

VOD 講義用字幕文の共起語グラフによる話題変化抽出

小山 登¹ 椎名 広光² 北川 文夫³

yamamame@gmail.com¹, shiina@mis.ous.ac.jp², kitagawa@mis.ous.ac.jp³

岡山理科大学大学院 総合情報研究科¹

岡山理科大学 総合情報学部^{2,3}

1 はじめに

現在, Web 教材を用いた e-Learning, すなわち WBT (Web-Based Training) と対面授業を組み合わせたブレンディッドラーニングやスライドと講義の動画を配信する VOD (Video On Demand) による e-Learning 講義などさまざまな大学で e-Learning の利用が行われている. 岡山理科大学では VOD による e-Learning 講義を 2004 年度から行っている. その中では VOD 教材を見かえすのに, タイトルの内容から目的の教材を探す必要があり, 復習が難しい状態にある.

目的の教材 VOD を検索する方法として, これまで教材として提供されている VOD システムの映像に付加されている字幕を利用して, 検索語の頻度のヒストグラムに対して統計的处理方法である混合正規分布モデルの EM アルゴリズムを用いた最尤推定を行う方法 [1] を提案した.

それに対して本研究では, 講義の字幕の各文に対して重要語とその共起語のグラフを作成し, 字幕文に対応した共起語グラフの変化から話題を検出する手法を提案する. また, 共起語グラフの変化は 2 つの手法で評価し, 評価の方法を比較する.

2 e-Learning 講義システムと検索機能

VOD 講義の実行画面は図 1 のような構成で, 左上に講師の動画, 左下にそのセクションの内容を表示する. 画面の右側に講義資料となるスライドを表示する構成になっており, ボタンで他のスライドに切り替えることができる. 1 回の講義は 3 つのセクションに分かれており, 1 つのセクションは 20 ~ 30 分程度となっている. また, 各セクションの最後に講義内容に関する課題があり, 講義内容の理解を確認するために用いられている.

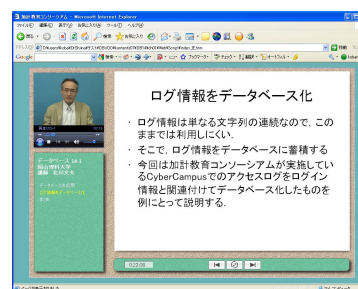


図 1: VOD 講義画面

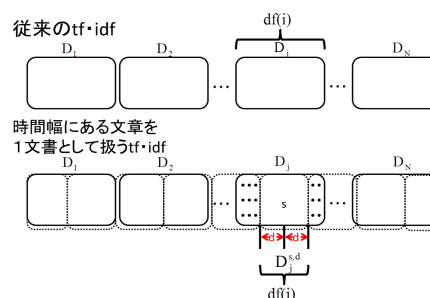


図 2: $tf \cdot idf$ の説明

この VOD 講義の字幕に対して処理を行い, 講義のセクションに対して話題を検出し提供する.

3 話題の検出

本研究では字幕データの各文に対する重要語とその共起語から作成されるグラフを利用し, 話題の変化を検出する.

3.1 字幕の重要語

VOD 講義の中で各文の代表となる語を重要語として処理を行い, その変化から話題の変化を検出する. そのために, 各文の話題に対応する重要語の算出が必

D(収入, 検索エンジン)=2

↔

検索エンジンの/主な/収入源は/三つほど/あります。

図 3: 文節の差

要であると考えられる。本研究での $tf \cdot idf$ の定義 (図 2) は、重要語を求める対象の講義 D_j の文 s の前後 d 時間の文章を 1 つの文書 $D_j^{s,d}$ とみなしているのが特徴で次のように定義する。

[文の前後を用いた $tf \cdot idf$ の定義]

単語 w_i が文書 $D_j^{s,d}$ に含まれる数 $tf(i, D_j^{s,d})$, 時間区間 $2d$ で区切った文書数 N , 単語 w_i を中心とした時間区間 $2d$ で区切った文書の数 $df(i)$ とするとき、

$$tf \cdot idf(i, D_j^{s,d}) = tf(i, D_j^{s,d}) \cdot \log \frac{N}{df(i)}.$$

3.2 重要語に対する共起語の共起度

各字幕文の $tf \cdot idf$ によって求められた重要語の共起語とその共起度を求める。本研究で共起語とはある単語が現れた文に同時に現れる名詞としている。

