語間の関係性を考慮したサブトピック抽出法に関する一検討

萩原 一貴† 波多野 賢治††

† 同志社大学大学院文化情報学研究科 †† 同志社大学文化情報学部

a) *hagi@ilab.doshisha.ac.jp* b) *khatano@mail.doshisha.ac.jp*

概要 情報技術の発展により，ユーザはインターネットを通じて容易に情報行為を起こすことが可能になった．ユーザは Google 検索や Yahoo! 検索に代表される検索エンジンにクエリを入力ことで，目的とする情報の探索を行っているが，曖昧なクエリを用いて検索した場合，ユーザの意図にそぐわない検索結果が提示されてしまう．そこで，我々は Web ページ内に出現する語間の関連性を考慮し，そのクエリのサブトピック，つまり，ユーザの求めるより詳細な内容を推定し，要求に見合う情報を提示する手法の一検討を行う．

キーワード サブトピックマイニング，語の関連性，クエリ拡張

1 はじめに

近年，情報技術の発展によりユーザはインターネットを通じて容易に情報行為を起こすことが可能になった．ユーザは，検索エンジンに対して要求するクエリを入力することで，情報を得てきたが，入力されたクエリが曖昧なワードであった場合に表示される検索結果は必ずにもユーザが求める情報を含むとは考えにくい．例えば，ユーザが「大阪のデパート」というクエリを用いて検索した場合,「大阪のデパート」に関連する Web ページが表示されるが，大阪のデパート自体の詳細が知りたいのか，関連するニュースや他人の意見が知りたいのかなど，複数の選択肢が存在するため，一度の検索で満足する結果を得ることができるとは考えにくい．

そこで本研究では，与えられたクエリに対して Web ページ内の語間の関係性を考慮したサブトピックを抽出し，ユーザに提示するための手法について検討する．これにより，ユーザは的確に求める情報を得る事ができるようになると考えられる．

2 関連研究

自然言語処理の分野において語の意味を表現する手法として注目を集めている研究の一つに，Mikolov らの研究 [1] が挙げられる．

これは,「同じ文脈に存在する語は同じ意味を持つ」という考え方に基づいており，非線形統計モデルであるニューラルネットワークを用いて単語を概念ベクトルとして表現することが可能になる [3]．また，word2vec では語から周辺の語の出現確率を推定するために Skip-gram や CBOW(Continuous Bag-of-Words) と呼ばれる学習モデルが使用されている．例えば Skip-gram モデルでは，目的関数

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

を最大化させることを考える．ここで，$w_1, w_2, w_3, ..., w_T$ は与えられる語の並び，c は学習するコンテキストのサイズを表す．ここで，$p(w_{t+j}|w_t)$ は式 (2) のソフトマックス関数式を用いて，定義され，v_w と v'_w はそれぞれ w に対する入力と出力を表し，W は単語数を表す．

$$p(w_O|w_I) = \frac{\exp({v'_{wO}}^{\top} v_{wI})}{\sum_{w=1}^{W}\exp({v'_w}^{\top} v_{wI})} \tag{2}$$

また，これをオープンソース化したソフトウェアに word2vec [1] がある．

word2vec[1] を応用した研究は多数されており，その中の一つにユーザの検索意図推定を目的とした Song らの研究 [2] では，クエリログからのサブトピックの抽出と意図の推定の 2 ステップに分けた提案を行っている．Song らは，検索を行った際のクエリのミスマッチを軽減させるため，語の意味関係に基づいて新たな語をクエリに追加し，サブトピック生成のために語のクラスタリングを行っている．彼らはサブトピックの生成には一般的な k 平均法を用いており，word2vec は単語のベクトルを学習するためだけに用いられている．また，k 平均法は偏ったクラスタリングに不向きであるとされており，Web ページのような分類対象数を確実に把握できない場合にはふさわしくないと考えられる．

3 提案手法

Song らは サブトピック生成のために k 平均法を用いていた．k 平均法によるクラスタリングの手順は，図 1 に示す通りであり，

[1] word2vec, https://code.google.com/p/word2vec/

1. ランダムに仮のクラスタの重心を指定した数だけ作成する．(ここでは，仮のクラスタの重心を 3 つ指定し，×で表す)
2. 全ての点と 1. で作成した仮のクラスタの重心との距離を求め，一番近い仮のクラスタの重心とつなげる．
3. それぞれの仮のクラスタにつながっている点の座標の平均値を求め，その点を新しいクラスタの重心として更新する．(ここでは仮のクラスタの重心を白い×，新しく更新されたクラスタの重心を黒の×で表す)
4. 1. から 3. を変化がなくなるまで繰り返す．

