カテゴリーに基づくリコメンド機能

東京都立戸山高等学校　SS情報1年　恩田唯衣

研究動機

現在のリコメンド機能にはもののカテゴリーのなかの相関関係だけで収まってしまっているのが現状だ。しかし、実際は映画とファッション、音楽とインテリアなどもののカテゴリー同士にも相関関係があるのではないかと思いこの研究に至った。

研究目的

カテゴリー間の相関関係を測りより広範囲でリコメンドできる数を増やす新たなリコメンド機能（図1）をつくることで、自分たち購買機会を増やし企業の売り上げにも貢献できる、より便利でスムーズなネットショッピングをするため。

実験１：データの相関度をJaccard係数を使って測る

実験内容　kaggleから引用した購買データ**67501979**件を対象に**615**カテゴリーに分類し**Jaccard係数**を用いて相関度を測る。

※　Jaccard係数とは、集合同士の類似度を測る計算式のことで自然言語学にも使用されることが多く、今回はカテゴリーを集合として扱うために使用した。

結果・考察

順位	組み合わせ	相関度
1	メモリ＆マザーボード	0.19
2	充電器系＆マザーボード	0.14
3	メモリ ＆ HDD	0.1
4	ブランケット ＆ 枕	0.09
5	横だな＆キッチンフード	0.09

※上位5つを抜き出したもの。１に近いほど正の相関が強い。

・実験の結果それぞれの組み合わせによってばらつきがでたものの同じカテゴリー内での相関関係が強く出た。
・また同じ結果から同じカテゴリーに所属するものは相関が高いと考えられる。したがって特に相関関係が強いものは0.1～0.2の値で出力されると考えられる。

実験２：Twitterのツイートから興味を算出する

実験内容　Twitter API を用いてTwitterのツイートを解析し名詞のみ取り出して興味を算出する。

自身のツイートを元にpython で MeCab を使い名詞を割り出しテキストマイニングを行う

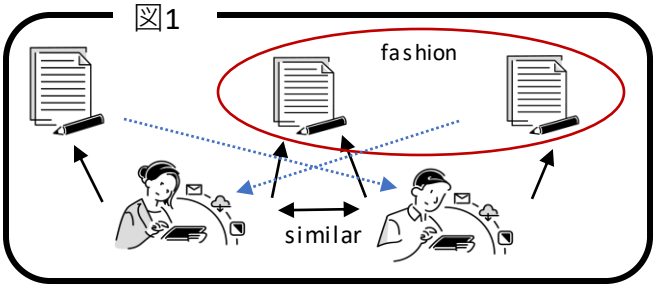
結果・考察

今回はツイート文から一見関係のないものの関係を探るためワードクラウドとして描画した。（図２）

(図２)

頻出単語からユーザーの興味や癖がうかがえると考えられる。この結果の場合「単語」や「インストール」などという単語をよく使用することが分かる。また、ワードクラウドに出現する単語の一致度で人の相性も算出できる可能性がある。

図1



今後の展望

今回の実験１から指標になる数値が明らかになったので次は実験に使う対象をもっと大まかにすることでカテゴリー間の相関関係を明らかにする。実験2では自分が単語だと認識していなかったものまで換算されているのでこの部分の対処、他のユーザーの嗜好も割り出すためターゲットを変えることのできるプログラムに書き換えを行いカテゴリー間の相関を考えていきたい。

参考文献（データ元）https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store
（プログラム参照）https://qiita.com/pontyo4/items/d0910514e1b9f46c1c98