単語間関係辞書を用いたテレビ番組検索

宮﨑 太郎　山田 一郎　三浦 菊佳　宮崎 勝　松井 淳　後藤 淳　住吉 英樹
NHK 放送技術研究所
{miyazaki.t-jw, yamada.i-hy, miura.k-ig, miyazaki.m-fk,
matsui.a-hk, goto.j-fw, sumiyoshi.h-di}@nhk.or.jp

1 はじめに

近年，インターネットを通じて動画の配信をするサービスが，多くのユーザに利用されるようになった．NHK でも，NHK オンデマンド[1]という動画配信サービスを提供しており，常時 5,000 本程度のコンテンツを配信している．このようなサービスでは，ユーザが見たい番組を探す際に，番組の検索技術が有用となる．

NHK オンデマンドの現状の検索機能では，番組タイトルや概要文の単語表記を手がかりとしてユーザが見たい番組を検索する．しかし，コンテンツの数が現状では数千規模と比較的少ないことと，検索の手がかりとなる概要文の文字数制限から，検索できる番組数が少ないという問題がある．このため，例えば「ガーデニング」や「高血圧」などの一般的な単語で検索した場合でも検索結果が 0 件となる場合がある[2]．また，「発熱」で検索するとドラマばかりが出力されるなど，ユーザが求める番組が出力に含まれない場合もある．

このような問題を解決するために，我々は単語間の関係を記した辞書（以後，単語間関係辞書）を用いた検索手法の研究を進めている [1, 2, 3]．本稿では，単語間関係辞書をたどり，検索する手法を提案する．これにより，以下の特徴と効果を持つ．

- 上位下位や原因結果などの多様な関係を用いてクエリが拡張されるため，検索対象の番組が少ない場合でも検索結果が出力可能となる
- より多くの関連単語が出現する番組に高いスコアを与えることで，クエリが主題に近い番組を上位に出力する

今回，提案手法について，実際のサービスに近い形で評価実験を実施し，検索結果の出力数が大幅に増加するとともに，検索の精度は okapi BM25[4] を用いたベースライン手法と同程度であることを確認した．

[1]https://www.nhk-ondemand.jp
[2]2015 年 12 月 18 日時点

2 提案手法

提案手法では，まずユーザが入力したクエリに対し，単語間関係辞書によりクエリと関連のある単語の集合を獲得し，クエリと単語の関係の強さを表す重みを計算する．次に，クエリと関連のある単語集合と重みを用い，クエリから番組へのスコアを計算し，降順に並べたものを検索結果としてユーザに提示する．

以下では，まず本稿で用いる単語間関係辞書について説明し，その後に提案手法の詳細について述べる．

2.1 単語間関係辞書の作成

単語間関係辞書は ALAGIN フォーラム[3]が公開している意味的関係抽出サービス [5] と上位下位関係抽出ツール [6] により取得した単語間の関係からノイズを除去 [7] することにより作成した．作成した単語間関係辞書の例を表 1 に，単語間関係辞書のデータ量を表 2 に示す．このサービスでは，約 6 億の web ページを解析して意味的な関係を持つ単語を出力するが，その関係の強さは出力されない．そこで，クエリと単語の関係の強さについては，外部のデータを用いて計算する必要がある．

表 1: 単語間関係辞書の例

単語 1	単語 2	関係名
コレステロール	高血圧	原因結果
ビール	大麦	材料
音楽	ゴスペル	上位下位

表 2: 単語間関係辞書のデータ量

データ種別	データ量
全関係数	10,125,818
異なり語彙数	3,466,416
異なり関係名数	94,191

[3]http://alagin.jp

2.2 クエリから番組へのスコア計算方法

クエリから番組へのスコアは以下の (a) から (d) の条件を満たすときに高くなるように設定した．

(a) クエリと番組の距離が近い (図 1-(a))

(b) クエリと番組をつなぐパスが多い (図 1-(b))

(c) 経由する単語間の類似度が高い (図 1-(c))

(d) 経由するノードにつながる単語数が少ない (図 1-(d))