1 つのセクションの字幕全体 D , 単語 w_i , 単語と共起する語 w_j , w_i が出現する文 S_i , w_i と w_j の文節の差 $D(w_i, w_j)$, w_j の頻度 $frq(w_j)$ とするとき、共起度 $Cov(w_i, w_j)$ を次の式で表す。

$$Cov(w_i, w_j) = \sum_{s_i \in D} \sum_{w_j \in s_i} \frac{\sqrt{frq(w_j)}}{D(w_i, w_j) + 1}$$

3.3 共起度の計算例

共起度の計算例を岡山理科大学サイバーキャンパス 2007 年度データベース 14 回目の講義の重要語「検索エンジン」を用いて説明する。

(1) 文節の差は、図 3 の例文では、 w_i を「検索エンジン」、 w_j を「収入源」として、その文節の差は $D(\text{「検索エンジン」}, \text{「収入」}) = 2$ となる。「検索エンジン」の共起語となる名詞は { サイト, キーワード, ..., 予想 } の 90 個となる。

(2) 重要語「検索エンジン」に対して共起する各名詞の出現頻度は、{ 26, 18, ..., 1 } と計算される。

(3) 名詞ごとに共起度を計算する。共起語として「収入源」を選んだ場合、

表 1: 共起度 (重要語: 検索エンジン)

| 順位 | 単語 | 共起度 | 頻度 | 共起頻度 |
|----|---------|-------|----|------|
| 1 | Google | 6.752 | 15 | 5 |
| 2 | 検索 | 5.196 | 12 | 3 |
| 3 | キーワード広告 | 3.431 | 15 | 7 |
| 4 | サイト | 2.694 | 26 | 2 |
| 5 | 入り口 | 2.578 | 3 | 5 |
| 6 | 仕組み | 2.364 | 14 | 6 |
| 7 | 最初 | 2.333 | 3 | 6 |
| 8 | 信頼 | 2.205 | 4 | 3 |
| 9 | 手法 | 2.167 | 7 | 2 |
| 10 | ソフトウェア | 2.036 | 4 | 7 |

$frq(\text{「収入」}) = 5, D(\text{「検索エンジン」}, \text{「収入」}) = \{2, 19, 19, 6, 28\},$

$$Cov(\text{「検索エンジン」}, \text{「収入源」}) = \frac{\sqrt{1 \cdot 24}}{2+1} + \frac{\sqrt{1 \cdot 24}}{19+1} + \frac{\sqrt{1 \cdot 24}}{19+1} + \frac{\sqrt{1 \cdot 24}}{6+1} \approx 0.273$$

となる。

重要語「検索エンジン」の共起語を共起度順に並べたものを表 1 に示す。共起度順では、「検索」や「キーワード広告」の共起度が高い。講義「データベース」の 14 回目の内容は「検索サイトの収入」に関する内容であり、そのため「検索エンジン」と「Google」「キーワード広告」等が講義内で共起度が高くなっていると考えられる。

3.4 共起語グラフ

共起語の作成法は、各文の重要語を初期節点とし、そこから共起度の高い共起語を辺で接続する。本研究では、共起語を 2 段階まで、共起度が高いものを取り出し、共起語を辺で接続する。このとき、同じ語が複数入らないようにする。共起語グラフの作成の手順を以下に示す。

(1) 各文から $tf \cdot idf$ の最も高い名詞を初期節点とする。

(2) 初期接点と共起する名詞を共起度の高い単語を取り出し、節点とする。

(3) (2) で得られた節点の語に対して共起度順に共起する名詞を共起度順に取り出し、節点とする。

(4) 得られた共起語を共起度の合計の高いものに、共起しているそれぞれの節点を辺で接続する。

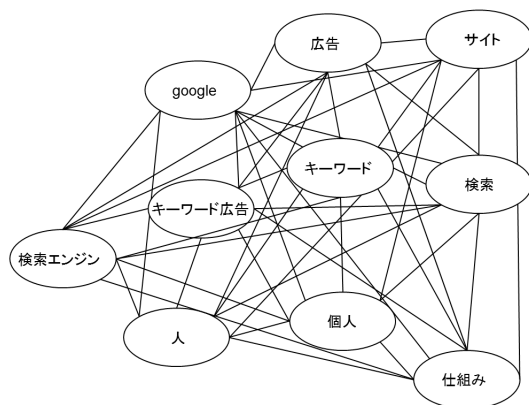


図 4: 共起語グラフ (検索語:「広告」)

