ISS-P-063

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

階層構造の中に文書が整理分類され保存されている場合、一つの定まった観点からの分類に限られる。例えば、住宅によって最初に分類し、さらにリフォームで分類するなどである。一度分類を決めた後での変更は難しい。このような問題はファイル構造が典型的な例でデータベースの分野で検討されてきた。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。格納されているオブジェクトの分類の場所を変更するのは、クラスタリングによって行われる。しかし、あるカテゴリーの中が二つに分かれている、または、あるカテゴリーの一部が他のカテゴリーの一部と結びついている場合は、文書の部分集合の和集合から最も相応しいカテゴリー名を付して、新しいカテゴリーを作ることになる。

問題は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。カテゴリー名も変わってしまい、理解が容易でなくなる。見つからなければ存在しないと解釈される。本研究では、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して一部を提示することで、全体の概念階層の直観的な理解を助けるシステムを提案する。特に、多量の文書の検索に有効である。従来研究の違いとしては，ニュースの記事ならば、国内ニュース、政治ニュース、国際ニュースなどに分類するのが通常行われている。しかし、コロナウイルスのように特定のトピックになると突然に階層が変わる。利用者の求めている文書が突然に変わる。このために、ベイジアンネットワークを用いて、記事を分類する。その文書は新鮮さがあり、時間でその価値が変わる、例えば、ニュース記事もその一つ。Yahoo の知恵袋のようなサービスもその一つである。

2. 研究方法

Yahoo!知恵袋の「住宅」カテゴリーの中にある「リフォーム」「DIY」といった 15 のカテゴリーの中にある，2019 年 10 月 22 日に取った各 10 個の質問文の全単語を，プログラミング言語「AWK」を用いたプログラミングによって抽出し，計 150 個の単語の名称と数をまとめたリストを作り，それを利用して，15 のカテゴリー同士の相関をとる．そしてその相関関係より，各カテゴリーの関係の強さを調べる．

3. 研究結果

15 個のカテゴリーの相関関係を示した図を図 1 に示す．

リフォーム	DIY	引っ越し	家具	収納	住宅ローン	新築マンション	新築一戸建耐震	中古マンション	中古一戸建	注文住宅	土地	不動産	賃貸
1	0.895347	0.893308	0.84244	0.907462	0.900871	0.892636	0.887216	0.877629	0.918045	0.900035	0.910647	0.916591	0.910879	0.784804
0.895347	1	0.828715	0.840343	0.89078	0.86583525	0.855423586	0.863897	0.866451	0.87657	0.879559	0.884075	0.882145	0.876165	0.764727
0.893308	0.828715	1	0.796901	0.874131	0.87581152	0.851523003	0.869585	0.825639	0.881367	0.872594	0.856524	0.897041	0.888215	0.772043
0.84244	0.840343	0.796901	1	0.843975	0.84561757	0.830448748	0.855314	0.823729	0.834932	0.844143	0.86083	0.834274	0.852005	0.734687
0.907462	0.89078	0.874131	0.843975	1	0.90472902	0.88952033	0.907231	0.877748	0.915084	0.91655	0.897992	0.793561	0.903991	0.778685
0.900871	0.865835	0.875812	0.845618	0.904729	1	0.89596883	0.909821	0.880398	0.910601	0.928602	0.919502	0.910666	0.926029	0.796311
0.892636	0.855424	0.851523	0.830449	0.88952	0.89596883	1	0.896932	0.865467	0.914273	0.906341	0.883229	0.889078	0.898839	0.766277
0.887216	0.863897	0.869585	0.855314	0.907231	0.90982126	0.896932379	1	0.876217	0.909713	0.915153	0.910863	0.902485	0.899644	0.787676
0.877629	0.866451	0.825639	0.823729	0.877748	0.88039817	0.865467178	0.876217	1	0.888287	0.871548	0.897749	0.86553	0.875162	0.75041
0.918045	0.87657	0.881367	0.834932	0.915084	0.91060074	0.914272978	0.909713	0.888287	1	0.917755	0.902889	0.909192	0.912057	0.781148
0.900035	0.879559	0.872594	0.844143	0.91655	0.9286022	0.906340843	0.915153	0.871548	0.917755	1	0.911656	0.912553	0.92495	0.793386
0.910647	0.884075	0.856524	0.86083	0.897992	0.91950248	0.883228536	0.910863	0.897749	0.902889	0.911656	1	0.888842	0.915882	0.787759
0.916591	0.882145	0.897041	0.834274	0.793561	0.91066605	0.889077899	0.902485	0.86553	0.909192	0.912553	0.888842	1	0.911168	0.796238
0.910979	0.876165	0.888215	0.852005	0.903991	0.92602907	0.898838728	0.899644	0.875162	0.912057	0.92495	0.915882	0.911168	1	0.793561
0.784804	0.764727	0.772043	0.734687	0.778685	0.79631123	0.766277088	0.787676	0.75041	0.781148	0.793386	0.787759	0.796238	0.793561	1

図1．15 カテゴリー同士の相関関係

図を見ると，賃貸のカテゴリーは，他のカテゴリーに比べて全てのカテゴリーとの値が小さく，「賃貸」というカテゴリーが他のカテゴリーと関係性が低く，内容が独立しがちであることが分かる．反対に，「住宅ローン」というカテゴリーは他のカテゴリーと全体的に相関係数が高いことから，関係性が高く，内容が他のカテゴリーと似ていることがわかる．

4. 今後の課題

「賃貸」というカテゴリーの関連性が低かったので，その原因を追究したい．また，可能であればデータ数を追加して，より細かい相関係数の値を出したり，クラスタリングを行って，類似度を算出し，別の方法で分析を行いたい．

5. まとめ

「住宅ローン」という単語は，一戸建ての家やマンションを購入する際に多用する言葉なので，それらのカテゴリーとの関連性が高いと思っていたが，実際の相関関係は，中古の一戸建て，マンションは高い値を示したが，新築一戸建て・マンションとはあまり高い値を示さなかった．今回の経験を糧にして，今後の研究に生かせればと思う．

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 閲覧日 2019 年 10 月 22 日

[2] ベイジアンネットワーク入門(1) 須鎗 弘樹 2003 年