5C-5

検索ワードに合わせた推薦理由単語の抽出

宮崎　太郎† 　山田　一郎† 　三浦　菊佳† 　松井　淳†
宮崎　勝† 　住吉　英樹† 　加藤　直人† 　田中　英輝†

NHK 放送技術研究所†

1．はじめに

近年，推薦技術が実運用システムで利用されるようになり，NHK においても，放送番組のオンデマンドサービス（NHK オンデマンド [1]: NOD）などで使用されている[1]．NOD では，ユーザが選択した番組を推薦・提示することにより，ユーザにより多くの番組に接触する機会を与えている．しかし，この推薦機能では，番組のタイトルとサムネイル，概要文を提示するのみで，何故その番組が推薦されたのかユーザが瞬時に把握できないこともある．そこで，我々は番組の推薦理由を提示する技術の研究を進めている．

本稿では，電子番組表（EPG）で配信されている番組概要文を対象とし，検索ワードに対して推薦された番組の推薦理由となる単語を概要文中から抽出する手法について述べる．

2．推薦理由単語抽出

推薦理由となる単語は，番組概要文中の重要な単語であり，かつ，ユーザが入力した検索ワードに関連が深い単語である必要がある．そこで，本稿では，

（1） 概要文中での単語重要度
（2） 検索ワードと概要文中に出現する単語の関連度

の２つのスコアを求め，それを融合することで推薦理由となる単語を抽出する．以下で，それぞれのスコアを求める手法について述べる．

2．1 概要文中での単語重要度（CoM）

概要文中での単語重要度を設定する手法として，単語の意味的中心性(Centeredness of Meanings: CoM)を用いた手法を提案する．ここでの意味的中心性とは，その単語が文全体の表す意味にどれだけ近いかを表す．文中の単語重要度を設定する手法としては Okapi BM-25[2]がよく知られている．しかし Okapi BM25 は単語の出現回数のみに基づいた指標であり，単語の意味を考慮することができない．それに対し，本手法では，文全体の意味的な重心に近い単語を抽出することが可能となる．

意味的中心性の尺度として，文中に出現する各名詞について，文中の他の全名詞との間の類似度を計算し，その平均値をスコアとする．スコアが高い単語は，文中に類似した意味を表す単語を多く持つため，文全体が表現している内容を代表する単語であると考える．単語間の類似度には文脈類似度[3]を用いた．

2．2 検索ワードと概要文中に出現する単語の関連度（QWS）

ユーザが入力した検索ワードと，番組概要文中に出現する名詞との間の類似度を計算し，この類似度をスコアとする．これにより，ユーザが検索ワードとして入力した単語と関連が深い単語が抽出できる．以下ではこの手法を Query-Word Similarity(QWS)と呼ぶ．単語間の類似度には CoM の計算と同様，文脈類似度を用いた．

2．3 手法の融合(CoM+QWS)

CoM，QWS はそれぞれ単独で用いても推薦理由単語を抽出できるが，この２つの手法を組み合わせることで，２節で述べた推薦理由となる単語の条件を満たすことができる．

CoM，QWS の結果の融合には，それぞれの手法で得られたスコアの和を用いた．

3．評価実験

3．1 評価実験条件

提案手法の有効性を確認するために，人手で抽出した推薦理由単語を用いた評価実験を行った．

評価実験では，NHK で放送した 1 週間分の番組概要文から作成した検索ワード-番組対を使用した．まず 1 週間分の番組概要文に含まれる名詞から 1 単語抽出し，その単語を検索ワードとして用いた番組検索を行った [2]．このときの検索ワードと最上位で検索された番組の対を，検索ワード-番組対とする．検索ワード-番組対は 100 組作成した．検索ワードを「魚介」とした場合の検索ワード-番組対の例を表 1 に示す．EPG には長さの異なる 2 つの概要文が含まれているが，今回の実験ではその両方を用いた．この検索ワード-番組対を 3 名の被験者に提示し，推薦理由となる単語を 5 単語以内で抽出してもらい，3 名が抽出した単語の和集合を正解データとして用いる．ただし，検索ワードと同一の単語は，今回の評価実験では正解から除外した．

CoM，QWS で用いる文脈類似度の計算には，高度言語情報融合フォーラム（ALAGIN）[3] で公開されている文脈類似語データベースを使用した．

Extracting Keywords corresponding to the query
†NHK Science & Technology Research Laboratories

[1] https://www.nhk-ondemand.jp
[2] 番組検索には NOD で使われている Okapi BM25 を用いた
[3] http://alagin.jp

表１　検索ワード「魚介」の場合の検索ワード-番組対（下線は正解データに含まれる単語）

番組タイトル	番組概要文（short）	番組概要文（long）
きょうの料理ビギナーズ「フライパンごはんで世界の味を楽しもう」	実りの秋は、ごはんが美味しくなる季節。今回はフライパンを使ったごはんものレシピがテーマ。スペイン料理のパエリヤ風や、人気のシンガポールチキンライスを紹介する	実りの秋は、ごはんがおいしくなる季節。今回は、フライパンひとつでできる、世界のごはんものレシピがテーマ。米を香味野菜と一緒に炒めてから、あさりのゆで汁を加えて炊く、魚介と野菜のおいしさをたっぷり味わえるスペイン料理の「魚介のパエリヤ風」の作り方を紹介。そのほか、鶏肉を大きいまま入れて炊く人気の「シンガポールチキンライス」のレシピも紹介する。

