Vol.2012-NL-207 No.12
2012/7/27

ランダムウォークを利用した番組類似性評価

山田一郎† 宮崎勝† 住吉英樹† 古宮弘智† 田中英輝†

テレビ番組を配信するオンデマンドサービスでは、ユーザへの番組推薦機能が重要となる。そこで本稿では、テレビの電子番組表(EPG)に含まれる番組概要文を利用して効果的な関連番組を推薦するための番組類似性評価手法を提案する。提案手法では、番組概要文に含まれる単語をノードとし、Web から取り出した単語間の関係（因果関係や上位下位関係など）で各ノードを結合したグラフ構造を生成する。このグラフ構造のノード間の到達可能性をランダムウォークにより評価する事によって、番組間の類似性を評価する。実験では NHK オンデマンドで実際に提示された関連番組を対象とした類似性評価を行い、提案手法の結果は従来手法と比較して人手による類似性評価結果に近いことを示す。

Calculating Similarity between TV Programs Using a Random Walk Algorithm

ICHIRO YAMADA† MASARU MIYAZAKI† HIDEKI SUMIYOSHI†
HIRONORI FURUMIYA† HIDEKI TANAKA†

This paper presents a novel method of calculating similarity between TV programs by using summaries as a part of Electronic Program Guide (EPG). Most previous methods used statistics such as *tf-idf* based cosine measure of word vectors, whose words are appeared in the summaries. However these approaches were not effective for calculating similarity between TV programs because broadcast summaries are too short to obtain reliable statistics. Our method generates a graph structures whose vertexes are TV programs and words. These words are connected by word relations which are extracted from Web automatically. Similarity between two TV programs is calculated based on the relativeness of two TV program's vertexes in the graph structure by using a random walk algorithm. Through experiments, our method showed effectiveness of calculating similarities between two TV programs compared with the baseline approaches.

1. はじめに

放送局では、ブロードバンド回線を通じて放送したテレビ番組を配信するサービスを行っている。今年 4 月からは、民放キー局の番組を対象とした「もっと TV[1]」というサービスが開始され、今後、テレビ番組に対する VOD サービスの拡大が期待できる。このようなテレビ番組を配信する VOD サービスでは、ユーザへの番組推薦機能が重要となる。NHK が提供している「NHK オンデマンド[2]」という動画配信サービスには、選択した番組に関連する番組を提示する機能があり、ユーザは提示された関連番組を通して嗜好に合った番組を芋づる式に探し出すことができる。しかし、ユーザが選択する関連番組はその提示順位に大きく依存し、たとえ関連番組として提示されても下位の順位にある番組は選択されにくい(図 1)。より良い VOD サービスを実現するためには、どのような関連番組を提示するか、また、提示する関連番組をどのようにランキングするかが重要となる。

これまで我々は、テレビの電子番組表(EPG)中の番組概要文に現れる単語を手掛かりとして、関連番組を提示する手法を提案している[1]。しかし、この手法では単語表記の

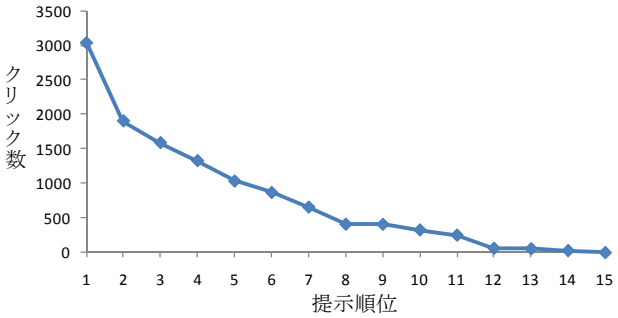


図 1. 関連番組の提示順位に対するユーザの選択数
(2010 年 9 月～2011 年 5 月における
NHK オンデマンドのクリック数)

完全一致を手掛かりとしているため、類似性評価の際に問題が生じることがある。例えば、以下の 2 つの文は同じような内容を述べているが、表記が一致する名詞や動詞が出現しないため類似していないと判定されてしまう。

[文 1] 生活習慣病の治療のポイントを伝える。
[文 2] 高血圧を改善するための減塩や薬物療法の進め方など、視聴者からの疑問に答える。

もし、「高血圧」が「生活習慣病」の一つであり、「減塩」や「薬物療法」が「治療」の一つであることが分かれば、この 2 つの文は類似していると判断できるかもしれない。

† NHK 放送技術研究所
Japan Broadcasting Corporation, Science and Technology Research Laboratory
[1] http://mottotv.jp/
[2] http://www.nhk-ondemand.jp/

