6N-01

共起と出現確率と単語間距離に基づく連語抽出は，ユーザプロファイルの自動名づけを可能にする

望月 啓太† 清水 康太郎† 毛利 瞳† 金道敏樹†

金沢工業大学工学部情報工学科†

1. はじめに

我々は，コンテンツベースの推薦システムに用いられるユーザプロファイルに対する自動名付け機能を実現することを目指している．

推薦システムは利用者にとって有用と思われる対象，情報，または商品などを選び出し，それらを利用者の目的に合わせた形で掲示するために利用される[1]が，現状ユーザプロファイルに対する自動名付けは，

・ユーザの興味の可視化
・ユーザプロファイルの相互作用の利便性向上
・ユーザプロファイルの興味毎の分割
・ユーザの興味の遷移の可視化

につながる有用な機能である(図 1)にも関わらず，残念ながら行えていない．従来研究も，「Paragraph Vector に基づくニュース記事分類のための自動ラベリング」[2]のような限定的な研究があるだけである．

図 1　嬉しさのイメージ

今回，複数の単語を組合せ，連語を探す方法で，ユーザプロファイルに対する自動名付けアルゴリズムに取り組んだので報告する．アルゴリズムの開発にあたり，我々は，情報フィルタ INSOP を用いた論文推薦システムを前提とした．INSOP は，軽量コンパクトなアルゴリズムであり，また，プロファイルの分割が可能な点が魅力的である．推薦システムのコンテンツとして「論文」を選択した理由は，論旨がはっきりしている文章から検証を開始したかったことによる．

2. 提案手法

我々が提案する自動名づけアルゴリズムは，連語インデキシングと代表連語抽出からなる(図 2)．代表連語抽出は，2 つのサブアルゴリズムからなる．第 1 アルゴリズムは，出現確率を用いるものであり，第 2 アルゴリズムは，単語間距離を用いるものである(図 3)．

2.1. 連語インデキシング

論文におけるユーザプロファイルを名付けるためのラベルに適するようなキーとなる概念を表す連語の抽出アルゴリズムを図 4 に示す．

Two connected nouns supported by occurrence probability, co-occurrence and relative location in documents enables automatic labeling of user profiles
Mochizuki Keita† Shimizu Kotaro† Mouri Hitomi† Kindo Toshiki†
Information and Computer Science, the College of Engineering, Kanazawa institute of Technology†

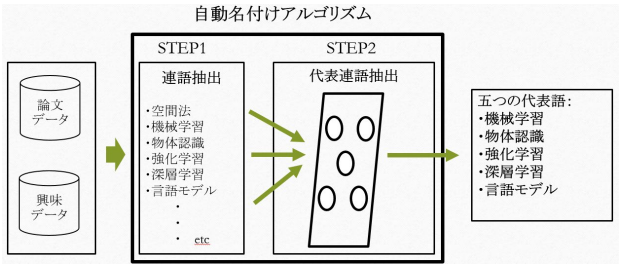

図 2　自動名づけアルゴリズムの概略図

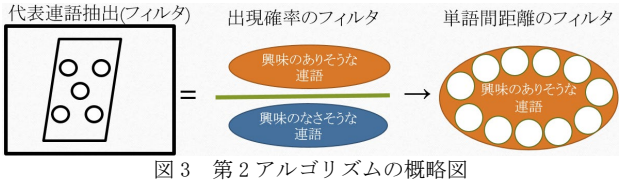

図 3　第 2 アルゴリズムの概略図

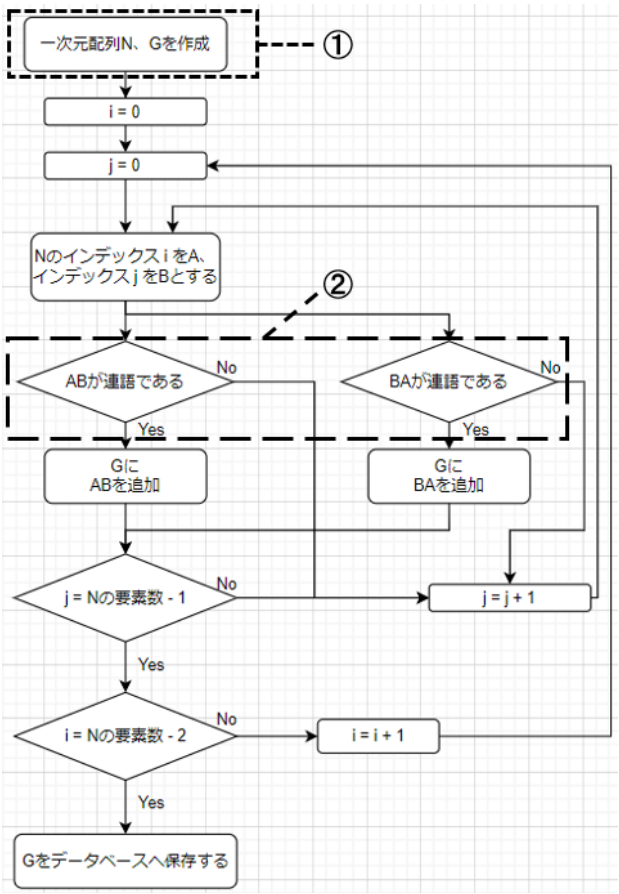

図 4　連語抽出アルゴリズムのフローチャート

図 4 のフローチャートに沿って説明を行う．まず，①内の一次元配列 N は，ある論文の名詞だけで構成された

共起データにおける他の名詞との結びつく数(リンク数)が多い名詞が配列の先頭になるようにデザインされた配列である．また，①内の一次元配列 G は，抽出された連語を保存するための配列である．次に②内で行われる「BA, AB の連語判定」では，Wikipedia を活用している．

