3Q-01

Word2Vecによるユーザの嗜好を考慮した検索補助法の提案

DAO VAN TUAN† 佐藤 浩‡
防衛大学校理工学科研究科情報数理専攻†
防衛大学校電気情報学群‡

１．はじめに

爆発的に増加する Web 上の情報からユーザに適合した情報を取得するために，近年の検索エンジンでは検索のパーソナライズ化が行われることが多い．この手法は検索履歴を考慮することで検索結果を並び替えるが限られた期間の履歴に基づいたカスタマイズのため，時期によっては適合性が上がらない．また，ユーザには並び替えた検索結果のみが提示されるため，どのように自身の嗜好が反映されたかを容易に理解できない [1].

本研究では，個人の検索履歴に加え，過去に作成した文書などのデータを利用することで時期に限らないパーソナライズ化を実現するとともに，Word2Vec により作られるベクトル空間内での単語間の距離関係に基づいたパーソナライズ化によりユーザの嗜好を明示的にした Web 検索補助法を提案する．

２．Word2Vec

Word2Vec は Mikolov らにより発表された単語群のベクトル化手法である [2]．Word2Vec を自然言語処理の分野で応用した例に，日本語動詞・形容詞に関する類似度データセットの構築 [3]やイベント情報の分類 [4]など数多くの利用例がある．

本研究では Word2Vec を利用し，単語ベクトル間の距離を考慮することで，各被験者によって適合性が高くなるような検索語の拡張（関連語の提示）を行う．

３．提案手法

本研究では，個人の語彙を構成する単語が言語全体のベクトル空間の中でどのような分布を作るかに着目する．言語全体からなるベクトル空間を作るために，本研究では Word2Vec を用いる．言語空間上における人の語彙分布はその個人の特徴，あるいは個性であると考えられるため，これを検索補助の手がかりとして利用する．

前節で述べたように，本研究ではユーザが入力する検索語の拡張（関連語の提示）を行う.提示される単語の出力手順を以下に示す.

１）学習済みの Word2Vec を用い，検索語とのコサイン類似度が上位 5 件の単語を取得する．

２）上記の類似語と被験者の語彙分布の平均ベクトル間のユークリッド距離を計算する．

３）２）で得られたユークリッド距離を，検索語と平均ベクトルとのユークリッド距離と比較し，その差が小さい順に，類似語の再ランキングを行う．

４．実験

本実験では，2016 年 10 月 wikipedia 日本語版の約 4 億単語から 1/100 の量を抽出したものに被験者の語彙を加えたものを言語全体のコーパスとして用いた。コーパスの作成においては MeCab の分かち書き機能を利用して品詞を区切った．MeCab の辞書として MeCab-ipadic-neologd を用いた．また，個人の語彙は、検索履歴と過去に作成した文書から作成した．

Word2Vec のオプション [4] を以下に示す.

・モデル：Skip-gram
・最大ウィンドウサイズ：8
・ネガティブサンプル：25
・階層的ソフトマックス：なし
・最低出現回数： 1

被験者は防衛大学校の学生とし,検索語は、被験者が調べたい内容から作成する.

ここでは例として,“マッカーサー文民統制などの関連事項”を検索することを考える.この場合,検索語を“マッカーサー”とした.

Word2Vec のコサイン類似度で得られた類似語のランキングと提案手法による再ランキングの結果を表１に示す.

表１　提案手法による類似語の再ランキング

類似語	コサイン類似度	ユークリッド距離の差
ダグラス・マッカーサー	**0.93685（１位）**	0.30783（２位）
アメリカ陸軍士官学校	0.92542（２位）	0.65942（４位）
トルーマン	0.91551（３位）	**0.15766（１位）**
アイゼンハワー	0.89267（４位）	0.98560（５位）
ルーズベルト	0.89183（５位）	0.45785（３位）

検索語「マッカーサー」に対し，提案手法の１位である「トルーマン」は，コサイン類似度の１位である「ダグラス･マッカーサー」と比べ検索の目的に合致した拡張であると言える．被験者も実際そのように判定した.

比較のため，検索語「マッカーサー」に対する代表的な検索エンジンによるクエリ拡張を表２に示す．これらの拡張は，提案手法に比べ，被験者の検索目的に関連していないことが確認できた.

表２　他検索エンジンの結果

Google	Yahoo	Bing
通り	元帥	元帥
ギャレッジ	昭和天皇	道路
ガレージ	名言	名言
昭和天皇	道路	通り
パーク	ガレージ	昭和天皇

また,同じ検索語:“開校祭”,“防衛大学校”に対し，複数の被験者に提示される単語を表3 に示す．提示された単語は被験者によって異なる結果になった．被験者 1，2，3 は提示された単語が検索目的に近いと判断したが，被験者 4 は，検索語「防衛大学校」に対する「現在地」は検索に合致しないと判断した．この理由として，被験者 4 の語彙数の不足が考えられる.

表３　各被験者による出力結果

	被験者 1	被験者 2	被験者 3	被験者 4
開校祭	棒倒し	天気	防衛大学校	横須賀
防衛大学校	防大	2398686	陸上	現在地

５．おわりに

本研究では Word2Vec による検索語と被験者の語彙中心との距離に着目することで，適合性の高いクエリ拡張を実現する検索補助法を提案した．実験により，個人の語彙の平均ベクトルが重要な役割を果たすことが分かったが，コーパスにおける個人の語彙の割合が学習モデルにどの程度影響を与えるかを検討する必要があることが明らかになった．今後の課題として，学習モデルの精度向上および，実機への実装を行う.

参考文献

[1] 水野 淳太，村田 祐一，勝屋久：ユーザの嗜好を反映したクエリ拡張を用いた情報検索・推薦システムの開発，楽天研究開発シンポジウム 2009.

[2] Tomas Mikolov， Kai Chen， Grag Corrado，Jeffery Dean，“Efficient Estimation of Word Represention in Vector Space” Cornell University Library arXiv.org，arXiv;1301.3781v3[cs.CL]，2013.

[3] 堺澤勇也，小町守，“日本語動詞・形容詞類似度データセットの構築”，言語処理学会第 22 回年次大会， pp. 262-265，2016

[4] 小野 良太，川村 秀憲，“ディープラーニングによるイベント情報分類器に向けた Word2Vec の活用検討”，人工知能学会，pp. 1-6， 2016.

[5] 西尾泰和， “Word2Vec による自然言語処理，” O’ Reilly Japan， 2014.