1S-05

ユーザの興味を利用した学術論文閲覧支援の一手法

岩本 拓実† 金澤 輝一‡ 上野 史† 太田 学†

岡山大学 大学院自然科学研究科† 国立情報学研究所‡

1. はじめに

筆者らは，タブレット端末のカメラ機能を用いて紙媒体の論文のテキストをリアルタイムに検出し，論文中の重要語や Web の関連情報等を表示する学術論文閲覧支援ブラウザを開発している[1]. 学術論文中の未知語を Google 等で検索できるこのブラウザでは，ユーザの興味に沿った検索結果が得られないこともある．本研究では，ユーザが過去の閲覧論文で興味を持った語句を利用し，Google の検索結果を個人化する手法を提案する．

2. 学術論文閲覧支援ブラウザの概要

開発している学術論文閲覧支援ブラウザ[1]の概要について述べる(図 1).

論文の任意のページを画面に写してカメラアイコンを押すと，撮影した静止画が表示される．その後解析アイコンを押すと，重要度の高い語句が青く網掛けされる．また論文画像中の語句をタップすると，下方の情報提示部にその語の重要度や出現頻度といった解析結果(赤)や，Wikipedia(橙), Weblio(緑)，Google(青)での検索結果が表示される．編集アイコンを押し論文画像中の任意のテキストを指でなぞることで，マーカを引くことができる．備忘録生成アイコンをタップすると，閲覧論文の重要語などを保存でき，タブ選択アイコンから保存した情報を選択，閲覧できる．本研究は，ユーザが引いたマーカ中の語句を興味を持った語句として利用し，閲覧中にタップして検索した語句の検索結果を個人化することを目的とする．

3. 検索結果の個人化手法

ユーザが興味を持った語句を利用した個人化手法について述べる．3.1 節では，クラスタリングによる興味語句の決定手法について述べる．3.2 節では，興味語句を用いた検索結果の個人化手法について述べる．

3.1 興味語句の決定

ユーザが閲覧した論文でマーカを引いた語句を興味語句候補とし，興味語句候補のベクトルを，その興味語句候補中の単語の平均ベクトルとして求める．この単語ベクトルは，GloVe(注 1)のそれを用いる．同一単語のベクトルが存在しなくても原形の単語ベクトルが存在すれば，それを代わりに

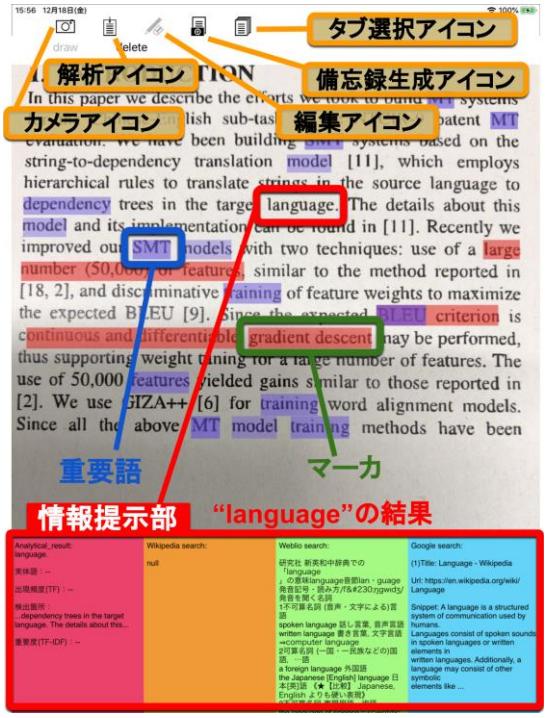

図 1 学術論文閲覧支援ブラウザの画面(論文は[2])

する．またサブワードのベクトルが存在する場合はそれらの平均ベクトルを代わりにする．例えば“Chinese-English”のベクトルは，サブワード“Chinese”と“English”の平均ベクトルとして求まる．原形やサブワードの単語ベクトルも存在しない場合は，Google でその単語を検索し，1 件目の Web ページのタイトルとスニペットに含まれる名詞の平均ベクトルとする．

次に，興味語句候補を k-means 法でクラスタリングする．事前実験から，クラスタの要素数が平均 3 語以上になるようにクラスタ数kを決定する．要素数の多い順に 3 つのクラスタから，各重心に最も近い興味語句候補を 1 つずつ取得し興味語句 1，興味語句 2，興味語句 3 とする．

3.2 興味語句を用いた検索結果の個人化

まず，検索語句と 3.1 節で決定した興味語句を用いて，次のクエリ Q_1〜Q_5 を Google で検索しそれぞれ上位 5 件のタイトルとスニペットを得る．

Q_1 「検索語句」
Q_2 「検索語句 AND 興味語句 1 AND 興味語句 2 AND 興味語句 3」
Q_3 「検索語句 AND 興味語句 1」
Q_4 「検索語句 AND 興味語句 2」
Q_5 「検索語句 AND 興味語句 3」