例として、データベース 14 回目の講義の中から、重要語「広告」から作られるグラフを図 4 に示す。

3.5 話題の変化の検出

話題の検出には、連続する 2 文の共起語グラフを比較し、話題の変化を検出する。変化の検出法は本研究では次の 2 つの手法を比較する。

(1) グラフの同じ語の節点を接続する手法 (グラフの接続法)。

(2) グラフの節点の総次数である次数中心性の対応分析 [4] を用いる手法 (次数中心性の対応分析法)。

3.5.1 グラフの接続法

文の重要語から作られるグラフと次の文の重要語から作られるグラフの同じ語を辺でつなぎ、グラフ同士をつなぐ辺の数が少なければ、話題の変わり目とする手法である。重要語「広告」の共起語グラフ (図 4)、次の文の重要語「キーワード広告」の共起語グラフを接続したグラフを図 5 に示す。図 5 では、左右にそれぞれの文の重要語から、作られたグラフがあり、それぞれの共起語グラフの一致する語を辺で接続している。それぞれの共起語グラフが 10 個の節点を持ち、つなぐ辺が 8 つあるため共起語グラフの半分以上が同じ語で変化が少なく、同じ話題の文であると考えられる。

3.5.2 次数中心性の対応分析法

グラフの節点の総次数である次数中心性を用い、対応分析によって共起語グラフの距離を計算する。この距離が大きいとき、話題の変わり目とする手法である。

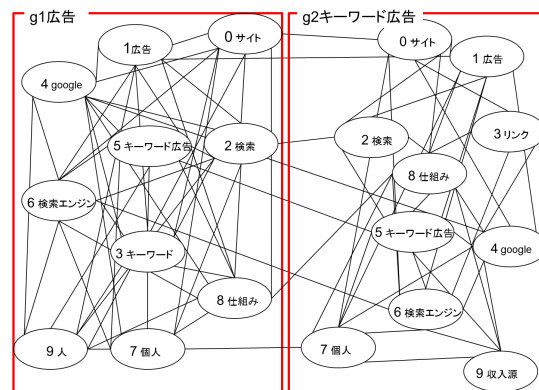


図 5: 接続したグラフ (検索語:「広告」「キーワード広告」)

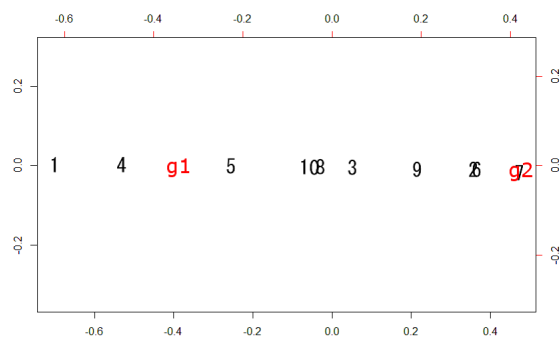


図 6: 対応分析による節点とグラフの差

例として、重要語「広告」と重要語「キーワード広告」に対して次数中心性に対する対応分析を行った結果を図 6 に示す。図 6 は、各接点と 2 つのグラフの距離を表したものである。他の文に適用したものと比較して距離が大きくないため、同じ話題であると考えられる。

3.5.3 手法の比較

評価実験として、グラフの接続法と次数中心性の対応分析法の時系列変化を比較した。対象とするデータは講義「データベース」14 回目の 2 セクション全体に対して文ごとのそれぞれの手法での変化を調査した。図 7、図 8 の横軸は文の番号を表している。枠で示した部分は人手で作成した話題の区間を示している。

図 7 は共起語グラフの接続の時系列での変化を表し、縦軸は 2 つのグラフ間の共通単語数を表している。この手法では、時間ごとに大きくグラフが変化して変わり目がわかりにくくなっている。しかし、最後の検

