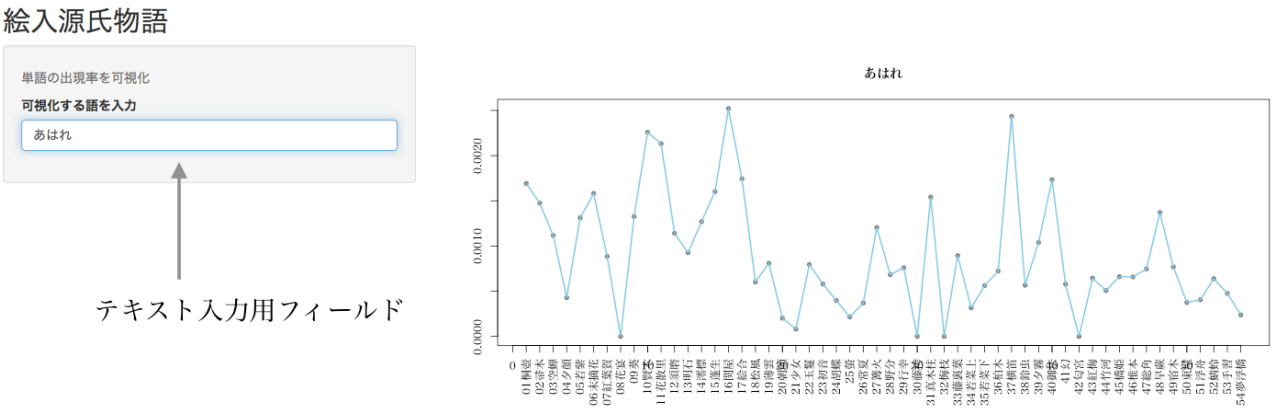

図 3　語の出現傾向の可視化

4 http://chamame.ninjal.ac.jp/

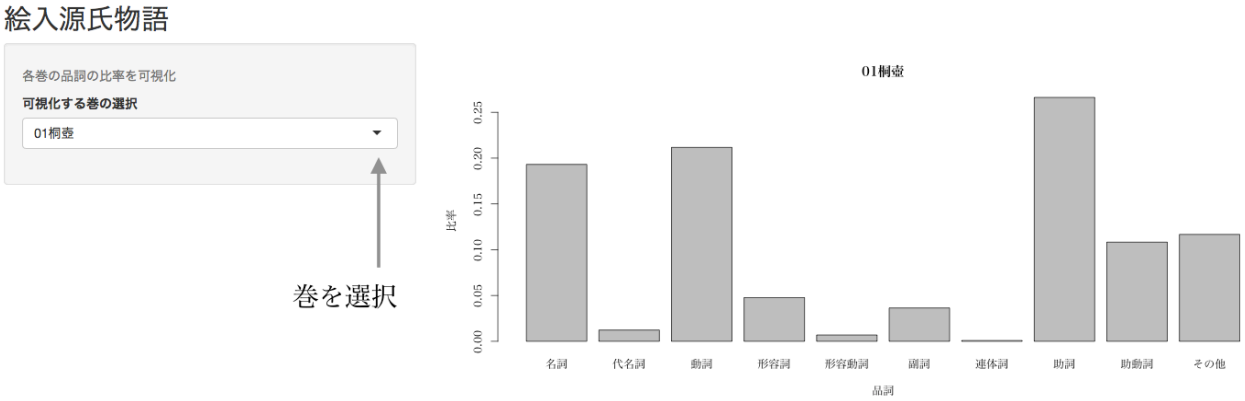

図 4　品詞の構成比率の可視化

しか実行できない。そこで、今後の展望として、主成分分析やクラスター分析などの多変量解析を実装することで、『絵入源氏物語』の諸巻の文体的特徴による分類など、より実践的な統計処理が可能な web アプリケーションを開発し、公開する予定である。

参考文献

1)　橋本雄太. (2015). 人文学資料オープンデータの可能性と現状. 情報の科学と技術, 65(12), 525-530.
2)　Open Knowledge Foundation. (2012). オープンデータとは何か? – Open data handbook. <http://opendatahandbook.org/guide/ja/what-is-open-data/> (参照 2016-10-04)
3)　村上征勝. (2002). 文化を計る−文化計量学序説. 朝倉書店.
4)　金明哲. (2009). 文章の執筆時期の推定―芥川龍之介の作品を例として―. 行動計量学, 36(2), 89-103.
5)　計量国語学会. (2009). 計量国語学事典. 朝倉書店.
6)　金明哲,張信鵬. (2013).テキスマイニングツール MTMineR のコンセプトと機能. 日本行動計量学会大会発表論文抄録集, 41, 360-363.
7)　樋口耕一. (2015). フリーソフトウェア「KH Coder」による計量テキスト分析: 手軽なマウス操作による分析からプラグイン作成まで. 研究報告人文科学とコンピュータ (CH), 2015(9), 1-2.
8)　樋口耕一. (2014). 社会調査のための計量テキスト分析―内容分析の継承と発展を目指して. ナカニシヤ出版.