

シナリオプランニングにおける合意形成支援のための関連概念可視化の提案 Visualization of Relevant Concepts for Consensus Building Process in Scenario Planning

張 确 軒[†]
Quexuan Zhang

大 澤 幸 生[†]
Yukio Ohsawa

1. はじめに

企業や組織は事業計画を行う際、シナリオプランニング [1] という手法を活用している。シナリオプランニングとは、ステークホルダー間のコミュニケーションにより、問題発見を促進し、将来起こり得る事象を考慮しながら解決案を策定するプロセスである。シナリオプランニングには、まずシナリオの創出が必要である。Innovators Marketplace on Data Jackets [2] というゲーム形式のシナリオ創出支援技法では、参加者が複数のデータのメタ情報 (DJ) の関係図を見ながらシナリオを創出し、要求を提示したり解決策を提案する。それにより、従来にない課題や解決案が数多く考案されている。しかし、複数の領域を横断した意見交換では、シナリオにより得られた課題の中の概念に対して参加者間の概念理解に矛盾が生じたり、時間の制約から参加者間で気づき得なかった重要な課題が未検討のままシナリオが策定されてしまう危険性がある。その結果として、課題の認識が不完全となり、本質的な解決策に至らなくなる可能性がある。

本研究では、シナリオプランニングにおいて、課題の検討と計画の策定を支援することを目的とした、関連概念を可視化する手法を提案する。シナリオの関連概念の発見には、潜在意味解析 (LSA) を使用し日本語版 Wikipedia から学習された意味空間を利用する。可視化には、DBpedia でカテゴリなどの情報を検索し、得られたリソースを利用する。関連概念の可視化を用いることで、シナリオプランニングの参加者に問題発見や解決案策定のヒントをもたらし、外部情報の検索時間を節約することが期待できる。

2. 意味空間

2.1 LSA

LSA [3] はベクトル空間モデルを用い、トピックの集合を含む意味空間を生成し、その関係を分析する手法である。意味空間は、文書-単語マトリクス (出現行列) を構築し、その出現行列に特異値分解 (SVD) などの階数低減手法を行うことで得られる。出現行列とは各行が各単語に対応し、各列が各文書に対応した疎行列である。この行列の各要素はその単語に対応した文書での $tf-idf$ 値を用いられることが多い。階数の低減により、データの格納に所要するメモリを低減したり、ノイズ (非重要語) を除去したり、文書の関連する単語を獲得することができる。

元の出現行列は M (単語数) \times N (文書数) 行列の場合、SVD を通じて、 $M \times L$ (トピック数) の単語トピック-単語、 $L \times L$ の文書トピック-単語トピック、 $L \times N$ の文書-文書トピックの三つの行列を生成する。コサイン相関量などの関連度計算手法によって、それらの行列から異なる文書間や異なる

単語間の相関度を得ることで、文書群や単語群のクラスターが得られる。さらに、新しい文書がクエリとして与えられた際、その文書を意味空間に変換することで、意味空間内での文書群と比較して関連の文書を検索することができる。そこで、情報検索の領域では、LSA は潜在意味索引 (LSI) とも呼ばれている。

2.2 Wikipedia

本研究では、意味空間の学習には日本語版 Wikipedia¹ をコーパスとして利用する。Wikipedia は自由に編集に参加できる多言語のインターネット百科辞典である。日本語版には現在 97 万本以上の記事がある。文書量や網羅した領域が多く、人間の知識によって編集されてきたことで、統計に基づく LSA における意味空間の学習にとって適切なコーパスだと考えられる。

また、全ての Wikipedia の記事にはタイトルがあるので本稿ではタイトルを概念と定義する。

2.2.1 事前処理

Wikipedia をコーパスとする前に、日本語版 Wikipedia の dump データ² を獲得し、事前処理を行う。

合意形成支援の役に立たない記事は事前に除去する。不要記事は以下の条件を満たす。

- 曖昧さ回避ページ
- 年、月、日、世紀、元号などの時間に関連する記事
- あるテーマの一覧ページ
- 標準の名前空間³に属されない記事

日本語の文書に含まれる単語群を獲得するため、形態素解析 (MA) を行う。本研究では MA ツールとして MeCab⁴ を使用する。MA の精度を高めるため、記事の Raw データの中のノイズを除去する。本稿では、以下のデータをノイズと見なし、除去した。

- Wiki markup やフォーマットコード
- テンプレートやカテゴリ、ファイルなどの記述
- URL
- 編集者のコメントや脚注

また、文中の括弧は MA の品詞判断にノイズを与えるため削除されるが、括弧の中の内容は抽出され、新しい文としてその文の後ろに追加される。

さらに、Wikipedia のタイトルは辞書に登録され、文書の単語群には、一般名詞と Wikipedia のタイトルのみ用いる。

2.2.2 意味空間の構築

事前処理済みの Wikipedia 記事を用いて、行に MA から得た単語を対応させ、列に記事を対応させ、要素に単語の記事中の出現回数を示した出現行列を構築する。次に、その行列の各要素に対して $tf-idf$ 値を計算し、SVD を行うことで、意味空間が得られる。本稿では、2014 年 12 月 30 日の dump を利用し、総出現数は 10 回未満かつ文書総数の

[†] 東京大学, the University of Tokyo

1. <http://www.wikipedia.com>
2. <http://dumps.wikimedia.org/jawiki/>
3. <https://ja.wikipedia.org/wiki/Help:名前空間>
4. <http://taku910.github.io/mecab/>