そこで本稿では、単語間の関係(因果関係や上位下位関係など)を Web から抽出し、この単語間の関係を利用した文書間の類似性評価手法を提案する。提案手法では、2 つの文書に出現する名詞間を自動獲得した関係で結ぶことによりグラフ構造を生成し、ランダムウォークを用いてグラフ上のノード間の関連度スコアを算出することにより、2 つの文書の類似性を評価する。実験では NHK オンデマンドで実際に提示された関連番組を対象とした類似性評価を行い、提案手法の結果は従来手法と比較して人手による類似性評価結果に近いことを示す。さらに、利用する関係の種類や関係数による類似性評価結果の違いを考察する。

2. 関連研究

従来、ユーザが選択した商品の関連商品を提示する推薦システムの研究は盛んに取り組まれている。これらの推薦システムで用いられるアプローチは、対象コンテンツの内容に基づいてユーザに商品を推薦する内容ベースフィルタリングと、類似する嗜好を持つユーザの情報を利用して商品を推薦する協調フィルタリングの 2 種類に分類できる。

内容ベースフィルタリングは従来の情報検索技術を利用するもので、コンテンツやユーザプロファイル、さらにはユーザが入力するクエリー間の類似度をもとに、商品をユーザに推薦する。Goto らは、放送番組の概要文に含まれる単語に対して Okapi-BM25[2]によって重み付けを行うことにより複数の番組間の類似性を評価し、ユーザが一つの番組を選択した際に、類似する別の番組を推薦する手法を提案している[1]。奥らは、時間帯や天気、同伴者、予算などのユーザの状況に応じて変化するユーザプロファイルをモデル化し、状況依存型の推薦システムを提案している[3]。

近年、推薦システムとして協調フィルタリングを利用した手法が実用システムとして効果を上げている。Amazon では、ユーザの購入履歴を基にアイテム間の類似性を評価する Item-to-Item Collaborative filtering を行い、「この商品をチェックした人はこんな商品もチェックしています」といった商品の推薦を行っている[4]。Koren らは、ユーザと商品を行と列としてユーザの嗜好を数値化した行列を作り、この行列を特異値分解して欠損値(ユーザの嗜好が不明な商品に対する嗜好の評価値)を推定する Matrix Factorization を利用したアルゴリズムを提案し、NetFlix が主催したコンテストで優秀な成績を挙げている[5]。

本稿では内容ベースフィルタリングを拡張し、商品に関するテキストに含まれる単語表記が一致しなくても適切に商品間の類似性を評価する手法を提案する。本手法は協調フィルタリングにおいても、ユーザの商品に対する嗜好を数値化した行列を作る処理などで有用である。Matrix Factorization を利用したアルゴリズムにおいても、その予測モデルに取り入れることができる。また、内容ベースフィルタリングと協調フィルタリングの両方を用いるハイブリッドなアプローチも提案されており[6]、本手法はこのようなアプローチにおける精度向上にも繋がると考えられる。

3. 単語間関係の獲得

本章では、提案する文書間の類似性評価手法で利用する単語間の関係の獲得手法について説明する。単語間の関係として以下の 4 種類を利用する。

上位下位	例：「生活習慣病／高血圧」
原因結果	例：「かび／ニオイ」
名物	例：「六義園／しだれ桜」
材料	例：「ビール／麦」

これらの関係は、ユーザの興味を惹く関連番組検索に有用と期待でき、実験的に選択している。例えば「六義園」に関する番組に関心のあるユーザは、他の「しだれ桜」の名所に関する番組にも興味を持つ可能性があるため「名物」の関係は関連番組検索に有用と考えられる。

また、「対象(entity)／属性名(attribute)／属性値(value)」という 3 項からなる属性関係を利用して、上記 4 種類以外の関係も獲得する。属性関係は、「対象」と「属性値」に該当する 2 つの単語が「属性名」という関係を持つと解釈できる。例えば「七人の侍／キャスト／三船敏郎」という属性関係では、「七人の侍」と「三船敏郎」の関係は「キャスト」と判断できる。

これらの関係獲得のために ALAGIN Forum[3] で公開されているツールを利用する。以下に、このツールで使われている各関係獲得手法の概略と獲得した関係の精度の調査結果を記す。

3.1 Wikipedia からの関係獲得

単語の上位下位と属性関係の獲得では、Wikipedia を利用した上位下位関係抽出ツール[4] を利用する。日本語 Wikipedia には現在約 80 万本の記事が存在する。上位下位関係抽出ツールは、この記事の階層的なレイアウト構造やカテゴリタグ、さらには記事中の第一文(見出し語の定義文)を利用する[7]。例えば、「生活習慣病」を記事タイトルとするページには「高血圧」、「糖尿病」という単語が見出し語として存在する。Wikipedia の見出し語となっている単語(例えば「高血圧」)が、レイアウト構造で上位にある記事タイトルや見出し語(例えば「生活習慣病」)と上位下位関係にあるか否かを教師有りの機械学習で判定することにより、大量の上位下位関係を高精度に獲得することができる。

