ベイズの定理を用いた個人的嗜好の推定

小堀端　智耶†　　佐藤　寛修†　阿部清彦††
†　関東学院大学工学部情報ネット・メディア工学科　††　関東学院大学理工学部情報学系

1.　はじめに

本研究では、閲覧した Web ページから興味のある単語を抽出し、特定のページに対しどれくらい興味があるのかをベイズの定理を用いて計測する方法について検討する。

2.　興味の度合いの計測

2.1　興味のある単語をまとめたデータベース

閲覧した Web ページのテキストを品詞分解して得られた単語をデータベースに格納する。品詞分解は形態素解析ツールである「Chasen」で行った[1]。

上記の単語から名詞以外の単語、google での検索件数が 1 億以上の単語、tfidf 値[2]を用いて検出した情報量の低い単語を削除した。残った単語を興味のある単語として定めた。

2.2 ベイズの定理を応用した公式

データベース内に格納された単語と特定のページ内に含まれる単語の類似性を計測するため、本研究ではベイズの定理を応用した次の公式を利用した。

$$P(prf \mid page) = \frac{P(page \mid prf)P(prf)}{P(page \mid prf)P(prf) + P(page \mid \overline{prf})P(\overline{prf})} \quad (1)$$

式(1)において P(prf) は興味の有る単語をまとめたデータベース内で特徴的な単語が含まれる割合を示す。P(prf|page) は求めたい値で、特定のページ内の特徴的単語がどれくらい興味の有るデータベース内の特徴的単語と類似しているかを示している。

3　評価実験

3.1　実験方法

閲覧した Web ページを元に興味のある単語を格納したデータベースを作成する。無作為に Web ページを選定し、あらかじめその Web ページに対する興味の有無を被験者が決定しておく。その後データベースとその Web ページに対する類似性を計測し、興味の有無と計測結果との相関性を検討する。

3.2 興味の有無と類似性の計測結果の相関関係

興味がある Web ページの計測結果を表 1 に、興味のない Web ページの計測結果を表 2 に示す。興味のある Web ページと興味のない Web ページの計測結果に大きな差が無く、興味の有無と提案した手法による計算結果の相関関係は見られなかった。データベースに不要な単語が多いことが原因と考えられる。

4.　今後の課題

解析に不要な単語の削除の手法を改良し、興味の有無と計測結果の相関関係を検討する。

表 1　興味のある Web ページの計測結果

ページ名	単語数	計測結果
バイオハザード	1020	0.174
もののけ姫	1052	0.259
ベイズの定理	430	0.372

表 2　興味のない Web ページの計測結果

ページ名	単語数	計測結果
江藤智	877	0.079
手野のスギ	401	0.208
DORBERMAN	510	0.418

参考文献

[1]「Chasen's Wiki」
< http://chasen.naist.jp/hiki/ChaSen/>
(2016/01/05 閲覧)

[2]金田哲他：統計科学のフロンティア 10 言語と心理の統計 ことばと言語の確率モデルによる分析、岩波出版　(2003)