以下でスコアの計算方法の詳細を述べる．

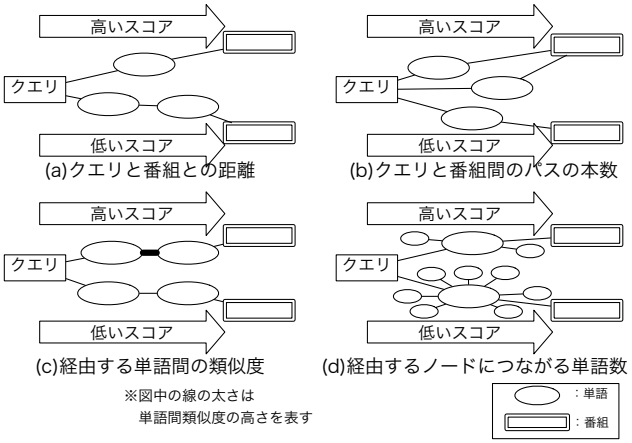

図 1: 4 つの条件によるスコア設定

2.2.1 単語間関係辞書の関係の重み計算

まず，ユーザから入力されたクエリを始点として単語間関係辞書をたどり，各単語との関係の重みを計算する．

クエリ q 中の単語 c_0 から単語 $c_1, c_2, \cdots, c_{n-1}$ を経由して c_n に至るパスの重み $w_{path}(q, c_n; c_0, c_1, \cdots, c_{n-1})$ は以下のように計算する．

$$w_{path}(q, c_n; c_0, c_1, \cdots, c_{n-1}) = \prod_{i=0}^{n-1} w_{edge}(c_i, c_{i+1}) \quad (1)$$

ここで，$w_{edge}(c_i, c_{i+1})$ は単語間関係辞書で直接つながる 2 単語間の重みを表し，その値は 0 から 1 の範囲である．この相乗をとることにより条件 (a) を表す．$w_{edge}(c_i, c_{i+1})$ は以下のように計算する．

$$w_{edge}(c_i, c_{i+1}) = \sqrt[3]{sim(c_i, c_{i+1})^2 \frac{1}{log(\max(|c_i|, |c_{i+1}|))}} \quad (2)$$

$sim(c_i, c_{i+1})$ は単語 c_i と c_{i+1} の間の類似度で，本稿では Word2Vec[8] により求めた単語の分散表現のコサイン類似度を用いた．また，$|c_i|$ と $|c_{i+1}|$ はそれぞれ単語間関係辞書中で単語 c_i，c_{i+1} とつながる単語の数である．(2) 式中の $sim(c_i, c_{i+1})$ が条件 (c)，$1/log(\max(|c_i|, |c_{i+1}|))$ が条件 (d) にそれぞれ対応する．なお，本稿では単語間関係辞書をたどる際の計算量を考慮し，$n < 3$ の条件を与えた．

単語間関係辞書では，単語から単語へのパスは複数存在する場合がある．最終的に用いる単語 c の重み $w_{dict}(q, c)$ には，クエリ q から単語 c に至る全てのパスのうちで重みが最大のものと，単語 c 自体の重みである $IDF(c)$ との積をとったものを用いる．

$$w_{dict}(q, c) = IDF(c) \cdot \max_{r \in R}(w_{path}(q, c; r)) \quad (3)$$

$$IDF(c) = log \frac{|D|}{|\{d : c \in d\}|} \quad (4)$$

ここで，R はクエリ q から単語 c に至る全てのパスの集合である．なお，$IDF(c)$ は外部のテキストから計算し，$|D|$ は総文書数，$|\{d : c \in d\}|$ は単語 c を含む文書数である．

$w_{dict}(q, c) > 0$ となる c の集合がクエリと関係のある単語の集合であり，それぞれの単語の重みが $w_{dict}(q, c)$ である．これらを用い，クエリから番組へのスコアを計算する．

2.2.2 クエリから番組へのスコア計算

クエリから番組へのスコア $score(q, P)$ は，番組概要文中に出現する全単語について，単語間関係辞書をたどった際の重み $w_{dict}(q, c)$ と，番組概要文中での単語の重要度 $w_{prog}(c, P)$ を用い，以下のように計算する．

$$score(q, P) = \sum_{c \in W} w_{dict}(q, c) \frac{w_{prog}(c, P)}{log|W|} \quad (5)$$

ここで，W は番組 P に出現する全単語の集合で，$|W|$ は番組 P の概要文に出現する単語の総数である．クエリ q と番組 P に含まれる単語 c とをつなぐパスが多いほど高い値をとり，条件 (b) に対応する．なお，番組概要文中での単語の重要度は単語 c の分散表現 $\boldsymbol{c}$ と，番組 P に出現する全単語の分散表現の和 $\boldsymbol{P}$ のコサイン類似度により計算する．

$$w_{prog}(c, P) = \frac{\boldsymbol{c} \cdot \boldsymbol{P}}{|\boldsymbol{c}| \cdot |\boldsymbol{P}|} \quad (6)$$