また、上位下位関係抽出ツールは記事の階層的なレイアウト構造を利用して属性関係を獲得することもできる。例えば、「七人の侍」を記事タイトルとするページには、その

[3] http://alaginrc.nict.go.jp/
[4] http://alaginrc.nict.go.jp/hyponymy/index.html

見出しに「キャスト」と「三船敏郎」が存在する。記事タイトル(例えば「七人の侍」)を「対象」、上位下位関係があると判定された単語対(例えば「キャスト」と「三船敏郎」)を「属性名」と「属性値」とすることにより、大量に属性関係を獲得することができる[8]。

3.2 Web テキストからの関係獲得

原因結果、名物、材料の関係の獲得では、ALAGIN フォーラムで公開されている意味的関係抽出サービス[9]を利用する。このサービスは、少数のシードとなる単語ペアを入力として与えることにより、シードの単語ペアに類似した関係を持つ大量の単語ペアを約 6 億の Web ページに含まれるテキストから獲得できる。この処理では、単語の意味クラスと文脈パターンを利用する。例えば、「X が Y の原因となる」という文脈パターンに出現する X と Y には原因結果の関係がある可能性が高い。さらに、X と Y が属する意味クラスを制限し、例えば、シードの単語ペアに「細菌／悪臭」がある場合は、「細菌」と同じ意味クラスに属する単語と「悪臭」と同じ意味クラスに属する単語のペア(例えば「カビ／におい」)にも同様の関係がある可能性が高い。この意味的関係抽出サービスでは、分布類似度を用いた手法[10]によって単語の意味クラスを自動獲得している。

意味的関係抽出サービスにより獲得される関係には明らかな誤りや曖昧な関係が含まれる。例えば「カビがアレルギーなどの症状の原因となる」という文を「X が Y の原因となる」という文脈パターンに照合すると、X=「カビ」、Y=「症状」となり、「症状」を修飾する「アレルギーなどの」という重要な文節が抜けてしまう。この文から、「カビ→症状」という因果関係が抽出されるが、この関係はカビが何の症状を引き起こすのか分からないため有用でない。そこで除外する単語リストを関係ごとに手作業で作成し、除外単語が X、Y のどちらか一方でも出現する関係を意味的関係抽出サービスの出力から除外した。表 1 に原因結果の関係において問題となった単語例を示す。

表 1. 原因結果の関係で問題となる単語例

	問題となる単語例
原因	病気,欠乏,投与,遅れ,差,干渉,付着,乱用,破裂,傾向,状況,誤り,取りすぎ,誤認,消失,混在
結果	症状,問題,影響,被害,歪み,ダメージ,誘発,変化,効果,歪,失敗,現象,作用,支障,制限,異変

3.3 単語縮退による上位下位関係獲得

複数の形態素から構成される複合名詞は、その末尾に出現する形態素が上位語になる可能性が高い[11]。例えば、「炎症」は「出血性炎症」の上位語となっている。そこで、3.1 節、3.2 節の処理で獲得した単語と、NHK オンデマンドの番組概要文に出現する単語を対象として、単語縮退による上位語を生成する。この処理では、複数の形態素から成る単語に対して先頭から形態素を除外した語を生成し、その語が辞書の見出し語に含まれる場合に、元の単語の上位語とする。含まれない場合には、さらに次の形態素を除外した語を生成して辞書との照合を繰り返す。例えば「出血性炎症」は、「出血/性/炎症」と 3 つの形態素に分かれる。先頭の形態素を除外した語「性炎症」は辞書の見出し語に含まれないため、次の形態素を除外した語「炎症」が「出血性炎症」の上位語とみなされる。実験では、辞書の見出し語として日本語 WordNet[12]に登録されている名詞を利用した。

3.4 獲得した単語間関係の精度・カバー率調査

上位下位関係抽出ツールを利用して、2007 年～2011 年の 5 年分の Wikipedia のダンプデータから上位下位関係と属性関係を獲得する処理を行った。また意味的関係抽出サービスでは、各関係数個程度の単語ペアのシードを入力とし、得られた結果の上位ペア(信頼性の高い処理結果)を、再度入力として意味的関係抽出サービスに与えることによって、原因結果、名物、材料の関係を獲得した。ツールにより獲得した関係例を表 2 に、関係数とその精度を表 3 に示す。

