4Q-07

議事録閲覧支援のための議論のクラスタリング

安藤 聡志† 井上 慧†† 松原 茂樹†† 長尾 確††
†名古屋大学 工学部電気電子・情報工学科
††名古屋大学 大学院情報科学研究科

1 はじめに

研究活動では，議論は頻繁に行われる．そして，議論の内容は議事録を閲覧することで想起できる．

しかし，議事録は1つの会議ごとにまとめられており，それが閲覧者にとって常に読みやすい形態であるとは限らない．実験方法など，特定のトピックについて，議事録にまたがって閲覧したい場合には必ずしも適さない．

本論文では，複数の議事録をそれぞれ内容の観点から分割し，内容ごとに分割された議論のクラスタリングを行う手法の提案をする．提案手法では，クラスタリングで用いる議論内容の特徴情報として，書記が記録したテキストに加え，議論で用いられたスライドの情報を使用する．スライドを用いる議論では，その内容がスライドの内容に影響を受けるためである．

2 議論セグメントとクラスタリング

ディスカッションマイニング [1] とは，映像，音声，テキスト，議論構造などのメタデータを獲得・記録することで，再利用可能な会議記録を作成する技術である．

ディスカッションマイニングによって作成された記録は会議コンテンツと呼ばれ，議論セグメントと呼ばれる単位で構成される．会議中の発言者は議論セグメントを作成することで新たな話題について話すことを明示し，議論する．

議論セグメントには，会議中の発言を書記が記録したテキストが含まれる．また，会議コンテンツには会議中に用いられたスライド情報が含まれており，発言時に表示されているスライドと対応付けられている．

本研究では，議論セグメントによって分割された会議コンテンツに含まれる話題ごとの内容を，議論セグメントが含む情報を用いてクラスタリングすることを目的とする．図1に提案手法の概要を示す．

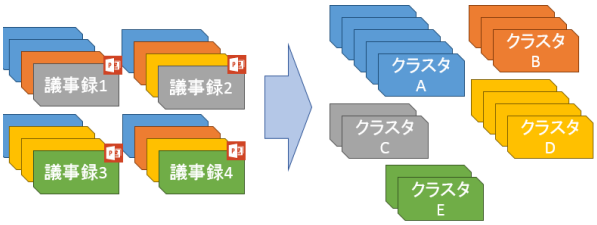

図 1: 議論セグメントのクラスタリング

3 議論セグメントのクラスタリング手法

本研究では，議論セグメントを文書とみなし，クラスタリングを行う．

これまでに様々な文書クラスタリング手法が存在するが，いずれも文書の特徴情報を利用する．本研究における独自の特徴情報として，発表時に用いられたスライド情報を使用する．スライドを見ながらの議論では暗黙のうちにスライドに書いてある内容が発言で省略され，発言内容を書記が書き記したものだけでは議論セグメント，すなわち文書の特徴情報が不足すると考えられるためである．

ベースとなるクラスタリング手法として Non-negative Matrix Factorization(NMF)[2] を利用する．NMF は次元縮約法の一種であり，類似した手法である Singular Value Decomposition(SVD) に基づく手法と比較して，文書ごとの各クラスタへの帰属度を示す行列 (以下，結果行列) が疎になりやすい．

議論セグメントの特徴として，1つの議論セグメントが複数の話題からなることは少ないことが挙げられる．また，文書クラスタリングの結果行列が疎になる場合，ある文書が異なるクラスタに同時に属することが少なくなるということが言える．よって，複数の話題を持ちにくい議論セグメントの理想的なクラスタリングにおいては，ある議論セグメントが異なるクラスタに同時に属することが少なくなると予想されるため，結果行列が疎になる方が理想的なクラスタリング結果に近いと考えられる．

