字幕付きＶＯＤ講義における発話内容のトピック分析

中村　槙吾[1]　　小林 伸行[2]　　椎名 広光[3]　　北川 文夫[4]
[1] 岡山理科大学大学院 総合情報研究科 情報科学専攻
[2] 山陽学園大学 総合人間学部 生活心理学科
[3,4] 岡山理科大学 総合情報学部 情報科学科
626653@teamgear.net[1], koba_nob@sguc.ac.jp[2], shiina@mis.ous.ac.jp[3],
kitagawa@mis.ous.ac.jp[4]

1　はじめに

現在，インターネット環境を利用して講義を行うVOD 講義 [1] が多くの大学で行われている．しかしながら，現状のシステムではVOD の内容に対する検索機能がほとんど作成されていないため，講義を何度も見たりしなければならない．

字幕データに対する検索語の出現頻度をもとにしたトピックの区間推定 [2] においては，検索語と検索語に共起する単語で推定区間の部分的に一致する関係が見られた．

そこで検索語と検索語と共起語の散布図を工夫して表示する情報の可視化を行うことで，目的のトピックを見つけられるのではないか考えられる．本研究では散布図を表示するにあたって，検索語と共起語の関連の大きさを示す共起度 [3] を定義し，共起度から共起距離を計算することで，その共起距離を散布図に加えることでトピックの区間を反映させている．

また，特定の区間におけるトピックの調査として，1 枚の講義スライドが表示されている範囲のキーフレーズ抽出 [4]，それに加えて講義全体におけるトピックの調査に LDA[5] によるトピック分析を行った．

2　単語共起度

2.1　共起度の定義

共起語は検索語と同じ文中に存在する語である．本研究では，関連したトピックに出現する語であると考えられる検索語に対して，字幕中でどの程度共起しているかを共起度として表し，共起度の高いものを検索に用いる共起語とする．

1 つのセクションの字幕全体 D，検索語 w_i, 検索語と共起する語 w_j，w_i が出現する文 S_i，w_i と 1 つの

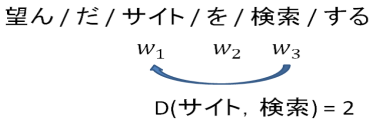

図 1: 共起語距離

w_j の文節の差 $D(w_i, w_j)$，w_j の頻度 $freq(w_i)$ とするとき，共起度 $Cov(w_i, w_j)$ を次の式で表す．

$$Cov(w_i, w_j) = \frac{1}{N(w_i, w_j)} \sum_{S_i \in D} \sum_{w_j \in S_i} \frac{\sqrt{freq(w_j)}}{D(w_i, w_j)+1}$$

2.2　共起度の計算例

検索語「サイト」を w_i として共起度の計算例を示す．

(1) 対象の講義のセクションの字幕から w_i と共起する名詞を w_j として取り出し，その文節の差を求める．文節の差は，図 1 の例文では，w_i を「サイト」，w_j を「検索」としている．サイトの場合の共起語となる名詞は { 検索, キーワード, $\cdots$, 機械 } となる．

(2) 共起する名詞に対して，対象の講義のセクションでの頻度を計算する．$\{31, 26, 24, \cdots, 1\}$ となる．

(3) 名詞ごとに共起度を計算する．共起語を「検索」とすると，

$freq$(“検索”) $= 31$,
D(“サイト”, “検索”) $= \{2, 2, 19, \ldots, 40\}$,
Cov(“サイト”, “検索”) $=$
$\frac{1}{9}(\frac{\sqrt{31}}{2+1} + \frac{\sqrt{31}}{2+1} + \frac{\sqrt{31}}{19+1} + \ldots + \frac{\sqrt{31}}{40+1}) = 1.169$

となる．

共起度は，単語間の距離を表していないので，共起度を利用した単語間距離 $Cd(w_i, w_j)$ を定義する．

$$Cd(w_i, w_j) = \frac{Cov(w_i, w_i)}{Cov(w_i, w_j)}$$

表 1: 共起語例　(検索語 :「キーワード」)