表 2. 獲得関係例

単語 X	関係名	単語 Y
筆記具	[上位下位]	シャープペンシル
映画	[上位下位]	七人の侍
炎症	[上位下位]	出血性炎症
アレルゲン	[原因結果]	気管支ぜんそく
エルニーニョ現象	[原因結果]	水温上昇
調布	[名物]	深大寺そば
長谷寺	[名物]	あじさい
ウイスキー	[材料]	大麦麦芽
パンナコッタ	[材料]	ココナツ
伊勢市	[学校]	御薗小学校
J・D・サリンジャー	[作品]	A boy in France

表 3. 獲得した単語間関係数と精度

関係名	獲得関係数	精度
上位下位(WikiPedia)	8,591,469	90.0%*
上位下位(単語縮退)	1,347,382	82.5%
属性	5,213,455	94.0%*
原因結果	77,636	75.0%
名物	183,093	49.0%
材料	49,711	73.0%

* 上位下位と属性に対する精度は文献からの引用

上位下位(単語縮退)、原因結果、名物、そして材料に対する関係獲得結果の精度は、獲得された関係からそれぞれ 200 ペアをランダムサンプルし、1 人の評価者(著者)による

判定で算出した。

獲得した全関係に出現する異なり単語数は 3,458,913 語であった。NHK オンデマンドでこれまでに公開された 25,769 個の番組概要文に出現する名詞(異なり数 94,456 語)を取り出し、これらの名詞に対して、獲得した関係に出現する名詞のカバー率を調査した。結果を表 4 に記す。

表 4. 獲得した関係に出現する名詞のカバー率

関係	カバー率
全関係	72.8% (68,726/94,456)
上位下位(単語縮退)以外の関係	47.7% (45,042/94,456)

表 3 から自動獲得できる単語間関係は 6%～51.0%の誤りが含まれることが分かる。また表 4 に示す全関係のカバー率は 72.8%と高い値であるが、これは、番組概要文に出現する名詞を縮退して生成した上位下位関係を含むためで、上位下位(単語縮退)を除いた関係では半分に満たないカバー率であることが分かる。現状では改良の余地が残されているが、異なり単語数が約 346 万語、獲得した全関係数が約 1,546 万と大規模なものであり、テキスト間の類似性評価における効果は期待できる。

4. 番組間の類似性評価

本章では、前章で獲得した単語間の関係を利用して NHK オンデマンドにおける番組概要文間の類似性を評価する提案手法を説明する。

4.1 番組概要文

NHK オンデマンドに登録されている番組概要文の平均文字数は 170 文字、含まれる平均名詞数は 26 語であった。本稿では、この番組概要文に含まれる名詞を手掛かりとして、番組間の類似性を評価する。

4.2 関連度スコア計算

2 つの番組の類似性を評価するため以下の手順で 2 つの番組を結ぶグラフを生成する。

1. 番組タイトルと、番組概要文に含まれる名詞をノードとし、番組タイトルのノードと各名詞のノードをエッジで結合。
2. 類似性比較対象の番組に対しても、1.と同様に番組タイトルのノードと各名詞のノードをエッジで結合。
3. 1.と 2.の名詞のノードを、前章で生成した単語間関係により連結。例えば、番組概要文に「生活習慣病」と「たばこ」という名詞のノードがあり、「生活習慣病／高血圧」、「高血圧／喫煙」、「喫煙／たばこ」といった関係が獲得されている場合は、4 つの名詞ノードを順に結ぶエッジを生成。

生成したグラフに対して 2 つの番組がどの程度強く連結されているかを評価することによって、2 つの番組の類似性を評価する。この処理では、Web 検索などで使われているランダムウォークのアルゴリズムの一つの Green Measures[13]を利用する。この手法では、一つのノードから別のノードへ遷移する確率を行列 M で表現し、(1)式で Green matrix を定義する。

$$G := \sum_{t=0}^{\infty}(M^t - M^{\infty}) \tag{1}$$

ここで、M^t は t 回目のランダムウォークのステップにおける遷移行列を示す。行列 G の i 行 j 列の要素は、ノード i とノード j がどの程度関連するかを示す値と解釈できる。最終的な関連度スコアは、Green matrix を利用した(2)式により定義される。

$$S(i,j) := G_{ij}\log(1/\nu_j) \tag{2}$$

ここで ν は equilibrium measure と呼ばれ、任意のベクトルμに対し無限に遷移を繰り返した後に収束するベクトル($\mu M^{\infty} = \nu$)を示す。(2)式の対数の項は、ν の j 番目の要素の値が大きい(遷移を繰り返した後の最終状態としてノード j に収束する可能性が高い)場合に対する補正を与え、情報検索などで用いられる idf 値と同じような役割を果たす。