以上で得られたスコア $score(q, P)$ の降順に番組を並べて提示することで，検索結果の出力とする．

3 評価実験

提案手法の性能を評価するための評価実験を行った．評価データにはNHKオンデマンドのテキストを用い，また，被験者自身が作成したクエリにより番組検索を行い，その結果を評価することで，実運用に近い環境での評価とした．以下で，評価実験について述べる．

3.1 実験条件

評価データには2015年8月にNHKオンデマンドで公開されていた5,066番組に付与されたテキストを用いた．テキストは番組のタイトル，80文字以内の短い概要文，200文字以内の長い概要文で構成されており，検索にはこれらの全てを使用した．

検索の際に使用するクエリは被験者自身が作成した．そのクエリから提案手法と，3.2節で述べるベースライン手法との2つの手法で検索を実行し，検索結果の上位10位までを被験者に提示する[4]．なお，手法とクエリによっては検索結果が10件に満たない場合もあるが，その場合は出力された全番組を提示する．被験者は提示された出力に対し，表3のように評価をつける．被験者は6名で，合計111のクエリで評価した．

形態素解析にはMeCab[9]を用いた．分散表現とIDFの計算にはwikipediaの2015年4月時点のデータを用いた．

表3: 評価値と内容

評価値	内容
4	関係がある
3	やや関係がある
2	あまり関係がない
1	関係がない

3.2 ベースライン手法

提案手法との比較に用いるベースライン手法として，単語の重み付けに一般に用いられるokapi BM25を使用した．評価データの番組タイトル，概要文に出現する各単語に対し，okapi BM25で重み付けする．このスコアを降順に並べたものを検索結果の出力とした．okapi BM25に用いるIDFの計算には，提案手法同様にwikipediaのデータを用いた．

3.3 実験結果

提案手法，ベースライン手法のそれぞれから得られた検索結果の出力数を図2に示す．出力数は最大が10であるのに対し，ベースライン手法では平均で6.77，提案手法では平均で9.78となった．提案手法では，評価実験に用いた全111個のクエリのうちの95%にあたる106個について，最大である10件の検索結果を出力した．また，検索結果が0件となるクエリがベースライン手法では19個あったのに対し，提案手法では1個であった．

ベースライン手法で検索結果が0件となった各クエリについて，提案手法で出力された番組の最大の評価値を表4に示す．提案手法では，それらのうちの約半数で評価値が3以上と評価された番組を出力できた．

それぞれの手法で出力した番組への評価値の平均[5]を表5に示す．表では，どちらかの手法で検索結果が0件となったクエリは除いている．出力された番組の評価値については，2つの手法でほぼ同等であった．

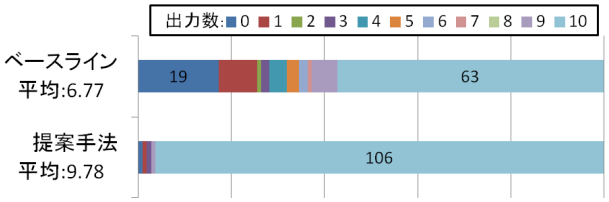

図2: 手法ごとの検索結果出力数

表4: ベースライン手法で検索結果が0件だったクエリに対する，提案手法の出力番組の最大の評価値

評価値	クエリ数
4	6
3	3
2	6
1	4

表5: 評価値の平均

	ベースライン	提案手法
評価値の平均	2.75	2.78

3.4 考察

今回の評価データのように比較的小規模なデータベースを対象とした検索では，ベースライン手法のような単語表記を手がかりとした手法の検索結果の出力数が少なくなりやすい．それに対し，提案手法では，単語間関係辞書をたどることで多くの単語を使って検索できるため，出力数を増加できた．また，ベースライン手法では検索結果を1件も出力できないクエリに対して，提案手法では約半数で有用な番組が出力できた．これより，提案手法は比較的小さなデータベースでの検索では特に有効であると言える．一方で，ベー

[4]被験者にはどちら手法から出力されたのかを伏せて提示した．

[5]2つの手法では検索結果の出力の数が違う場合があるので，その場合は少ない方に合わせた出力数で算出した．

スライン手法で十分な出力数が得られる大規模なデータベースでの有効性は，今後の検証が必要である．

今回用いた評価データは約 6 割 (3,020 番組) がドラマであるなど，偏りがあった．ドラマの概要文には幅広い単語が出現することから，検索結果にはドラマが出力されやすい．しかし，被験者が入力したクエリの大半がドラマを期待しておらず，両手法から多く出力されたドラマの大部分が低い評価値となった．それが評価値に差が見られなかった一因と考えられる．偏りの少ないデータでの評価も必要である．