10% (約7万3千) の文書で出現した単語を除去し、上位10万語かつ400トピックで構築された意味空間を用いる。

また、関連度の計算はコサイン相関量を用いるため、事前にベクトルの長さを計算しておけば、意味空間内の文書と関連度を計算する際の所要時間を短縮することができる。

2.3 シナリオ文

シナリオは文書として創出されるので、シナリオ文に対して、事前処理と同じMA手法によって単語群を抽出し、その単語群を利用してシナリオ文を意味空間に変換することができる。そして、関連するWikipediaの記事を検索することで、関連概念を取得できる。

以下のシナリオ文を例として関連概念を検索する。

「2020年の東京五輪開催が決まったことを受け、恩恵が予想される関連業界では早くも期待が高まる。五輪期間中の来場者数は延べ約1000万人。東京都と政府は観光客受け入れのためにインフラ整備を急ぐ考えだ。だが五輪開催は真夏の暑い時期。快適な移動手段や十分な宿泊施設をどう確保するか、過密都市ゆえの課題も多い。人手不足で施設の建設コストが膨張する懸念も浮上している。」⁵

この文をクエリとして、以下に示すような関連概念(関連度上位20語)を獲得できる。

マスダール・シティ, スプロール現象, 都市化, フランシスコ・ペレイラ・パソス, 配置計画, 交流居住, 南條道昌, 中心市街地, 都市計画, 農村戸籍, シムシティ frontier, クエンカ(エクアドル), 観光都市, ラインハルト・パウマイスター, 集団規定, 都市の鍼治療, 石川栄耀, アメニティ, 都市緑地法, 田園都市

意味空間はWikipediaの様々な領域の記事から学習し、得られたため、文書のタイトルは、一般意味上の概念だけでなく、都市名や人名などの固有表現も含まれる。そこで、Wikipediaのカテゴリ情報を利用して上位概念を獲得することで、上位概念による関連概念のフィルタリングやクラスタリングが可能となると考えられる。

3. DBpedia

DBpedia[4]は、Wikipediaから情報を抽出してRDFによって情報を構造化したものである。RDFでは、リソースに関する関係情報は主語、述語、目的語の3つの要素で表現される。クエリ言語SPARQLによって、関連の情報を検索することができる。

3.1 述語による検索

DBpediaでは、リソースはWikipediaの記事に対応している。述語dbpedia-owl:wikiPageIDを利用すれば、Wikipediaの記事のIDからDBpediaの対応したリソースへアクセスすることができる。dcterms:subjectとskos:broaderの2つの述語を用いれば、概念の所属するカテゴリと上位カテゴリを検索することができる。また、述語rdf:typeによって、リソース類型も獲得できる。

例えば、「マスダール・シティ」は「再生可能エネルギー」、「環境問題」、「アブダビ」、「各国のエネルギー政策」、「アラブ首長国連邦の計画都市」に所属する。そのリソース類型はgeo:SpatialThingとなる。「環境問題」の

5. <http://www.nikkei.com/article/DGXNZO59486280Q3A910C1EA2000> より

6. <https://ja.wikipedia.org/wiki/Category:総記>

上位カテゴリは「自然保護」、「環境」、「政治問題」、「社会問題」となる。

Wikipediaのカテゴリ情報は単純な階層関係ではなく、ネットワーク構造となっている。そこで、ボトムアップでの最短経路による3階層までの最小木の情報を利用する。

3.2 ノイズカテゴリ

Wikipediaのカテゴリ情報には、概念理解に対して有用でないと考えられるカテゴリが存在する。そこで、以下の条件を満たすカテゴリはノイズと見なして除去される。

- メタ分類(例:○○別の△△,○○に分類した△△,○○のカテゴリ,各○○の△△)
- Category:総記⁶に属される大部分のカテゴリ

4. 可視化の考案

シナリオ文をクエリとして得た関連概念集合と、関連概念のカテゴリ情報によって、各概念の可視化での配置位置を計算する。また、リソース類型やカテゴリ情報によってフィルタリングを行う機能を実装できる。可視化は以下の情報を表す。

(1) 概念定義の表示:Wikipediaの記事の概要情報から抽出された一言の定義を表示することができる。

(2) 概念とシナリオの関連度:図の中心はシナリオと見なされ、意味空間から得た関連概念の関連度によって概念の半径を計算する。

(3) 概念の所属カテゴリ:同じカテゴリに所属する概念に対して、同じ色で着色する。ただし、一つの概念が所属するカテゴリや上位カテゴリが複数存在する場合、着色の基準となるカテゴリはそれらのtf-idf値によって決める。

(4) 概念間の関連度:意味空間によって概念群のクラスタが得られる。関連度の高い関連概念を互いに近い場所に配置する。

5. まとめ

シナリオプランニングにおける合意形成を支援するため、Wikipediaの知識を意味空間に変換し、それによってシナリオの関連概念を提示する可視化手法を考案した。今後の課題として、可視化ツールの実装と支援効果を測定する検証実験の実施が挙げられる。また、Wikipediaの記事だけでなく、DJや従来の課題と解決案も文書と見なすことができる。それらを学習に用いた意味空間により、シナリオとの関係性を可視化することで、計画策定の支援ができると考えられる。

参考文献

- [1] Schoemaker PJH., "Scenario planning: a tool for strategic thinking." Sloan management review, Vol.36, No.2 (1995).
- [2] Ohsawa Y., Kido H., Hayashi T., Liu C., Komoda K., "Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data.", Knowledge-Based Information Systems in Practice (2015).
- [3] Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., Harshman R. A., "Indexing by latent semantic analysis.", JASIS, Vol.41, No.6 (1990).
- [4] Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P.N., Hellmann S., Morsey M., van Kleef P., Auer S., Bizer C., "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, Vol. 6 No. 2(2015).