順位	共起語 Word	頻度	共起度 $Cov(w_i, w_j)$	共起距離 $Cd(w_i, w_j)$
1	広告	24	1.658	2.074
2	サイト	32	0.792	5.438
3	一つ	4	0.667	6.646
4	ビジネス	6	0.612	7.327
5	関係	3	0.577	7.832
6	ページ	18	0.576	7.847
7	順位	13	0.563	8.058
8	検索	31	0.532	8.590

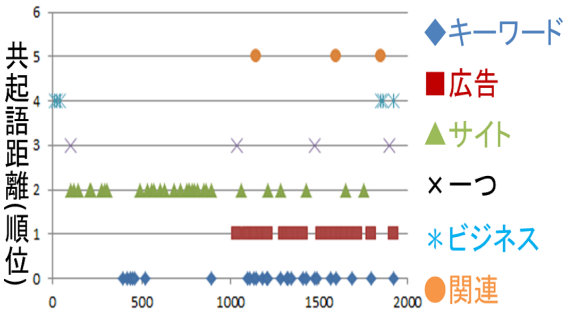

図 2: 検索語 :「キーワード」の共起語の共起度の順位による散布図

検索語を「キーワード」としたときの共起語を表 1 に示す.「広告」と「サイト」との共起度が高くなっているのが分かる．これは対象としている講義がデータベースに関するものであり，講義の中で「検索エンジンの収入」,「Google」,「キーワード広告」などの言葉が出現しているためである．

2.3　共起語の出現分布

共起語の講義中の出現分布を散布図で表す．散布図は，横軸を時間とし，縦軸を共起度順と共起度とした 2 種類がある．例として検索語を「キーワード」としたときの縦軸を共起度順とした図 2，縦軸を共起度を距離とした図 3 を示す．

3　キーフレーズ抽出

共起距離による分析では特定の単語とその単語に関連するトピックの区間を調べることができた．しかしある区間において，どのようなトピックが展開されているかを調べることができない．そこで，Yahoo!デベロッパーネットワークで公開されているキーフレーズ抽出 API[4] を利用し，特定の範囲におけるトピックを調査する．キーフレーズ API は，文章から特徴と

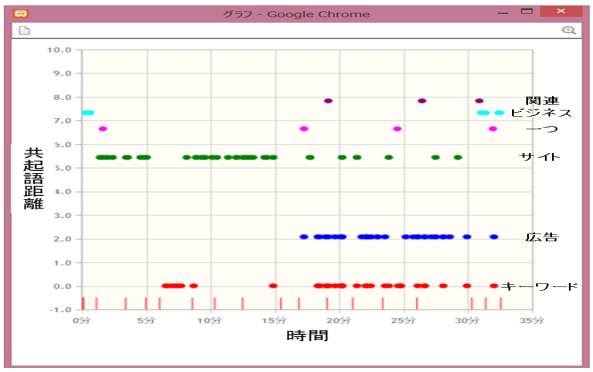

図 3: 検索語 :「キーワード」の共起語の共起距離による散布図

表 2: キーフレーズ抽出　(スライド単位)

スライド 7		スライド 8	
フレーズ	スコア	フレーズ	スコア
リンク	100	対偽装順位付け	100
サイト	82	ソフトウェア	53
エステ	77	手法	51
丸印	69	リンク	45
ユーザ	62	索引サイト	44
偽装リンク	61	呼び方	43
実際	60	ページランク	36
プログラム	54	人手	36
たくさん	54	評価	36
検索エンジン	51	ところ	36
いくつか	48	正しいサイト	34
左側	42	実際	33

なるキーフレーズを抽出し，スコアを付けて結果を返す API である．

本研究では，講義スライドが 1 枚ごとの範囲の字幕に対してキーフレーズ抽出を適用した．スライドごとのフレーズとスコアの結果の一部を表 2 に示す．表 2 ではスコアの高いフレーズがそのスライドのトピックに大きく関係している．スライド 7 では「リンク」のスコアが一番高く，その次に「サイト」が高い．そのことからあるサイトとあるサイトのリンクの仕組みについてトピックが展開されていることが推測できる．スライド 8 では「対偽装順位付け」のスコアが高く，「検索サイト」や「ページランク」など，出現しているフレーズから Web サイトの順位付けの仕組みについてのトピックだと推測できる．

