Web内容マイニングによる個別知識の共通単語を介した相関ルールからの発想支援

1615052　山元悠貴　　情報基盤工学講座　指導教員 奥原浩之

要約

本研究では，Google検索を用いた複数のキーワードからの発想支援について考える．既存の発想支援システムでは1つのキーワードから発想を広げていくものが多かった．そこで，複数のキーワードから関連する単語を用いてランク付けすることで，より効率の良い発想支援につながると考えた．

キーワード：テキストマイニング，共起ネットワーク，ベイジアンネットワーク，発想支援

1 はじめに

現在，情報処理技術の発達に伴って，コンピュータが人間の創造的な問題解決・思考活動を支援する発想支援システムの研究が進んでいる．これからの時代はよりアイデア発想が重要になってくると考える．創造活動をする際，人間は言葉で表現することが多く，その言葉を用い，自分の発想を整理し，修正を行う．実際，認知心理学では，人間にとって思考と言語は深い関連があると言われている[1].

そのため，人工知能が発想支援を行うためには，人間の言葉を知る必要がある．しかし，自然言語を機械に理解させるのは非常に困難であり，機械独特の自然言語の分析方法が用いられている[2].

発想を支援することは，人間からアイデアが出やすくすることと考えられる．そのため，機械が発想支援を行うには，人間が考えるために使う手段である自然言語の分析方法を利用する必要がある．そこで現状の発想支援について，複数のキーワードからの発想支援に着目した．

本研究では，複数のキーワードについてのネットワークの作成とキーワードの関連語をランキング形式で表示する事を考える．

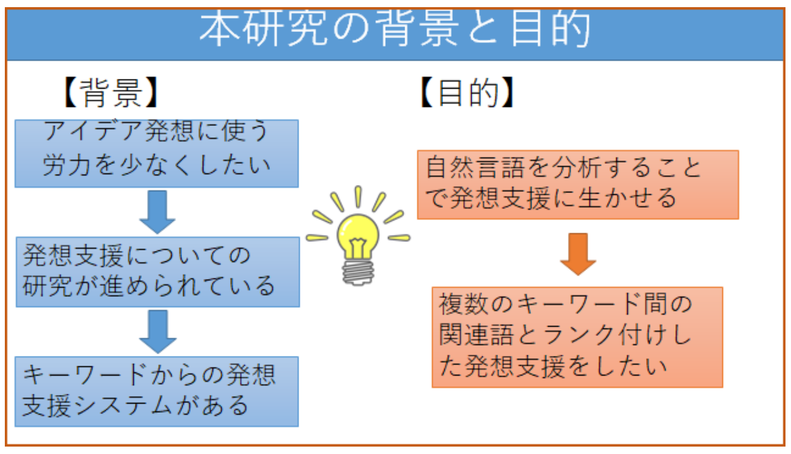

図1 研究背景と目的

2 発想支援システムとは

2.1 サイバー空間からのテキストデータ収集

現代社会においてインターネット上の情報量は莫大になっており，今後も増え続けることが予想される．このインターネット上の情報を収集して分析することで発想支援に生かせると考える．発想支援において重要なことはキーワードからより関連度の高い単語をより多く表示させることである．そこで，より良いデータを多く収集するためにサイバー空間からテキストデータを収集することとする．

今回，キーワードごとにGoogle検索から上から何件分かのURLを取得し，そのURLからテキストを抽出してそのテキストに対して自然言語処理を行い，必要な単語を取り出す．その単語群から共起ネットワークを作成する．