式(2)を利用して 2 種類の類似性評価手法を提案する。1 つ目の手法では、$S(p_1, p_2)$ を直接用いることにより 2 つの番組 p_1, p_2 の類似度 $S_{dierct}(p_1, p_2)$を定義する。

$$S_{direct}(p_1, p_2) = S(p_1, p_2) \tag{3}$$

ノードを結ぶエッジに与える重みは(4)式、(5)式の値とする。この値は、遷移確率を表す行列 M の各要素となる。

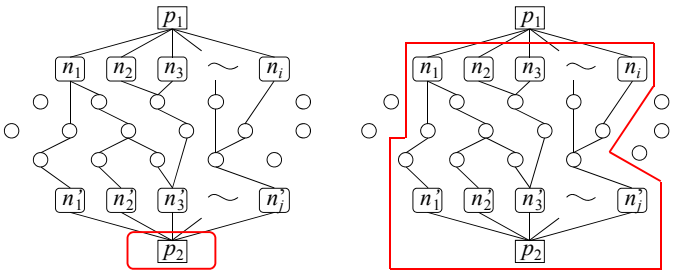
[番組から名詞へのエッジ]

$$e(p_i, n_j) = tf(n_j)idf(n_j)/Z_{p_i} \tag{4}$$

[名詞間のエッジ]

$$e(n_i, n_j) = 1/Z_{n_i} \tag{5}$$

手法 1: p_2 に与えられる値のみを利用　　手法 2: $p_1 \rightarrow p_2$ の経路上にある全ノードに与えられる値を利用



[p] : 番組p　　[n] : 番組概要文に出現する名詞n
○ : 自動獲得した単語間関係に出現する名詞

図 2. 2 つの手法における関連性スコア計算の概略

ここで、*tf*(*n*)は名詞 *n* の番組 *p* における出現頻度、*idf*(*n*)は名詞 *n* の全番組における逆文書頻度、そして、Z_p、Z_nは、それぞれ、*p*、*n* から他のノードへのエッジの重みの合計を示す。

2 つ目の手法では、p_1 から p_2 へのパス上にある全ノードに与えられた $S(i, j)$の値の合計を利用し、2 つの番組 p_1, p_2 の類似度 $S_{related}(p_1, p_2)$ を(6)式で定義する。

$$S_{related}(p_1, p_2) = \sum_{v \in vertex(p_1, p_2)} S(p_1, v) \tag{6}$$

ここで *vertex*(p_1, p_2)は、p_1 から p_2 へのパス上にある全ノードを示す。この手法においても、ノードを結ぶエッジに与える重みは(4)式、(5)式の値とする。図 2 に 2 つの手法における関連度スコア計算の概略を示す。

5. 実験

提案手法の有効性を示すために、NHK オンデマンドに登録されている番組を対象とした関連番組のリランキング実験を行った。まず、2010 年 9 月から 2011 年 5 月までに登録されていた 25,769 番組から、以下の制約のもとで 352 番組をランダムにサンプルした。

- 番組タイトルが同じ番組は取り出さない(例えば「NHK スペシャル」は 1 番組のみサンプル)
- 関連番組を 2 番組以上持つ

次に、NHK オンデマンドで提示された 352 番組の関連番組を対象として、筆者を含まない 3 名のアノテータにより、サンプルした番組とその関連番組間の類似性をランキングする作業を行った。各番組に関する関連番組は Okapi-BM25 を利用した手法[1]をベースに抽出され、一つの番組に対して平均 10.4 個の関連番組が提示されていた。3 名のアノテータが付与したランキング結果の相関を確認するため、(7)式に示す同順位を考慮した順位相関(Spearman's rank correlation)を利用する。

$$\rho = T_x + T_y - \sum D^2 \Big/ 2\sqrt{T_x T_y} \tag{7}$$
$$T_x = \left(N^3 - N - \sum_{i=1}^{n_x} (t_i^3 - t_i)\right) \Big/ 12$$
$$T_y = \left(N^3 - N - \sum_{j=1}^{n_y} (t_j^3 - t_j)\right) \Big/ 12$$

ここで、*D* は比較する 2 つのデータにおける順位の差、*N* はデータ数、n_x、n_yは 2 つのデータの同順位の個数、t_i、t_j は同順位の大きさを示す。アノテータによるランクの順位相関平均は 0.565 であった。これは、一定の一致度であったと解釈できる。最終的に 3 名のアノテータが付けた類似性のランクを平均し、平均ランクの昇順に類似すると判断したデータを基準とする。このデータと各手法によるランクを(7)式の順位相関により評価する。

