

階層からのカテゴリライズ分析

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、カテゴリの進化があった場合に、どのカテゴリ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の単語の類似度を確かめ、値が高かった 3 カテゴリ同士について、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。

$$\cos(X, Y) = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \cdot \sqrt{\sum_{i=1}^n Y_i^2}} \quad (1)$$

3. 研究結果

2019 年同士のカテゴリの単語の類似度を示した図を図 1 に示す。

2019															
2019 DIY	DIY	リフォーム	引っ越し	家具イン	暮れ暮れ	収納	住宅ロ	新築マン	新築一戸建	中古マン	中古一戸建	注文住宅	賃貸	土地	不動産
	0.28674	0.17607	0.19958	0.16423	0.25261	0.19784	0.27436	0.24066	0.38962	0.16198	0.15042	0.21651	0.18871	0.17043	0.20308
リフォーム	0.28674	0.29670	0.22021	0.17504	0.26871	0.16101	0.36356	0.26499	0.26054	0.21447	0.30356	0.21746	0.18461	0.17009	0.18146
引っ越し	0.17607	0.29670	0.32872	0.13486	0.20766	0.23302	0.21172	0.19294	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	
家具インテリア	0.19958	0.25261	0.17474	0.16072	0.25301	0.20729	0.24337	0.20274	0.16832	0.19635	0.18736	0.18866	0.10379	0.19246	
暮れ暮れ	0.16423	0.17504	0.13486	0.16072	0.25301	0.20729	0.24337	0.20274	0.16832	0.19635	0.18736	0.18866	0.10379	0.19246	
収納	0.25261	0.26871	0.20766	0.23302	0.21172	0.19294	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	0.18146	0.20308	
住宅ロ	0.19784	0.16101	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	
新築マンション	0.24066	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	
新築一戸建て	0.38962	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	
中古マンション	0.16198	0.21447	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	0.12348	
中古一戸建て	0.15042	0.30356	0.21262	0.19635	0.18736	0.18866	0.10379	0.19246	0.16198	0.21447	0.20217	0.18146	0.20308	0.12348	
注文住宅	0.21651	0.18461	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
賃貸	0.18871	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
土地	0.17043	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
不動産	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	

図1. 2019 年のカテゴリの類似度

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 年同士のカテゴリの単語の類似度を示した図を図 2 に示す。

2020															
2020 DIY	DIY	リフォーム	引っ越し	家具イン	暮れ暮れ	収納	住宅ロ	新築マン	新築一戸建	中古マン	中古一戸建	注文住宅	賃貸	土地	不動産
	0.28574	0.22742	0.22206	0.22405	0.27129	0.19026	0.20565	0.20461	0.17389	0.20493	0.17348	0.14137	0.20638	0.18243	0.20728
リフォーム	0.28574	0.26801	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261	0.19261
引っ越し	0.22742	0.26801	0.31827	0.18527	0.25673	0.23674	0.20473	0.20744	0.20123	0.3029	0.35247	0.31353	0.24855	0.20349	0.21826
家具インテリア	0.22206	0.19261	0.18527	0.18527	0.25673	0.23674	0.20473	0.20744	0.20123	0.3029	0.35247	0.31353	0.24855	0.20349	0.21826
暮れ暮れ	0.27129	0.19261	0.18527	0.18527	0.25673	0.23674	0.20473	0.20744	0.20123	0.3029	0.35247	0.31353	0.24855	0.20349	0.21826
収納	0.19026	0.20565	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461
住宅ロ	0.20565	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461	0.20461
新築マンション	0.17389	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348
新築一戸建て	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493	0.17348	0.20493
中古マンション	0.20461	0.3029	0.35247	0.17519	0.26163	0.24479	0.18499	0.38293	0.24676	0.31299	0.27165	0.25192	0.33457	0.20131	0.27205
中古一戸建て	0.17348	0.3029	0.35247	0.17519	0.26163	0.24479	0.18499	0.38293	0.24676	0.31299	0.27165	0.25192	0.33457	0.20131	0.27205
注文住宅	0.35247	0.31353	0.24855	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349
賃貸	0.24855	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349
土地	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349	0.20349
不動産	0.20638	0.18243	0.20638	0.18243	0.20638	0.18243	0.20638	0.18243	0.20638	0.18243	0.20638	0.18243	0.20638	0.18243	0.20638

図 2. 2020 年のカテゴリの類似度

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 年のカテゴリの単語の類似度を示した図を図 3 に示す。

2019															
2019 DIY	DIY	リフォーム	引っ越し	家具イン	暮れ暮れ	収納	住宅ロ	新築マン	新築一戸建	中古マン	中古一戸建	注文住宅	賃貸	土地	不動産
	0.28674	0.17607	0.19958	0.16423	0.25261	0.19784	0.27436	0.24066	0.38962	0.16198	0.15042	0.21651	0.18871	0.17043	0.20308
リフォーム	0.28674	0.29670	0.22021	0.17504	0.26871	0.16101	0.36356	0.26499	0.26054	0.21447	0.30356	0.21746	0.18461	0.17009	0.18146
引っ越し	0.17607	0.29670	0.32872	0.13486	0.20766	0.23302	0.21172	0.19294	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	
家具インテリア	0.19958	0.25261	0.17474	0.16072	0.25301	0.20729	0.24337	0.20274	0.16832	0.19635	0.18736	0.18866	0.10379	0.19246	
暮れ暮れ	0.16423	0.17504	0.13486	0.16072	0.25301	0.20729	0.24337	0.20274	0.16832	0.19635	0.18736	0.18866	0.10379	0.19246	
収納	0.25261	0.26871	0.20766	0.23302	0.21172	0.19294	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	0.18146	0.20308	
住宅ロ	0.19784	0.16101	0.16812	0.25706	0.21262	0.29348	0.12348	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	
新築マンション	0.24066	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	
新築一戸建て	0.38962	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	0.26054	
中古マンション	0.16198	0.21447	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	0.12348	0.20217	0.18146	0.20308	0.12348	
中古一戸建て	0.15042	0.30356	0.21262	0.19635	0.18736	0.18866	0.10379	0.19246	0.16198	0.21447	0.20217	0.18146	0.20308	0.12348	
注文住宅	0.21651	0.18461	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
賃貸	0.18871	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
土地	0.17043	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	0.17009	
不動産	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	0.20308	0.18146	

図 3. 2019、2020 年のカテゴリの類似度

・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) 須舘 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