図2にデータ収集に用いるテキストマイニングの説明を示す．

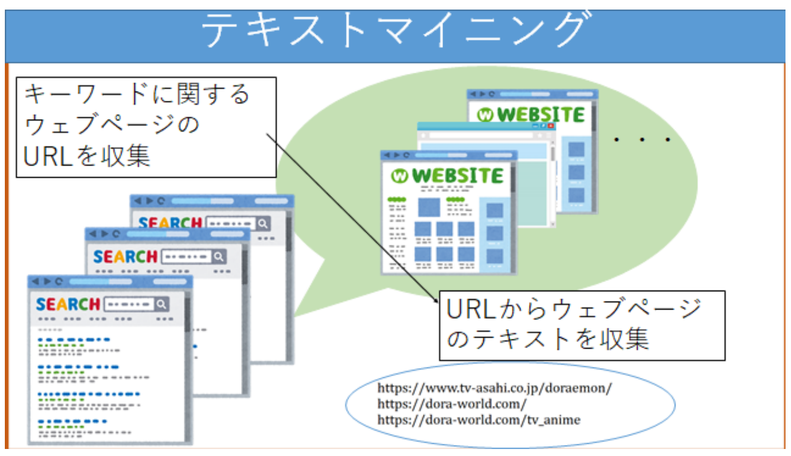

図2 データ収集の流れ

2.2 テキストマイニング

テキストマイニングとは，大量かつ多量なデータを様々な観点から分析し，役に立つ情報を取り出そうとする技術である．

インターネット上のテキストを用いることで大量のデータを活用することができる．まず，収集した文章に対してHTMLタグやJavaScriptのコードを取り除くクリーニング処理をする．その次に形態素解析を行う．形態素解析とは文を形態素ごとに分解する技術である[3]．発想支援において必要な名詞や動詞に分解することである．また，助詞の「は」や「が」助動詞の「です」は不要なので取り除く．そして単語群に対して正規化を行う．正規化することで文字種を統一できる．半角と全角や数字を統一することで同じ単語として扱うことができる．

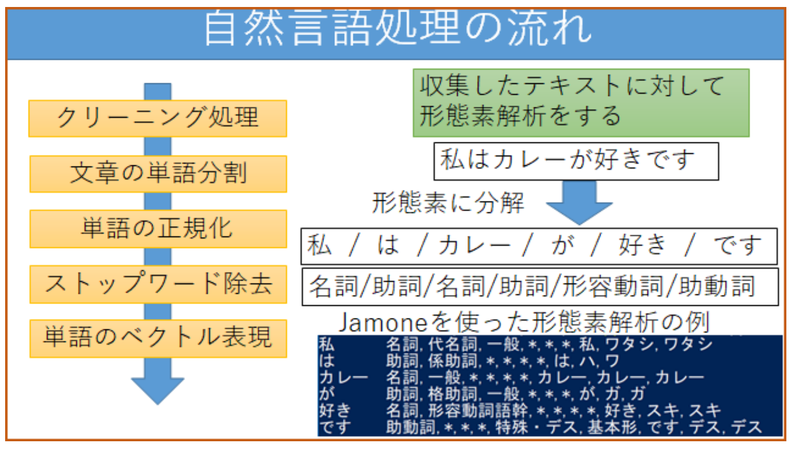

図3 自然言語処理の流れ

2.3 既存の発想支援と課題

既存の発想支援システムにはAIブレストスパークのひらめきマップや，ブレストアイデアがある[4]．ひらめきマップはキーワードを入力するとその単語に共起する関連語を表示して発想支援に役立てるというものである．ブレストアイデアはキーワードを用いたワード，フレーズの生成をするシステムである．また，KH coderを用いた階層的クラスタ分析を用いて単語の関連性を分析した研究がある[5]．課題としては表示される単語のランク付けがされないことが挙げられる．

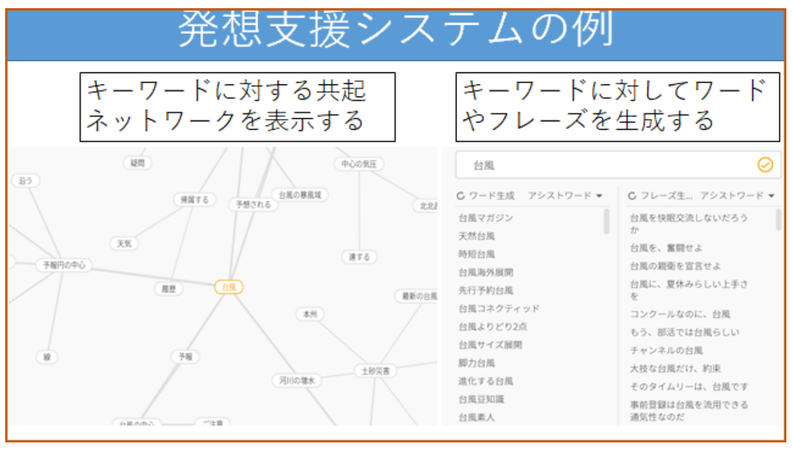

図4 既存の発想支援システムの例(キーワード：台風)