5.1 ベースライン手法

本節では、提案手法と比較する 3 つのベースライン手法を紹介する。

Okapi-BM25 を利用した手法

Goto らの手法[1]では文書 p_1 に対する文書 p_2 の類似性を、Okapi-BM25 の指標を利用した(8)式で評価する。

$$S_{BM}(p_1, p_2) = \sum_{n \in p_1} idf(n) \cdot \frac{tf_{p_2}(n) \cdot (k+1)}{tf_{p_2}(n) + k \cdot (1 - b + \frac{|p_2|}{avgdl})} \cdot \frac{(k'+1) tf_{p_1}(n)}{k' + tf_{p_1}(n)} \tag{8}$$

ここで、*idf*(*n*)は単語 *n* の逆文書頻度、$|p_2|$は p_2 の文書長、*avgdl* は平均文書長、*k*、*k'*、*b* はパラメータであり、*k*=3.0、*k'*=100.0、*d*=0.75 を使用している。

***tf-idf* による手法**

tf-idf による手法では、文書 *p* に出現する単語 *n* に対して(9)式の重みを与えて文書を単語のベクトルで表現する。

$$w_{TFIDF}(n) = tf_p(n) \cdot idf(n) \tag{9}$$

文書 p_1 に対する文書 p_2 の類似性は、2 つの文書のベクトル間のコサイン類似度により評価し、(10)式の降順に類似していると判断する。

$$S_{TFIDF}(\vec{p}_1, \vec{p}_2) = \frac{\vec{p}_1 \cdot \vec{p}_2}{|\vec{p}_1 \| \vec{p}_2|} \tag{10}$$

単語間関係を利用した手法

自動獲得した単語間関係を用いて文書 *p* に出現する単語 *n* を拡張し、文書に出現する単語 $n \in p$ と、*n* と直接関係を持つ単語 n_{rel} を要素とするベクトルで文書 *p* を表現する。*n* に与える重みは(9)式、n_{rel} に与える重みは(11)式を用いる。

$$w_{rel}(n_{rel}) = w_{TFIDF}(n) / N_{rel}(n) \tag{11}$$

ここで、N_{rel}(*n*)は *n* と関係を持つ単語数を示す。文書 p_1 に対する文書 p_2 の類似性は、(10)式と同様に 2 つの文書のベクトル間のコサイン類似度により評価する。

5.2 各手法によるリランキング実験

ランダムサンプルした 352 番組とその関連番組に対して、前節で説明した 3 つのベースライン手法と 4.2 節で説明し

表 5. 各手法の評価結果

手法	rank correlation
ベースライン 1 (Okapi-BM25)	0.370
ベースライン 2 (*tf-idf*)	0.350
ベースライン 3 (単語間関係)	0.371
提案手法 1 ($S(p_1, p_2)$を直接利用)	0.351
提案手法 2 (経路上の全ノード利用)	0.423

表 6. 各手法のリランキング処理結果

【対象番組】 まる得マガジン エクササイズ股関節と脚

【番組概要文出現単語】 股関節, 脚, ストレッチ, 動作, あなた, 楽, 可動域, アップ, 姿勢, 股関節まわり, 5分, 速度, 歩幅, 運動, 筋肉, 範囲, 1日

リランキング対象番組タイトルと番組概要文に出現する単語	人手による順位 (スコア)	提案手法 2 の順位 (スコア)	ベースライン 2 の順位 (スコア)
まる得マガジン エクササイズ脇と背中 【単語】ストレッチ, 背中, 体, 両脇, 5分, コンディション, 緊張, 毎日, あなた, 頭, 運動, 姿勢, 筋肉, 脇, 1日	1 (1.333)	1 (0.913)	1 (0.328)
ためしてガッテン バナナ大革命！新食材宣言 【単語】バナナ, 果物, 野菜, 調理, 味, 大公開, 栄養, 満点, 手, 生産, 消費量ナンバーワン, 判明, 世界, 出荷, わたし, 簡単, 高級食材, 実力, 食材, ガッテン流バナナ調理術, 大部分, 方法	2 (1.667)	2 (0.582)	2 (0.0)
ふだん着の温泉 青森・下風呂温泉 【単語】湯治場, 体, 厳寒, 室町時代, 舞台, 海峡, 役割, 人々, 2つ, 共同温泉, ニシン漁, 戦前, 大湯, 井上靖, 地, 地元, 拳, 交換, 情報, 漁師, 漁師仲間, 津軽海峡, 小説, さまざま, 有名, 場, 大切, 下風呂温泉, 温泉, 芯, 新湯	3 (3.333)	3 (0.561)	2 (0.0)
あさイチ JAPAなび 京都 【単語】京都, 大八車, トーク, 巨大, 大覚寺, 秋, 水面, 大沢池, 旅, 着物, 真帆, 野菜農家, 逆さ紅葉, 幻想的, 一緒, 出会い, 精進料理, 人気, 京野菜, ライトアップ, 光景, 野口久子, 82歳, たんのう, お寺, 嵯峨野, 俳優, アンティーク着物巡り, 紅葉づくし, 13代目, 軽妙, 世界, 圧巻, 住職, 存分, 古都	4 (3.667)	4 (0.203)	2 (0.0)