提案手法はテキストのみを用いた検索手法であり，コールドスタートの問題は発生しない．そのため，放送前のテレビ番組を検索するなど，事前に評価を集めることができない場合にも使用できる．また，テレビ番組に特化した手法ではないため，テキストを用いた検索では幅広い応用が考えられる．

4 関連研究

近年のコンテンツ推薦の研究は，ユーザによる評価を基にした手法が中心である．NetFlix Prize で優秀な成績を挙げた Koren らの手法 [10] や，Amazon などのオンラインストアで多く用いられている協調フィルタリング [11] がその代表である．これらの手法はコンテンツからコンテンツへの関連度の計算に用いられるものであり，キーワード検索への応用が難しい．

単語表記が一致しない場合でも検索を可能とするためのクエリ拡張の手法としては，海野らの手法 [12] などが挙げられる．この手法では，2 言語間の対訳コーパスを用いて言い換え表現を獲得し，言い換え確率に基づく言語モデルを用いて検索性能を向上したが，言い換え表現以外へのクエリ拡張は難しい．

テレビ番組を対象とした検索手法としては，Goto らの手法 [13] と，Yamada らの手法 [1] が挙げられる．Goto らは，okapi BM25 を n-gram に拡張し，さらに複合語や固有名詞など，重要な役割を持つ単語に高い重みを与えることで，精度よく検索できる．しかし，検索結果の出力数を増加することはできない．Yamada らは，単語間関係辞書を用いてランダムウォークにより関連度を算出することで，高い精度で関連度が求められ，また検索結果の出力数を増加できる．しかし，ランダムウォークは計算コストが高いため，キーワード検索に用いる場合には計算時間の問題が生じる．

5 おわりに

本稿ではテレビ番組を対象に，クエリを始点として単語間関係辞書をたどる検索手法を提案した．

評価実験では，検索結果の出力数の平均が，okapi BM25 を用いたベースライン手法では 6.77，提案手法では 9.78 と，出力数が大幅に増加することを確認した．ベースライン手法では検索結果が 1 件も出力できなかったクエリに対しても，提案手法では約半数で有用な番組が出力できた．これらのことから，検索対象が数千規模の比較的小規模なデータベースでの検索について，提案手法の有効性を確認できた．

一方，検索結果の評価値については，今回の実験条件ではベースライン手法との差が見られなかった．今後，大規模なデータベースや，偏りが少ないデータベースを用い，実験条件を変えて評価を進める．また，番組から番組への関連度の計算への応用を検討する．

参考文献

[1] Ichiro Yamada, et al., “Measuring the Similarity between TV Programs using Semantic Relation,” in Proceedings of COLING 2012, pp.2945–2595 (2012).

[2] 三浦菊佳ほか, “単語間の意味的関係を用いた番組リンク生成,” 電子情報通信学会研究会, NLC2014-42, pp. 105–110 (2014).

[3] Masaru Miyazaki, et al., “My Health Dictionary - Study on Web Service using Program Information Data-hub as Linked Open Data,” in CEUR workshop Proceedings, vol. 1486 (2015).

[4] S. E. Robertson and S. Walker, “Okapi / Keenbow at TREC-8,” in Proceedings of TREC-8, pp.151–162 (1999).

[5] Stijn De Seager, et al., “Large Scale Relation Acquisition using Class Dependent Patterns,” in Proceedings of IEEE International Conference on Data Mining, pp. 764–769 (2009).

[6] 隅田飛鳥ほか, “Wikipedia の記事構造からの上位下位関係抽出,” 自然言語処理, Vol. 16(3), pp. 3–24 (2009).

[7] 山田一郎ほか, “Wikipedia を利用した上位下位関係の詳細化,” 自然言語処理, vol. 19(1), pp. 3–23 (2012).

[8] Tomas Mikolov, et al., “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781 (2013).

[9] Taku Kudo, et al., “Applying Conditional Random Fields to Japanese Morphological Analysis,” in Proceedings of EMNLP 2004, pp. 230–237 (2004).

[10] Yehuda Koren, et al., “Matrix Factorization Techniques for Recommender Systems,” IEEE Computer, pp. 42–49 (2009).

[11] Greg Linden, et al., “Amazon.com Recommendations,” IEEE Internet Comput., vol. 7, no. 1, pp. 76–80 (2003).

[12] 海野裕也ほか,“自動獲得された言い換え表現を使った情報検索,” 言語処理学会第 14 回年次大会, pp. 123–126 (2008).

[13] Jun Goto, et al., “Relevant TV Program Retrieval using Broadcast Summaries,” in Proceedings of the 14th ACM International Conference on Intelligent User Interface, pp.411–412 (2010).