4　LDA によるトピック分析

本研究では LDA（Latent Dirichlet allocation）によるトピック分析についても行った．LDA では単語の出現情報からその単語だけでなく，単語間の関連性か

ら潜在的なトピックを見つけることができる特徴を持ち，手順としては確率的な試行を利用して処理を行っている．本研究ではLDAによってトピック解析を200回行い，それぞれのトピック間のコサイン類似度によってクラスタリングを行う．その結果からトピックの特徴的なトピックのパターンを抽出する．

4.1 LDAの講義発話への適用

LDAとは複数のトピックで文書が構築されていることを想定したモデルである．そのため，講義の中でトピックを見つけるのに適したものと考える．

LDAはランダムにトピックを作るため，毎度同じ結果にはならない．そこで，繰返しトピックを生成し，真のトピックを解析する．

なお，LDAを行うに当たってVOD講義のセクションの字幕全体を取り出し，ChaSenによって名詞のみを抽出している．

4.2 コサイン類似度

1つのトピックにおける単語の出現確率をベクトルで表し，トピック$\boldsymbol{x}$の単語w_iの出現確率をx_i，トピック$\boldsymbol{y}$の単語w_iの出現確率をy_iとするとき，類似度$sim(\boldsymbol{x},\boldsymbol{y})$を次の式で表す．

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

4.3 クラスタリング

LDAで得られたトピックからコサイン類似度を計算し，その類似度からクラスタリングする．トピックを200件のデンドログラムを図4に示す．デンドログラムからクラスが5個になるように，中心となるトピックを1つ取り出したものが表4と表5である．それぞれのクラス内で並び順において中央となるトピックを代表値としたものが表4である．表5はクラス内のトピックでコサイン類似度を求め，その合計が一番大きなトピックをクラスの代表値としたものである．表3にその計算例を示す．

表4の①のトピックは，「キーワード」「広告」「検索」などの要素で構成されており，②のトピックと比較すると非常に近いことが分かる．④や⑤のトピックと比較すると比較した場合には関係が薄いことから，正しく分類されていると考えられる．表5においても同様に考えられる．

表3: トピック間のコサイン類似度の計算例

	トピック1	トピック2	…	トピックn	合計
トピック1	1	0.6	…	0.2	16.7
トピック2	0.6	1	…	0.2	18.0
…	…	…	…	…	…
トピックn	0.2	0.2	…	1	18.5

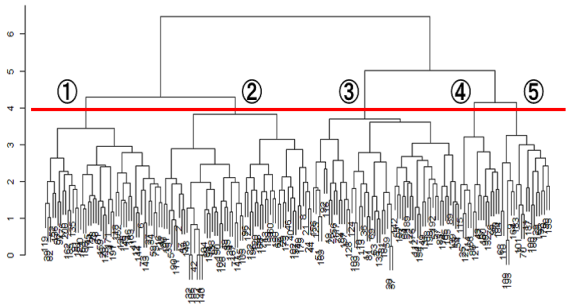

図4: LDAで作成したトピック200件のデンドログラム

表4: 並び順による代表トピック

トピック	①	②	③	④	⑤
要素	キーワード	キーワード	キーワード	キーワード	分類
	広告	広告	説明	円	説明
	検索	検索	検索	段階	セクション
	エンジン	エンジン	順位	順位	データベース
	ページ	ページ	対	ページ	ネットワーク
	収入	サイト	サイト	サイト	URL
	個人	手法	人	人	細分
	web	人	それ	件	化
	AdSense	店	偽装	一	風
	google	そこ	付け	二千	その他

表5: コサイン類似度による代表トピック

トピック	①	②	③	④	⑤
要素	キーワード	キーワード	キーワード	キーワード	セクション
	広告	広告	偽装	広告	AdSense
	検索	検索	索引	検索	収入
	エンジン	エンジン	登録	二千	者
	ページ	ページ	ページ	三千	公開
	google	サイト	サイト	google	個人
	表示	リンク	リンク	一	日本円
	順位	例	順位	例	アメリカ
	右	仕組み	付け	円	円
	付け	これ	サーバ	上位	web

図4に新たにトピックを1件追加し，デンドログラムを作り直したものが図5である．①〜⑤は図4とトピックの内容が近いクラスに振ったものであり，クラスに所属するトピック数が大きく変動していることから，データ数を増やす必要があると考えられる．