３．２　評価実験結果

評価実験では，比較のために Okapi BM25 で番組概要文中に出現する各単語に重みを与え，その降順に出力するベースライン手法の実験も行った．Okapi BM25 は通常，クエリ$Q=\{q_1,q_2,...,q_n\}$に対する文章 D の関連度を与えるものであるが，ここではクエリ Q の代わりに文章 D に現れる各名詞を用いて，名詞ごとの文章 D 中での重要度を求める．名詞 w の文章 D 中での重要度は下式により計算できる．

$$s_{BM25}(D,w)=IDF(w)\frac{f(w,D)(k+1)}{f(w,D)+k(1-b+b\frac{|D|}{avgdl})}$$

*IDF(w)*は単語 w が出現する文章数の逆数，*f(w,D)*は文章 D 中の単語 w の出現回数，*avgdl* は番組概要文の平均単語数，|*D*|は D 中の単語数，*k*，*b* は定数である．今回は k=2.0，b=0.75 とした．*IDF* と *avgdl* の計算には評価実験に用いた 1 週間分の番組概要文を用いた．

図 1 に評価実験結果を示す．横軸が各手法により出力した単語数を表し，縦軸が正解データと比較した場合の適合率を表す．CoM，QWS とも，単独で用いた場合でベースライン手法より良好な結果が得られた．また，CoM と QWS を融合した CoM+QWS では，それぞれの手法を単独で用いる場合よりも性能が向上した．CoM+QWS は，ベースラインと比較して，上位 1 単語を抽出する場合で 21.0 ポイント，上位 3 単語を抽出する場合で 13.1 ポイントの性能向上が得られた．

４．考察

表 2 に，表 1 のデータの場合に各手法から得られた単語の上位 5 単語までを示す．下線の単語は正解データに含まれているものを示す．

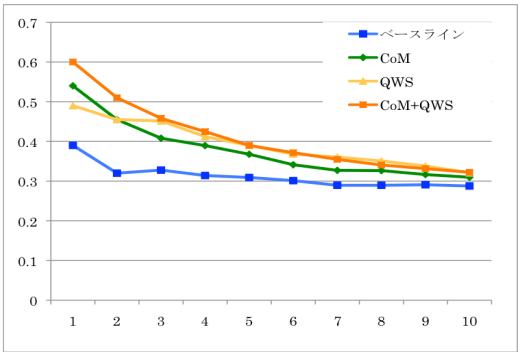

図１　評価実験結果

表２　各手法で得られた単語（上位５単語）

手法	抽出した単語
ベースライン	フライパン，パエリヤ，チキンライス，シンガポール，ごはん
CoM	ごはん，チキンライス，料理，野菜，レシピ
QWS	鶏肉，あさり，野菜，チキンライス，味
CoM+QWS	鶏肉，野菜，チキンライス，あさり，ごはん

CoM による手法は，「料理」のように，他の文にも多く出現するためにベースライン手法では抽出ができなかった単語も抽出することができた．それに対し，ベースライン手法は，「フライパン」や「シンガポール」のように，他の文にはあまり出現しないがこの概要文中では特に重要ではない単語を抽出してしまう場合が多かった．逆に「パエリヤ」のようなやや専門的な用語については，CoM ではうまく抽出できなかった．これは，文脈類似度の計算用のモデルの作成に用いた学習データにあまり出現しなかった単語について，文脈類似度が精度よく求められなかったことが原因と考えられる．

QWS では，「鶏肉」や「あさり」のように，検索ワードである「魚介」と意味が近い単語を抽出することができたが，「味」のように，検索ワードとの関連は深いが，概要文中ではそれほど重要ではない単語を抽出してしまうことが多かった．そこで，CoM と QWS を融合することで，２つの手法からともに上位で得られた単語を抽出することができ，良好な結果となった．

５．まとめ

本稿では，番組検索の際に，その番組が推薦された理由となる単語を抽出する手法を提案した．

評価実験では，単語抽出手法の上位３単語を出力する場合では，ベースラインとなる TF-IDF を用いた単語抽出との比較で，提案手法では 13.1 ポイントの性能向上が得られた．

今後の課題として，ユーザプロファイルを用いた推薦手法への本手法の適用や，抽出した単語を用いて推薦理由を自然文で作成することが挙げられる．

参考文献

[1] Jun Goto, Hideki Sumiyoshi, Masaru Miyazaki, Hideki Tanaka, Masahiro Shibata, Akiko Aizawa, “ Relevant TV Program Retrieval using Broadcast Summaries,” in Proceedings of the 14th ACM International Conference of Intelligent User Interfaces (IUI 2010).

[2] Stephen E. Robertson and Steve Walker, “ Okapi/Keenbow at TREC-8” In Proceedings of the 5th International Conference on Language Resources and Evaluation (2008).

[3] 風間淳一, Stijin De Saeger, 鳥澤健太郎, 村田真樹, “ 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成”，言語処理学会第 15 回年次大会 (2009).