た 2 つの提案手法を適用して関連番組のリランキング処理を行い、これらの結果とアノテータにより生成した基準データとの相関を、(7)式に示す順位相関により評価した。結果を表 5 に示す。

経路上の全ノードに与えられる関連度スコアを利用した提案手法 2 の順位相関が 0.423 と最も高く、この手法が他に比べて人手によるランキング結果に近いことが分かる。一方、番組を表すノード p_1 から p_2 への直接の関連度スコアを利用した提案手法 1 は良い結果が得られていない。Green matrix を利用した関連度スコアの値は、直接エッジで繋がれているノード間は大きい値となるが、間接的に繋がれているノード間では極端に小さな値となっていた。そのため、直接繋がれているノード(2 つの番組に共通する単語)のみに影響を受け、従来手法とほぼ変わらない結果となってしまったと考えられる。

番組「まる得マガジン エクササイズ股関節と脚」の 4 つの関連番組を対象として、ベースライン手法 2 と提案手法 2 によりリランキングした結果を表 6 に示す。表 6 において、人手によるスコアは 3 名のアノテータにより付けられた順位の平均、提案手法 2 のスコアは(6)式の値、そして、ベースライン 2 のスコアは(10)式の値を示す。ベースライン手法 2 では、「ためしてガッテン バナナ大革命！新食材宣言」、「ふだん着の温泉 青森・下風呂温泉」、「あさイチ JAPAなび 京都」の 3 つの番組に対してスコアが 0.0 となっている。これは、対象とする番組と共通する単語が出現していないことが原因となっている。一方、提案手法 2 では、これらの番組に対して類似度を示すスコアを与えることが出来ており、その結果も人手による順位と一致している。このように提案手法では、従来手法でスコアを与えることができなかった番組に対しての順位付けが可能となり、順位相関が向上したと考えられる。

5.3 関係種類別・関係数別の評価

前節において最良の結果であった提案手法 2 を用いて、利用する関係の種類別にランダムウォークを行った。アノテータにより生成した基準データとの順位相関の結果を表 7 に示す。上位下位関係だけを利用した結果が最良であり、次に名物関係が効果的であったことが分かる。ここで「関係なし」とは、全ての単語間関係を使わなかった場合で、番組タイトルノードと、番組概要文に出現する名詞ノード間がエッジで結ばれたグラフを利用する。「関係なし」の順位相関は 0.400 と、表 5 の各ベースライン手法と比較して高く、「同じ番組に出現する単語は類似した単語である」というヒューリスティックが利いたと考えられる。

また、全関係から一定の関係数を無作為抽出し、関係数による影響を提案手法 2 により行った。結果を図 3 に示す。

表 7. 使用した関係種類別の評価結果

関係の種類	rank correlation
上位下位(Wikipedia、単語縮退)	0.412
属性	0.398
原因結果	0.400
名物	0.405
材料	0.403
関係なし	0.400

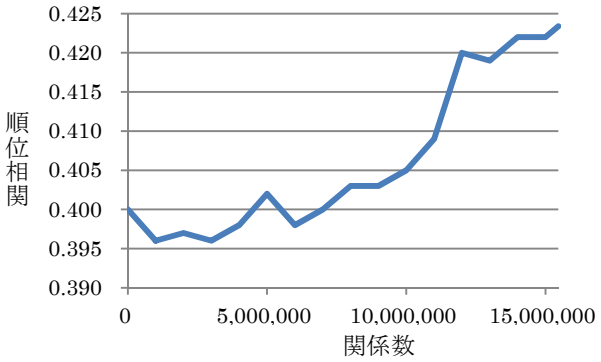


図 3. 関係数別の評価結果

関係数が 7,000,000 個あたりから順位相関の向上が見られ、関係数の増加に伴い順位相関も上昇している。現状では関係数は約 1,546 万個であるが、これを増加させることによりさらなる順位相関の向上が期待できる。