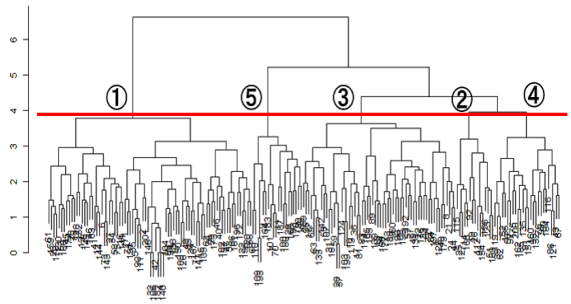

図 5: トピック 200 件のに新たに 1 つトピックに追加したデンドログラム

表 6: 字幕とクラスの対応

字幕番号	字幕	スライド番号	クラス
19	索引サイトは今のように、分類…	5	②
20	例えば、ショッピングというよ…	5	②
21	その他も同じ様に、大分類、小…	5	⑤
22	これに対して、検索エンジンと…	6	②
23	それで、ある量まで蓄積した段…	6	③
24	順位付けができたらば、検索者…	6	②

4.4 字幕とクラスの対応

講義中の字幕 119 件がどのクラスに分類できるか調査した．講義中の字幕の一つに,「順位付けができたらば、検索者があるキーワードで検索を入力した場合に順位付けの高いものから表示をする、というようなつくりになっています。」があり，この字幕から名詞を取り出すと「順位, 付け, 検索, 者, キーワード, 検索, 入力, 順位, 付け, 表示, つくり」となる．この字幕は図 4 におけるトピック 1（サイト, 検索, 順位, リンク, 付け, 情報, キーワード, これ, ページ, エンジン）に分類され，図 4 におけるクラス ② にあたる．

表 6 はスライドの変わるときの字幕とクラスの対応を一部取り出したものである．6 件の字幕のうち，4 件がクラス ② に分類されていることからスライドが切り替わっても同じような話題が展開されていると考えられる．このようにクラスと字幕全体を対応を表 7 に示す．③ のトピックにかなりの比重が置かれていることから，この講義の中心となるトピックだと考えられる．それに対し比重の少ない ④ や ⑤ のトピックは特徴的な内容を捉えたトピックだと考えられる．

5 まとめと今後の課題

本研究では検索語とその共起語の出現頻度を共起距離による関連度を表現した散布図の作成，講義スライドの検索語と共起語の関連性について調査，キーフレーズ抽出による特定の範囲におけるトピックの調査，LDA による講義全体のトピック分析を行った．それぞれについての課題を以下に示す．

表 7: クラスにおける字幕の分布

クラス	所属数
①	18
②	28
③	61
④	3
⑤	9
合計	119

(1) 散布図と関連性の調査では，調べたい単語の出現頻度だけでなく，それに関係する単語やトピックとその関連度が得られるようになった．しかし，今回作成したシステムは人が見てその解釈を行うものであり，機械的な処理による結果も出せるようにする必要がある．

(2) キーフレーズ抽出では特定の範囲におけるトピックを見つけることができた．重要な単語やそのスコアが得られることから，それぞれの区間におけるトピックが想像しやすいものとなった．

(3) LDA による調査ではセクションの内容を推察できるトピックが得られた．クラスタリングした結果を利用し，様々な範囲におけるトピックを調査できるようにする必要がある．

参考文献

[1] 北川，大西: “対面講義と e-learning(LMS + VOD) を併用した講義形式の実践と分析”, 日本教育情報学会学会誌 Vol.22 No.3 pp.57-66, 2007.

[2] 中村, 椎名, 北川: “ベータ分布による VOD 講義の話題区間の検出”, 第 64 回電気・情報関連学会中国支部連合大会論文集 pp.321-322, 2013.

[3] 北, 津田, 獅々堀: “情報検索アルゴリズム”, 共立出版, 2002.

[4] Yahoo!デベロッパーネットワーク, http://developer.yahoo.co.jp/sample/jlp/sample3.html

[5] Willi Richert, Luis Pedro Coelho: ”実績 機械学習システム”, 株式会社オライリー・ジャパン, 2014.