ISS-A-005

階層からのカテゴライズ分析

船山　貴由†　塩谷　勇†
† 法政大学理工学部創生科学科

1 研究目的

階層構造の中に閲覧されるための文書が整理分類され保存されている場合、長い間使い続けると、階層構造の見直しである分類の進化が必要になる。しかし，カテゴリーの進化があった場合に、どのカテゴリーにオブジェクトが移動したかが利用者に解らなくなり、必要なものが見つからない。本研究では、時間の経過に対するカテゴリー内の変化を比較し、エージェントが利用者に興味のあると思われるトピックに関するオブジェクトを事前分類するシステムを検討する。

2. 研究方法

yahoo 知恵袋の 2019 年 12 月、2020 年 12 月に存在したカテゴリーの一つの「住宅」の中から、「DIY」「リフォーム」といった 16 個のカテゴリーにある文章各 10 個から単語を取り出し、単語リストを作る。2019 年同士、2020 年同士、2019 年と 2020 年の二つの年度のカテゴリーの単語の類似度を確かめ、値が高かった 3 カテゴリー同士について、両者に共通する単語を調べることで、カテゴリー同士を再分類できるか検討する。

3.研究手法

コサイン類似度を(1)式で求める。

$$\cos(X,Y)=\frac{X\cdot Y}{|X||Y|}=\frac{\sum_{i=1}^{|V|}XiYi}{\sqrt{\sum_{i=1}^{|V|}Xi^2}\cdot\sqrt{\sum_{i=1}^{|V|}Yi^2}} \qquad (1)$$

3. 研究結果

2019 年同士のカテゴリーの単語の類似度を示した図を図 1 に示す.

図1. 2019 年のカテゴリーの類似度

類似度が高い上位三つの組み合わせは、
「新築マンション」「中古マンション」類似度:0.38826
「リフォーム」「中古一戸建て」　類似度:0.381345
「リフォーム」「新築マンション」　類似度:0.345308
だった。
2020 年同士のカテゴリーの単語の類似度を示した図を図 2 に示す.

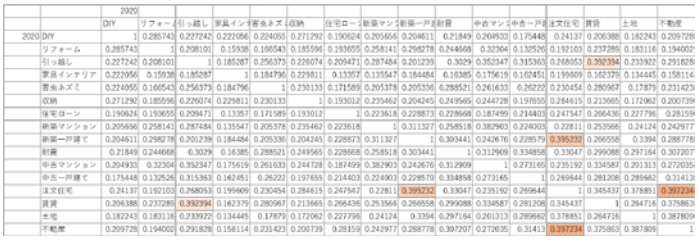

図 2. 2020 年のカテゴリーの類似度

類似度が高い上位三つの組み合わせは、
「不動産」「賃貸」　類似度:0.397234
「新築一戸建て」「注文住宅」類似度:0.395232
「賃貸」「引っ越し」　類似度:0.392394
2019 年、2020 年のカテゴリーの単語の類似度を示した図を図 3 に示す.

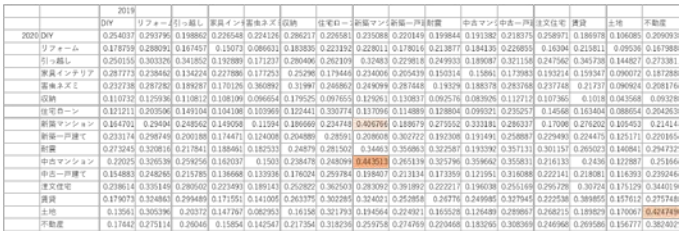

図 3. 2019、2020 年のカテゴリーの類似度

・2019「新築マンション」2020「中古マンション」
　類似度:0.443513
・2019「不動産」2020「土地」
　類似度:0.424749
・2019「新築マンション」2020「新築マンション」
　類似度:0.406766
となった。

4. 今後の課題

2019、2020、2019・2020 の類似度の高い上位 3 つのカテゴリー同士で、それぞれ主となっている単語を拾い、再分類を行えないか検討する。

参考文献

[1] https://chiebukuro.yahoo.co.jp/ Yahoo!知恵袋 － みんなの知恵共有サービス
[2] ベイジアンネットワーク入門(1) 須鎗　弘樹　2003 年
[3] Naive Bayes 分類における高頻度語と低頻度語の扱いについて