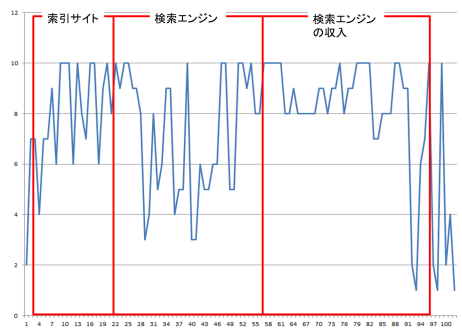


図 7: 字幕文におけるグラフの変化 (グラフの接続法)

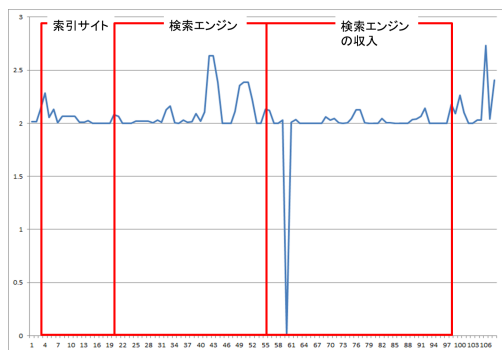


図 8: 字幕文におけるグラフの変化 (次数中心性の対応分析法)

検索エンジンの収入とまとめの間は大きく共通単語が変わっているため、話題が変化しているとわかる。

一方、図 8 は次数中心性の対応分析の時系列での変化を表し、縦軸は対応分析による 2 つのグラフの距離を表している。セクションの中盤と最後にグラフの距離が大きい文がある。セクションの中盤の話題の変化は、人手で作成した話題の変化のとおりに検索エンジンとその収入の話題の変化を表している。同様にセクションの最後は、検索エンジンの収入の話題が終わりまとめに入る話題の変化を表している。

どちらの手法でも、人手で評価された話題の変わり目とは違う場所に話題の変わり目が検出されている。しかし、実際の字幕では人が見ても話題が変わっているため、小さな話題の変化をとらえているためであると考えられる。また、検索エンジンの収入に関する話題では、重要語が同じ文がならんでいることがあるため、話題がつながっているような結果となっている。

本手法では 2 文の間の差を用いているため、重要語となる名詞が存在しない文や、意味のない文が含まれていることが問題となっている。この問題を解決するために、複数の文から話題の変わり目を検出する必要

があると考えられる。

4 まとめ

文の重要語とその共起関係から話題の変わり目を検出する手法を提案した。提案した話題の変わり目を検出するシステムは、共起関係をグラフ化しその変化から話題の変化を求めるシステムである。各文に対して、 $tf \cdot idf$ により、重要語を抽出し、その共起語からグラフを作成する。このグラフの節点の中心性からグラフの差を求め、話題の変化を検出する手法である。

このような手法は人間の感覚とは違いがあるかもしれない。しかし、大学の講義などのように、ある一定の目的をもった動画に対しては有効な検索方法であると考えている。また、話題は大きな話題の中に小さな話題が含まれて構成されている。このため話題の変化が小さいものを話題の変化としていしまう問題がある。また、意味をなさないような文や名詞が存在しない文を話題の変化として検出してしまう問題がある。このような問題を解決するのが今後の課題である。

参考文献

- [1] 小山, 椎名, 北川, 字幕付き VOD 講義に対する共起語による類似映像区間推定, 日本行動計量学会第 39 回大会抄録集, pp313-316, 2011.
- [2] Kobayashi, Koyama, Shiina and Kitagawa, Detecting Movie Segments Using Gaussian Mixture Models for VOD Lectures with Japanese Subtitles, JSiSE Vol10, 2012. (to appear)
- [3] 金森, 竹ノ内, 村田, パターン認識, 共立出版, 2009.
- [4] 鈴木, ネットワーク分析, 共立出版, 2009.
- [5] 伊藤, 藤井, 石川, 音声文書検索を用いたオンデマンド講義システム, 電子情報通信学会技術研究報告 SP 音声, Vol.101, No.523, pp.55-60, 2001.
- [6] 北, 津田, 獅々子, 情報検索アルゴリズム, 共立出版, 2002.