である [4]．

k 平均法を用いるメリットは，高速に実行可能である点と逐次データが追加されていくような場合に再度実行可能な点であるが，その一方で，単純な k 平均法を用いることのデメリットとして，乱数を用いて初期のクラスタ重心を決定しているため，同じようにクラスタリングを行ったとしても同じクラスタが抽出されるとは限らないというデメリットも存在している．

そこで，我々は Song らの手法とは異なり，k 平均法を用いてサブトピックを生成するのではなく，word2vec を使用したサブトピック手法の提案を行う．具体的に，与えられたクエリを用いて検索した結果得た Web ページ内の文に対して形態素解析を行い，word2vec を用いて元のクエリと近い意味関係にある語を抽出する．一つの語から複数の候補を提示することを考えるため，学習モデルには Skip-gram を用い，入力されたクエリから同一文脈中に存在する単語を推定し，元のクエリと組み合わせることでサブトピックを生成する．

4 おわりに

本稿では，ユーザの情報行為を支援するために word2vec を用いたサブトピック抽出法に関する検討を行った．従来の k 平均法に基づくサブトピック作成では，初期のクラスタの重心よって結果が変化する可能性があり，また，偏ったクラスタリングには不向きであるとされているため，純粋に元のクエリからの意味的な近さを求めることのできる word2vec を用いた手法は有用であると考えられる．

今後の課題として，クエリとして検索される語は，具体的な人物名や商品名などの固有名詞が多いと考えられるため，適切に形態素解析を行う為に辞書の整備をする必要があると考え足られる．また，word2vec を用いても一語で複数の意味を持つ多義語の解釈をすることは困難であると考えられるため，word2vec を適用させる前にユーザがどの意味で用いているのか明確にするといった対策を講じる必要がある．

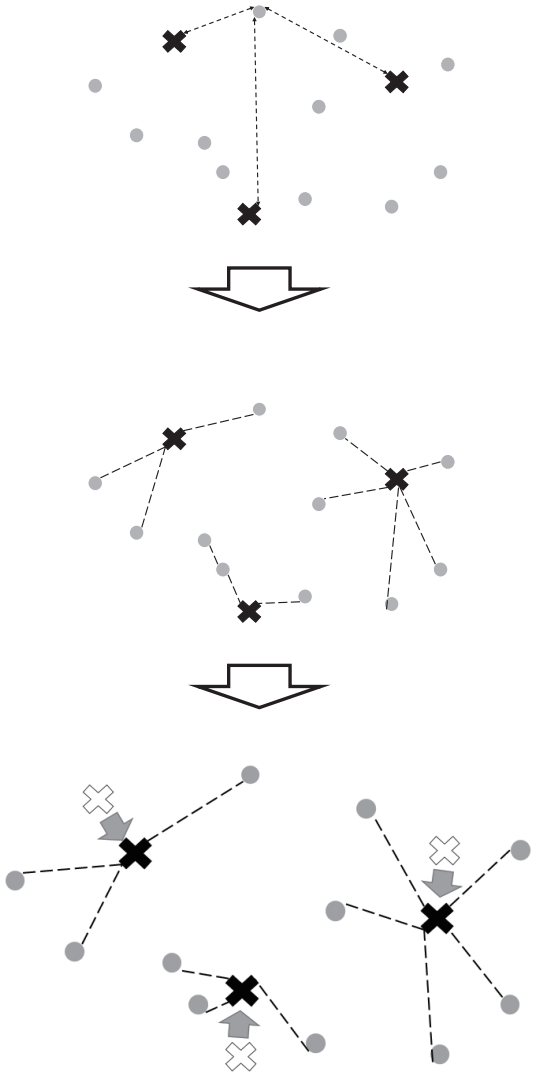

図 1 k 平均法によるクラスタリングの手順の例

謝辞

本研究は，日本学術振興会の科学研究費補助金 (26280115) の支援によるものである．

参考文献

[1] Milkov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distrinbuted Representations of Wods and Phrases and their Compositionality, In NIPS, March 2013.
[2] Song, W. Xu, W., Liu, L. and Wang, H.: CNU System in NTCIR-11 IMine Task,In NTCIR-11 IMine Task, December 2014.
[3] Hastie, T., Tibshirani, R., Ffriedman, J., 杉山将 監訳，井手剛 監訳，神嶌敏弘 監訳，栗田多喜夫 監訳，前田英作 監訳，井尻善久 他訳：統計的学習の基礎 - データマイニング・推論・予測 - ，共立出版株式会社，2014.
[4] 奥村学 ： 言語処理のための機械学習入門，コロナ社，2010.