3 データドリブンによる最適化

3.1 単語の分散表現ベクトル

自然言語処理の分野において，単語をベクトル表現するWord2Vecという技術がある．複数のキーワードの関連語がなかった場合にも発想支援できるように単語をベクトル表現することで近いベクトルの単語同士を関連語として見れるのではないかと考える．

Word2VecについてSkip-gramでモデル化する．Skip-gramでは，ある単語を入力した時，その周辺にどのような単語が現れやすいか予測することをモデル化することができる．

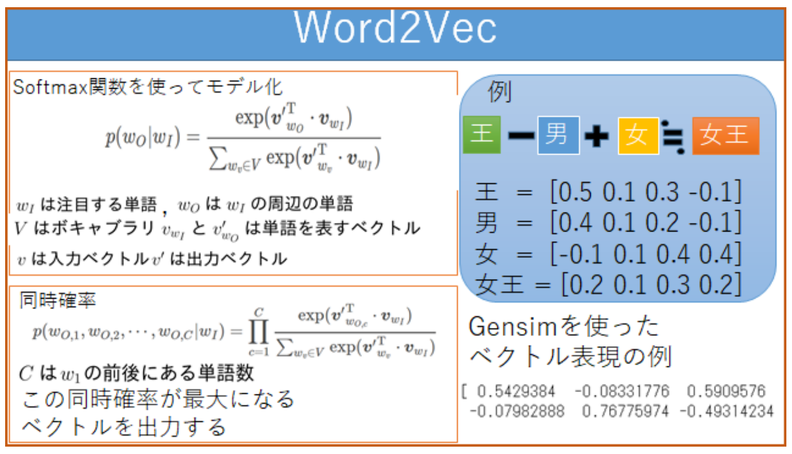

図5 Word2Vecのモデル化

3.2 複数のキーワード間の共起ネットワーク

ある単語とある単語が同時に出現することを共起するといい，文章において関係深い単語は共起することが多い．共起分析では単語同士のJaccard係数を比較したり，共起関係を持つ単語と単語を線で結んで描かれる共起ネットワークが利用される．文章また単語群に対して共起する単語をネットワークで表した共起ネットワークという．今回，キーワードごとに集めたテキストに対してそれぞれの共起ネットワークを作成する．

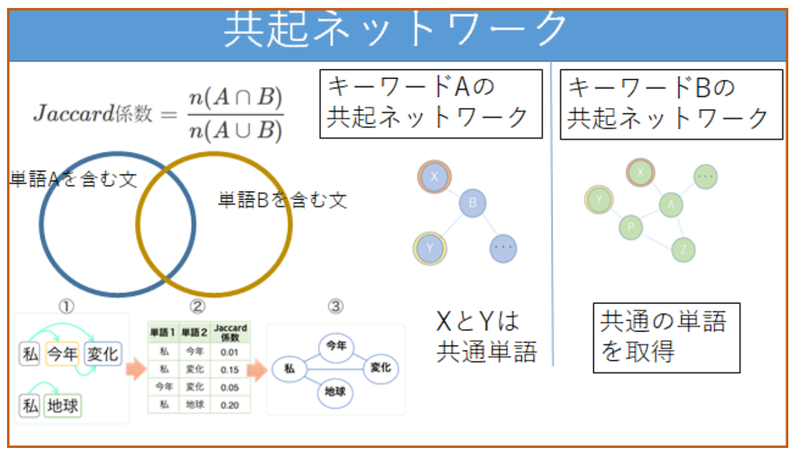

図6 複数の共起ネットワークにおける共通単語

3.3 単語の遷移による共通単語のランク付け

ベイジアンネットワークにより単語のつながりを可視化する．単語群から，単語の遷移をにカウントして隣接行列を作成する．Softmax関数から単語の遷移確率を求めてベイジアンネットワークを作成する．隣接行列をPandasで作成し，隣接行列からNetworkXというPythonのライブラリを使用してベイジアンネットワークを作成する．

それぞれのネットワークのある共通単語について，単語の推移確率の和をパスの長さで割った値を双方向から求め，その和をある共通単語のスコアとする．このスコアの高い単語をランク付けして共通単語とスコアの値を表示する．