(注 1)： https://nlp.stanford.edu/projects/glove/

A Method for Supporting of Browsing Scholarly Papers Using Users' Interests

Takumi Iwamoto† Teruhito Kanazawa‡ Fumito Uwano† Manabu Ohta†

† Graduate School of Natural Science and Technology, Okayama University

‡ National Institute of Informatics

次に Q_1〜Q_5 で得られた 25 件の Web ページのベクトルを，検索結果のタイトルとスニペットに含まれる名詞の平均ベクトルとし， Web ページのスコアSを式(1)で求める．

$$S = \sum_{i=0}^{3} (C_i * I_i) \qquad (1)$$

ここで Web ページの内容が，検索語句($i = 0$)および興味語句1,2,3($i = 1,2,3$)とどの程度類似しているかをコサイン類似度C_iで表す．興味度I_iは，その興味語句を含むクラスタの要素数が，興味語句候補の全体数において占める割合である．検索語句の興味度I_0は事前実験より 0.4 とする．最後に，スコアSの高い順に上位 5 件の Web ページを個人化した検索結果として返す．

4. 評価実験

4.1 実験概要

岡山大学の情報系の学部 4 年生 1 名と修士 1 年生 2 名の計 3 名が，学術論文閲覧支援ブラウザで論文を閲覧しながら興味のある語句や指定された語句を検索し，検索語句について興味のある情報が新たに得られるかを評価する．本実験の流れについて述べる．

(1) 被験者ごとに，過去の 3 本以上の閲覧論文から，興味のある語句を合計 17 語収集し興味語句候補とする．
(2) 被験者は，興味のある論文と指定の論文の 2 つの論文を学術論文閲覧支援ブラウザで閲覧する．ここで，興味のある論文は被験者に自由に選ばせる．また指定の論文は，3 名の被験者の興味とそれほど近くないと筆者が考える Soleimaninejadian らの論文[3]とした．
(3) 被験者は，興味のある論文で興味のある語句を 2 つ，指定の論文[3]で指定の語句を 2 つ，計 4 つの語句を検索する．
(4) 検索語句について 3.2 節のクエリ Q_1〜Q_5 の検索結果上位 5 件ずつ，提案手法による個人化結果上位 5 件，計 6 手法の結果がランダムな順にブラウザに表示される．被験者は検索結果の Web ページを閲覧し，検索語句について興味のある情報が新たに得られたかを次の 4 段階で評価する．
    4 点　非常に得られた
    3 点　ある程度得られた
    2 点　あまり得られなかった
    1 点　全く得られなかった

最後に，被験者の評価点とその検索結果の順位に基づき，検索結果の評価値Eを式(2)で求める．ここで，P_iは被験者が付けた i 番目の検索結果の評価点である．

$$E = \sum_{i=1}^{5} (P_i / i) \qquad (2)$$

表 1　被験者 A,B,C の検索語句と興味語句 1,2,3

A	検索語句(興味)	adversarial examples, data augmentation
	興味語句 1,2,3	LSTM, bibliographic elements, tokenizing
B	検索語句(興味)	unsupervised data augmentation, BERT
	興味語句 1,2,3	DBOW, Product2Vec, dmpv
C	検索語句(興味)	GLUE, dropping layers
	興味語句 1,2,3	CoNLL14, WordPiece, JFLEG
	検索語句(指定)	lifelog data, Big Five Traits

表 2　検索結果の評価値E(上:興味，下:指定)

	Q_1	Q_2	Q_3	Q_4	Q_5	提案
被験者 A	3.82	4	4.53	5.23	5.2	4.93
被験者 B	5.43	0	5.51	3.58	2.66	4.61
被験者 C	2.91	0	4.48	3.47	3.05	3.7
平均	4.05	1.33	**4.84**	4.09	3.64	4.41
	Q_1	Q_2	Q_3	Q_4	Q_5	提案
被験者 A	5.13	2.28	2.83	2.95	3.16	2.8
被験者 B	6.13	0	4.95	1.14	2.53	5.07
被験者 C	3.78	0	1.14	2.88	1.14	2.83
平均	**5.01**	0.76	2.98	2.33	2.28	3.56

4.2 実験結果

3 名の被験者 A,B,C の興味で選んだ 2 つの検索語句と興味語句 1,2,3，また指定した 2 つの検索語句を表 1，興味で選んだ語句と指定の語句による検索結果の評価値Eの平均を表 2 に示す．興味のある語句を検索する場合，Q_3，提案手法の評価値が高く，Q_1 はこれらより低かった．一方，被験者があまり知らないと考えて指定した検索語句では Q_1 が有効だった．Q_2，Q_4，Q_5は，Q_1やQ_3，また提案手法より低かった． Google の検索結果が得られない場合は被験者の評価値を 0 点としたため，Q_2 の評価値が最も低かった．

5. まとめ

本研究では，ユーザが閲覧した論文で興味を持った語句をクラスタリングし，要素数の多いクラスタ中の語句で Google の検索結果を個人化する手法を提案した．評価実験より，興味のある語句を検索する際に提案手法が有効だとわかった．

謝辞

本研究の一部は，科学研究費補助金基盤研究(C)(課題番号 18K11989) および 2020 年度国立情報学研究所公募型共同研究(20FC07)の援助による．

参考文献

[1] 岩本拓実，高須淳宏，太田学，“学術論文閲覧支援のための備忘録の設計,” FIT2019，E-014，2019.
[2] J.Ma and S.Matshoukas，“BBN’s Systems for the Chinese-English Sub-task of the NTCIR-9 PatentMT Evaluation,” Proceedings of NTCIR-9 Workshop Meeting，pp. 579-584，2011.
[3] P. Soleimaninejadian *et al.*，“THIR2 at the NTCIR-13 Lifelog-2 Task: Bridging Technology and Psychology through the Lifelog Personality, Mood and Sleep Quality,” Proceedings of NTCIR-13 Workshop Meeting，pp. 20-17，2017.