NMF は与えられた行列 X から以下の等式を成立させる行列 V を求めるアルゴリズムである．

$$X = UV^T \tag{1}$$

ここで，X は文書群全体の特徴を表す行列であり，V は行が文書，列がクラスタに対応した，文書ごとの各クラスタへの帰属度を示す結果行列である．

議論セグメントのクラスタリングを行う上で各議論

Discussion Clustering for Supporting Minutes Browsing
† ANDO, Satoshi(ando@nagao.nuie.nagoya-u.ac.jp)
†† INOUE, Kei(kinoue@nagao.nuie.nagoya-u.ac.jp)
†† MATSUBARA, Shigeki(matubara@nagoya-u.jp)
†† NAGAO, Katashi(nagao@nuie.nagoya-u.ac.jp)
Dept. of Information Engineering, School of Engineering, Nagoya University(†)
Graduate School of Information Science, Nagoya University(††)

セグメントに対応する特徴ベクトルの導出が必要となる．特徴ベクトルの各要素には議論セグメント内に出現した各名詞の tf-idf 値を用いる．具体的には発言内容を書記が記録したテキストから抽出して使用する．また，対応するスライド中に出現した名詞も議論セグメント内に出現した名詞として含めることで，スライド情報を議論セグメントの特徴情報として活用する．

提案手法の手順をまとめると以下のようになる．

1. 議論セグメントごとの特徴ベクトルを書記テキスト情報，スライド情報を用いて導出
2. 全ての議論セグメントの特徴ベクトルをまとめて特徴行列 X とする
3. NMF を用いて行列 X から結果行列 V を導出
4. 行列 V から各議論セグメントがどのクラスタに属しているかを閾値をもとに判断する

4 評価実験

3 章のクラスタリング手法を実装し，評価実験を行った．本実験ではクラスタ数を 8 と定めた．

4.1 実験方法

クラスタリングの対象となる会議コンテンツは著者らの研究室で行われたゼミ発表で作成されたものを用いた．正解データは該当するゼミの発表者が作成した．すなわち，各発表者自らが，担当した 3 つの会議コンテンツに含まれる全ての議論セグメントを 8 つのクラスに手動で分類した．各発表者が担当した議論セグメント数の平均は 59 であった．

クラスタリングの評価基準として純度 (P) を用いた．

$$P = \frac{1}{N}\sum_{i=1}^{k} max_h|C_i \cap A_h| \qquad (2)$$

C_i は結果のクラスタを，A_h は正解データのクラスタを示す．N は議論セグメントの合計数である．純度はどれだけ不適切な議論セグメント群を同一のクラスタに割り当てていないかを表す指標であり，クラスタリング結果としての質の良さを示す．

ベースラインはスライド情報を使用せずに行ったクラスタリング結果とし，提案手法と純度を比較した．

本手法ではランダムな初期行列をプログラムに与える必要があるが，初期行列の違いがクラスタリング結果に影響する．よって，クラスタリングを 20 回試行し，その平均を評価軸として用いた．

4.2 実験結果

実験結果を表 1 に示す．

表 1: 評価実験の結果

ベースラインの純度 (%)	65.9
提案手法の純度 (%)	68.7

結果の通り，提案手法を用いることで全ての実験対象とした会議コンテンツのクラスタリングにおいて純度の向上が見られた．また，ベースラインと提案手法の純度に統計的有意差があるかどうか有意水準 5%で t 検定を行ったところ，有意差があることが認められた ($p = 0.000054$)．

5 おわりに

本論文では発表時に用いられたスライド情報を使用することで，発表議事録中の議論内容のクラスタリングをより高精度に行う手法を提案した．

今後の課題としては，会議コンテンツ中にはスライド中の文章へのアノテーション情報や，議論セグメント内部の発言構造など，クラスタリング能力向上に利用できると思われる情報がスライド情報以外にも含まれているため，それらを利用してさらなる能力向上を図ることが考えられる．

参考文献

[1] 土田貴裕, 大平茂輝, 長尾確，対面式会議コンテンツの作成と議論中におけるメタデータの可視化，情報処理学会論文誌，Vol.51，No.2， pp.404-416， 2011

[2] X.Wei，L.Xin，G.Yihong，Document clustering based on non-negative matrix factorization，ACM SIGIR，pp.267-273，2003