図 4 のアルゴリズムは，対象の論文 P(i)， i=1, 2, …全てに対して適用され，各論文に対して，連語インデックス I(P(i))を作成する．

2. 2. 代表連語の抽出

代表語の抽出は 2. 1. で抽出した論文毎の連語インデックスから，ユーザプロファイルのラベルとなる連語を抽出することである．今回，複数の論文を 2 つの集合に分割したとき，一方の集合に現れる確率を用いた方式と，論文中の単語の出現位置を用いる方式を組み合わせてアルゴリズムを構築した．

確率を用いた方式では，任意に 2 分割した集合からその分割に必要な連語を抽出できる．

まず，連語インデックスに含まれる連語の集合 U を考える．次に論文集合 S を二つの集合 S1 と S2 に分割する．U の内，ある連語 a が S1 に出現する確率を求め，Ps(Probability score):$Ps = \Sigma P(S'1)/P(U)|a$として算出する．Ps が 1 に近ければ S1 を代表する連語，0 に近ければ S2 を代表する連語とできる(図 5)．

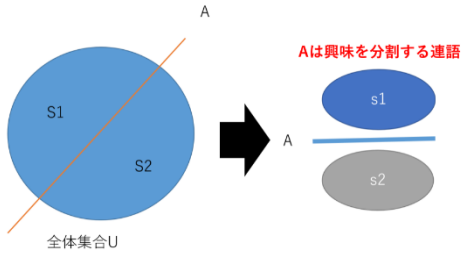

図 5　集合 U を 2 分割する連語の抽出

これにより，あるユーザの興味の有無を含んだユーザプロファイルが与えられたとき，上述の S1 に INSOP を用いてユーザが興味を持つもの，S2 を，興味をを持たないものとすれば，興味の内容に関わらず，ユーザが何に興味があり，興味が無いかを示す連語を可視化できる．

この出現確率を用いたフィルタリング機能だけでは時折，集合 S1 と S2 の関係を表さない連語が残る．その原因は連語の”自然さ”は Wikipedia の検索語であることで保証されているものの，共起しか調べていないため論文の内容を反映しているとは限らないからである．ここでは連語をつくる 2 つの単語が論文中に近い「距離」で現れるかどうかを調べて，距離スコア Ds(Distance score)とする．

名詞 A と名詞 B を組み合わせた連語を AB としたとき論文内の名詞 A，B の間に入っている名詞の数を単語間距離 d と設定し，d がより短い方がラベリングに妥当な連語であると考え Ds を$(Ds = \Sigma(100 - 5d)n, n = 出現回数)$と定義した．

連語の最終的なスコアは，両者の積（score × Ds）となり，その値が大きいもの 5 つを代表連語とする．

3. アルゴリズムの検証と考察

8 人のユーザに自動名づけアルゴリズムに対する満足度を回答して貰い，アンケートによるアルゴリズムの性能評価を行った．

自動名づけアルゴリズムに対するユーザの満足度の評価方法は，まず，8 人のユーザに，情報工学に関する 300 本の論文の内，人工知能に関する論文 10 本程度に対して興味の有無を回答してもらい，ユーザプロファイルを作成する．次に，ユーザプロファイルに対してアルゴリズムを適用し，アルゴリズムの出力である代表連語(5 つの連語)をユーザに見て貰い，その連語がユーザの興味を表現できているのかを 10 段階評価で評価をしてもらった(図 6)．

10 段階評価:ユーザの興味を表現する連語が
「1：全くない　～　10：妥当なものがある」

評価の結果，ユーザの評価の平均が 8. 25 点と高く，アルゴリズムを用いて抽出された連語の妥当性が示された．

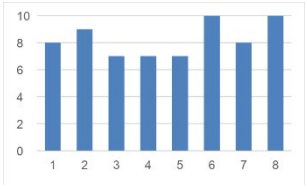

図 6　ユーザの連語に対する評価　{x: UserID, y: points}

UserID4 と 6 のユーザプロファイルからは以下の連語が抽出され，人工知能というジャンルの中でも興味の違いが確認でき，”machine”と”learning”それぞれ一語では意味が広く漠然としてしまうが，”machine learning”にすることで意味が判然となり連語の強みも確認できた(表 1)．

表 1　UserID4 と 6 の代表連語

4	information model	language model	object detection	learning model	model approach
6	reinforcement learning	machine learning	image fusion	user interface	change detection

上記の実験結果から，自動名づけアルゴリズムによるユーザプロファイルのラベリングは，有効であり，ユーザの興味を可視化ができると結論できる．

4. まとめ

ユーザの興味の有無を利用した自動名づけは，ユーザの興味をユーザプロファイルにラベリングすることで可視化できる．また，興味の可視化によりユーザプロファイルを利用したシステムの利便性向上のチャンスにつながることが示された．

5. 謝辞

「距離尺度を用いた意味のある連語の選択」部分において学部 3 年の毛利瞳さんに協力していただきました．心より感謝申し上げます．

参考文献

[1]　Toshiki KINDO et.al. Adaptive Personal Information Filtering System that organizes personal profiles automatically, Proceeding of IJCAI97, 1997.

[2]　斎藤大志，内田理．Paragraph Vector に基づくニュース記事分類のための自動ラベリング，電子情報通信学会総合大会，2017