KeyGraphを用いた時系列における聖書の比較解析による考察

新井智也† 伊藤誠†

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維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が、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で、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いまい、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では意味内容が異なる場合があり、任意の文章単位

Consideration by comparison analysis of Bible in time series that uses KeyGraph

†Tomonari Arai Makoto Ito, Chukyo University

での辞書の適用と文章の前後関係を考慮したキーグラフの必要があると考えた。

4．文章データの時系列処理

文章データの前後関係を考慮したキーグラフの提案として前処理の段階で任意のデータ範囲に辞書を適用し時間軸に従ってオーバーラップ的に文章データを処理し時間の流れに沿ってキーグラフによる有効グラフのような可視化を行いキーグラフ同士を比較できる機能を持ちたいと考えた、キーグラフを用いた代表的なツールとして「Polaris」「紙芝居キーグラフ」などがある、焦点を絞るにあたって深く掘り下げたいキーワードに対して特定の位置にキーワードを固定し複数の文章データを比較する必要が出てくると考えられるが Polaris には任意のキーワードを位置固定する機能があるが一時的なものであり固定位置を保存すること機能は無く、時間の流れを含めた複数の文章データの処理を行うことも現段階では出来ない、「紙芝居キーグラフ」ではデータ毎に細かく辞書の設定や複数のキーグラフを比較することは出来るが特定のキーワードに焦点をあててキーグラフを比較することは出来ない。

5.まとめ

キーグラフにおいて時系列で巨大なデータを扱うモデルとして聖書を題材として選択した場合の問題点課題を挙げ、時間の流れと文章データ間を挟んだ場合のデータ処理の必要性と文章の流れ時代によって変化する単語に対する分析と方向性を持ったキーグラフの有向グラフ化に対する提案を行った。キーグラフの有向グラフから得られるデータによってどんな気付きが起こるかは分析者に委ねられるが今後、実際の機能を組み込んだツールの開発をしていきたいと考えている。

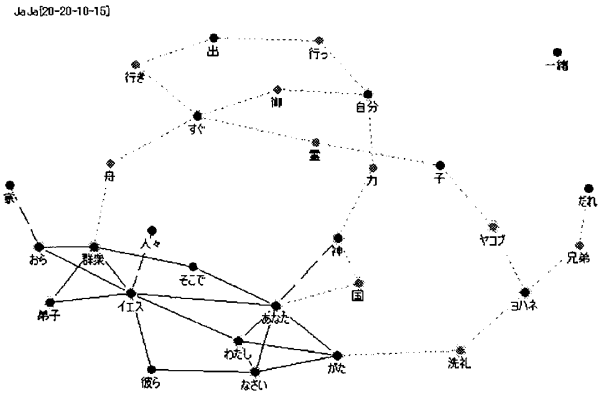

図２

参考文献

[1]『チャンス発見の情報技術―ポストデータマイニング時代の意思決定支援』大澤幸生 2003 東京電機大学出版局

[2]『チャンス発見のデータ分析―モデル化+可視化+コミュニケーション→シナリオ創発』大澤幸生 2006, 東京電機大学出版局

[3]『新共同約聖書』日本聖書協会

[4]『新改訳聖書第三版』新改訳聖書刊行会

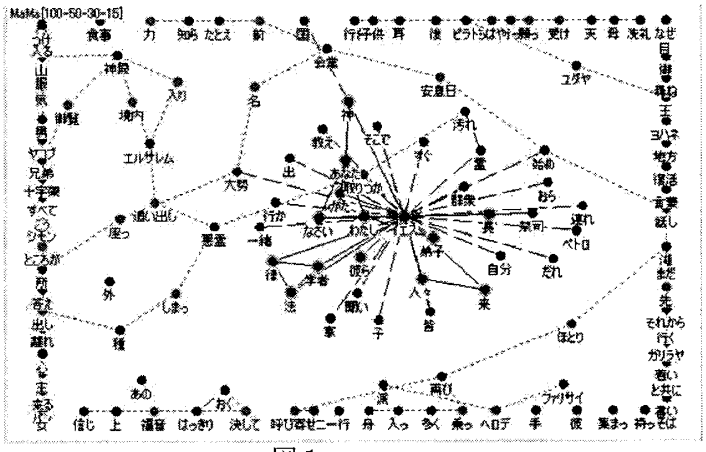

図１