5.4 エッジへの重み付けの考察

提案手法では、番組タイトルノードと番組概要文に出現する名詞ノード間のエッジに対して(4)式の重みを与えている。この式では単語の重要度を *tf-idf* 値により考慮しているが、単語の重要度を考慮しないで、番組タイトルノードから出るエッジ数で割った(5)式を使うこともできる。単語の重要度を考慮しないで提案手法 2 の実験を行った結果を表 8 に示す。

表 8. 単語の重要度を考慮しないエッジ重み付け手法の評価結果

手法	rank correlation
提案手法 2 (*tf-idf* を考慮しないエッジ重み付け)	0.427

単語の重要度を考慮しない表 8 の結果では、表 5 に示す単語の重要度を考慮した結果と同等以上の値が得られた。番組概要文内で類似する単語群のノードは直接的、または間接的にエッジで結合されているため、ランダムウォークを行うと、類似する単語群のノードには高い関連度スコアが与えられる。番組概要文内で類似する単語が多いほど、その番組の中心的な単語と考えられるため、*tf-idf* などの指標を用いてエッジに重み付けをしなくても暗黙的に単語の重要度が考慮されたと考えられる。

6. おわりに

本稿では、Wikipedia や Web テキストから自動獲得した単語間の関係を用いることにより、2 つの文書間の類似性を評価する手法を提案した。提案手法では、2 つの文書に出現する名詞間を自動獲得した関係で結んだグラフ構造を生成し、ランダムウォークによりグラフ上のノード間の関連度スコアを算出することにより、2 つの文書の類似性を評価した。評価実験では、2 つの文書を繋ぐ経路上の全ノードの関連度スコアを利用した手法による結果が、従来手法と比較して人手によるランキング結果に近いことを示した。関係種類別の実験では、上位下位関係が最も効果的であることが分かり、関係数を変えて行った実験では関係数の増加に伴い順位相関も上昇する傾向が確認できた。さらに、グラフ生成の際に *tf-idf* などの指標を用いてエッジに重み付けしなくても、ランダムウォークによって暗黙的に単語の重要度が考慮できることを確認した。

実験で利用した単語間の関係には誤りが含まれる。また現状では、Web や Wikipedia から獲得した関係だけでは番組概要文に出現する名詞の半分弱しかカバーできていない。今後、関係数を増加させ、さらには人手により関係のチェックを行い、文書間類似性評価処理の検証を進める予定である。

参考文献

[1] J. Goto, H. Sumiyoshi, M. Miyazaki, H. Tanaka, M. Shibata, and A. Aizawa: Relevant TV Program Retrieval using Broadcast Summaries, Proceedings of ACM on Intelligent User Interfaces(IUI), pp.411-412, 2010
[2] S. Robertson and S. Walker: Okapi/ Keenbow at TREC-8, In Proceedings of TREC-8, pp151-162, 1999.
[3] 奥健太,中島伸介,宮崎純,植村俊亮: 状況依存型ユーザ嗜好モデリングに基づく Context-Aware 情報推薦システム，情報処理学会論文誌. データベース, Vol. 48, SIG11 (TOD_34), pp. 162-176, 2007.
[4] Greg Linden, Brent Smith, and Jeremy York: Amazon.com recommendations, IEEE Internet Comput., vol7, no.1, pp. 76-80, 2003.
[5] Y. Koren, R. Bell and C. Volinsky: Matrix Factorization Techiques for Recommender Systems, IEEE Computer, pp.42-49,2009.
[6] P. Melville, V. Sindhwani: Recommender Systems, Encyclopedia of Machine Learning, Springer, 2010.
[7] 隅田飛鳥, 吉永直樹, 鳥澤健太郎: 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.
[8] 山田一郎, 橋本力, 呉鍾勲, 鳥澤健太郎, 黒田航, Stijn De Saeger, 土田正明, 風間淳一: Wikipedia を利用した上位下位関係の詳細化, 自然言語処理, Vol.19, No.1, pp. 3-23, 2012.
[9] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda and Masaki Murata: Large Scale Relation Acquisition using Class Dependent Patterns, In Proceedings of the IEEE International Conference on Data Mining (ICDM'09), pp.764-769, 2009.
[10] Jun'ichi Kazama and Kentaro Torisawa: Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations, In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), pp. 407-415, 2008.
[11] 黒田航, 李在鎬, 野澤元, 村田真樹, 鳥澤健太郎: 鳥式改の上位語データの人手クリーニング, 第 15 回言語処理学会年次大会発表論文集, pp76-79, 2009
[12] Francis Bond, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto: Japanese SemCor: A Sense-tagged Corpus of Japanese in The 6th International Conference of the Global WordNet Association (GWC-2012), 2012
[13] Ollivier Yann, Senellart Pierre: Finding Related Pages Using Green Measures: An Illustration with Wikipedia, Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, pp.1427-1433, 2007.