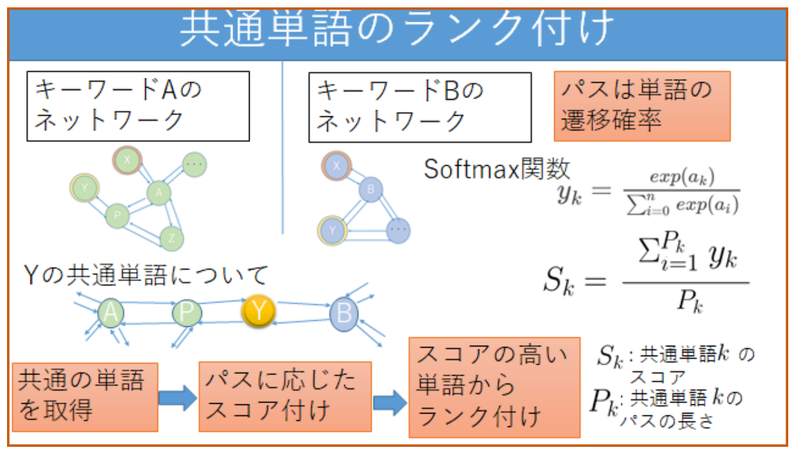

図7 提案手法の概要

4 数値実験ならびに考察

今回，キーワードごとにGoogle検索から上から20件分のURLを取得し，そのURLからテキストを抽出してそのテキストに対して自然言語処理を行い単語群からベイジアンネットワークを作成する．単語の遷移確率をテキストデータから算出し，隣接行列からベイジアンネットワークを作成する．単語の遷移確率の隣接行列は正しく表示できたが，networkXでベイジアンネットワークを作成したとき，エッジに重みがあり，双方向の遷移が存在する場合片方の確率遷移のみしか表示されなかった．

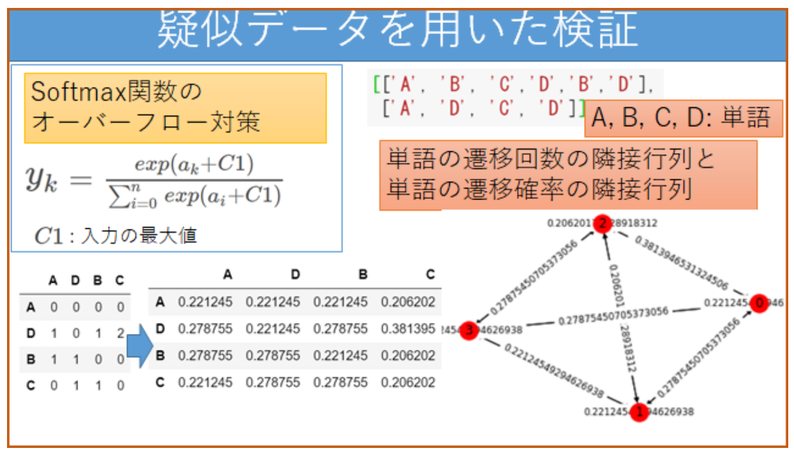

図8 疑似データによるベイジアンネットワークの可視化

5 おわりに

本研究では，複数のキーワード間の発想支援についてベイジアンネットワークを用いた単語の遷移確率を使ったスコア算出によるランク付けについて考えた．今後の課題はURL取得からランク付けまでのすべての工程をつなげることと，ネットワーク作成のとき共通の単語がなかった際にどうするか考える．また，システムの評価方法についての考察をする必要がある．

参考文献

[1] 國藤 進, “発想支援システムの研究開発動向とその課題”, 人工知能学会誌, pp. 552-559, 1993.

[2] イ スンジュ, “発想支援のためのテキストマイニング”, 人工知能学会 第25回 セッションID: 1P2-10in, 2011.

[3] “加藤 耕太 “Pythonクローリング＆スクレイピング ーデータ収集・解析のための実践開発ガイド”, 技術評論社, pp. 130-131, 2017.

[4] TIS 株式会社, “AI ブレストスパーク”, https://www.ai-b-spark.com/, 閲覧日: 2019年10月30日.

[5] 伊藤 順子・東 孝行・宗森 純, “単語共起度の低い単語を提示する発想支援システムの提案と適用”, 情報処理学会論文誌, pp.1528-1540, 2015.