1－10－3　図書館情報メディア系

SNS・AI

ソーシャルメディアにおけるノイズ除去を目的とした重複文書の検出技術の開発

図書館情報メディア系　准教授　関　洋平
図書館情報メディア研究科 博士前期課程 小邦 将輝

ソーシャルメディアの発展は利便性を向上させた一方，質の低いコンテンツをも生み出すこととなった．本研究では，人工知能を用いた重複文書の検出技術を開発し，ソーシャルメディアにおけるノイズの除去を目指す．

■ 研究背景

■ ソーシャルメディア (ユーザ投稿型コンテンツ) の発展
- 誰もが情報の発信者となる時代へ

日々の出来事を発信：Twitter　Facebook　Instagram

自らの成果物を発信：pixiv　YouTube　TikTok

- 細かな情報を手に入れられるように
  例：交通情報や店舗の営業情報，災害に関する情報
- 情報の取捨選択が容易に

課題

- 多くの情報が手に入れられる一方で，質の低い情報も溢れている

[1] 他のコンテンツと重複した情報

求める情報に辿り着けない可能性

[2] 出典・根拠などが不透明な情報

誤った情報に騙される可能性

本研究の目標: 重複コンテンツを検出し，人々がソーシャルメディアを利用しやすい環境づくりを行う

■ 本研究の目指すところ

文字列が一致した重複文書

→ 重複の検出が**容易**
→ 文字列の一致度を考慮

文字列は異なるが同じ意味合いの重複文書

@hogehoge
留年確定した人のことを『龍角散』って呼ぶのやめてあげてください
投稿日: 2018年6月3日

@hogehuga
留年確定した人のこと「りゅうかくさん」って呼ぶのやめろ
投稿日: 2019年1月28日

→ 重複の検出が**困難**
→ 意味的な一致度を考慮

アプローチ

■ 従来手法
文字n-gramを用いた重複検出
- 文字列が類似した重複文書の検出に成功
- 書き換え，言い換えへの対処に課題
  → 文字列が持つ意味を考慮する必要性

■ 提案手法
深層学習技術を用いた手法で重複検出
- 深層学習技術を用いてできること
  - 大量のテキストデータから，単語が持つ意味の学習
  - 文書が持つ特徴を理解することによる，文書の分類 など

  → 文字列が持つ意味を学習し，重複した文書を検出

多様なドメインに対して有効な重複検出技術の開発

■ 研究成果

[査読付き国際会議論文] Masaki Oguni, Yohei Seki, Risako Shimada and Yu Hirate: Method for Detecting Near-duplicate Recipe Creators Based on Cooking Instructions and Food Images, Proc. of 9th Workshop on Multimedia for Cooking and Eating Activities (CEA 2017), August, 2017, pp.49-54.
[査読付き論文誌] 安藤有生, 関洋平: 市民のツイートを用いた分散表現に基づく地名に対する市民の関心の傾向の可視化. 知能と情報（日本知能情報ファジィ学会誌）, Vol.30, No.6, pp.804-814, 2018年12月.
[招待講演] 島田理紗子, 小邦将輝, 平手勇宇, 関洋平: 重複する料理レシピを判別するためのコーパスの構築, Web インテリジェンスとインタラクション研究会 第6回ステージ発表（採択率 21.7%）, 2018年12月.
[ネットニュース] つくばサイエンスニュース: ネットに投稿された料理レシピの重複を精度良く検出, 2017年8月. (http://www.tsukuba-sci.com/?p=2685,)

その他の研究業績はこちらから
http://cu.slis.tsukuba.ac.jp/achievement.html

Difference

- ソーシャルメディアを対象として，重複コンテンツを検出する研究は少ない
  - 人工知能技術を用いて，多くの人々が用いるソーシャルメディアのUX改善
- 招待講演への採択や報道での取り上げ等，社会における高い評価の獲得

連絡先　①研究内容：関 洋平　yohei@slis.tsukuba.ac.jp
②産学連携：後藤秀利　goto.hidetoshi.fw@un.tsukuba.ac.jp
③事務局：産学連携企画課　tlo@ilc.tsukuba.ac.jp