

KeyGraph を用いた時系列における聖書の比較解析による考察
 Consideration by comparison analysis of Bible in time series that uses KeyGraph

新井智也 伊藤誠
 Tomonari arai Makoto ito

1. はじめに

近年、様々な情報の電子化が進むと同時にインターネットの普及により、WEB上には様々な知識と情報が蓄積されるようになった、その膨大且つ今後も爆発的に増大するであろう知識・情報から次の時代に役立つ人間の潜在的なニーズや活用方法を取り出そうという試みが企業、研究者によってなされている。本研究ではデジタル化された諸々のデータ、特に文章データに着目し蓄積されたデータに対して時間の流れと焦点の当て方を考慮したキーグラフについて考察した。

2. 背景と研究目的

インターネットが普及し多くの人が世界中の情報にアクセス出来るようになりメールや電子掲示板、BLOG、SNSなどの電子媒体やコミュニティーにおいて様々な意見を交換できるようになった、また書籍、資料などを初め様々なコンテンツを電子化しデジタルアーカイブする試みが各国、団体主導の下で活発に進められており電子図書館などネットワークを通して膨大な知識や情報を手に入るようになった。しかし、このようなデータの巨大集合やデータベースから次の時代に求められるものは何なのか？個人が意思決定において重要な判断材料となる事象・状況とは何のかが課題となってくる[1] またこのような膨大なデータを取り扱う場合、データマイニングの手法や焦点のあて方などが問題になる。

また、もう一つの課題として時代の流れと共に文化が多様に変化・革新されるように言葉も時代ごとに意味内容が変化し、新しい言葉となっていく性質を持つ長期的な時代に跨る文章データなどをキーグラフで取り扱う場合このような言葉の変化に個別に対応した機能が必要であると考えた。

3. 聖書を題材とした比較解析

聖書を分析するにあたってキーグラフを用いることにした、キーグラフは文章データの可視化手法の一つであり、その文章を形態素と形態素との関係性をマップ上に可視化するツールである。キーグラフは頻度の高いキーワ

ードを黒ノード(図1)とし黒ノード同士を共起確率に従いリンクでつなぎ黒ノードのグループ(島)を形成する、また赤ノードは頻度は低い、複数の黒ノードのグループ同士を橋渡しし連結するキーワードである。キーグラフにおいて価値の高い言葉とは、単に頻出で理解されやすい言葉の主張よりも頻度は低く理解されにくいゆえにこそ新たに理解する価値の高い言葉として潜在的な意義を見出すことにある。

3.1 聖書の特徴と分析手法

本研究では扱うデータとして聖書[3][4]を使用した。ここで言う聖書は一般的にキリスト教で正典とされているいくつかの文書群からなる旧約聖書、新約聖書のことを指す。聖書は世界各国 200以上の言語に翻訳されており多くの偉人、著名人に影響を与え続けている

聖書はそれぞれ異なった時代、場所、職業の人間、約40人によって書かれており、書かれている時代の間隔はおよそ1500年にも及ぶ、また書かれている形式も多様であり歴史書、物語から詩、歌、箴言(格言)、預言書など様々である。聖書をキーグラフで処理するには幾つかの問題がある、まず聖書自体のデータが巨大でありキーグラフの性質上十分な効果を発揮できない可能性が高い、これは上位頻度の単語同士が強くクラスタ化してしまうため、巨大なクラスタと小さなクラスタを結びつけるものや、共起確率が高くなる低頻度の単語同士での橋しか見つけられなくなってしまうからである。(図1)はキーグラフの標準ツールである「Polaris」を用いマルコの福音書のデータを解析パラメータ黒リンク赤リンク共に計算方法を共起頻度に設定し黒ノード数を100 黒リンク数50 赤ノード数30 緑ノード数15にパラメータを設定した解析結果である、先に述べている通り上位頻度の単語「イエス」が巨大なクラスタ化してしまい、小さなクラスタと結びつけるものしかみつけられなくなってしまうっておりノード数の多さもあってとても見辛いものとなっている。全体を分析するだけに限定すればノードの数を少なくし不要語の設定を行えばシンプルな解析結果を得ることは出来る(図2参照)、しかし、問題となるのは得られたデータから重要な部分を絞り込んで掘り下げることである。聖書の長い年月をかけて記述されており、その時間の流れを聖書は旧約聖書39巻、新約27巻聖書併せて66巻から成る文書群であるが書簡によって意図的に時系列的に記述されていない場合や、書簡同士の順序も時系列でない場合がある、基本的には時系列によって無視できない前後関係で構成される、あるいは前後関係は与えられていないが単語間の共起関係であるデータが混在していると言える、更に系列の最初と最後で単語間の因果関係の構造が変化していく特徴を持つ、書簡内や書簡を跨ぐ文章の場合、ある時代での単語の意味と違う時代の単語とは意味内容が異なる場合があり、任意の文章単位

† 中京大学

