ニュースにおける異メディア間の差異情報抽出手法

11571027　大中　ちづる（灘本研究室）

あらまし：本研究では，同じテーマのニュースに対して新聞の社説と Twitter を比較し，一方のメディアにしかない情報(差異情報)を抽出する手法の提案を行う．具体的には，まずユーザがニュースのテーマを入力する．次に入力したテーマに対して社説とツイートを取得し比較を行う．その比較した結果から差異情報を抽出する．

1. はじめに

現在，新聞やテレビ等のマスメディア，Twitter[1]や LINE[2]等のソーシャルメディア，Google[3]や Yahoo![4]等のニュースサイトなどの数多くのメディアが存在している．また，近年ではソーシャルメディアやニュースサイトのようなインターネット媒体から検索を行い，ニュースについての詳細を知ることが多い．それに伴い，ニュースに対して様々な意見を持つことが可能になっている．

一方，ニュースに対する意見が多様化することで，人は好きなサイトやニュースばかり見てしまうという傾向がある．これにより，フィルターバブルやエコーチェンバーが発生し，視野が狭くなり自分の考え方に固執してしまうという問題が生じている．

吉田ら[1]の研究では，各国の新聞記事同士を分析するシステムを提案している．これに対し，本研究では，ある同じテーマのニュースに対して異なるメディアから差異情報を抽出する手法を提案する．これにより以下の 3 つの効果が期待できる．

① ユーザが見ていないメディアからニュースに対する差異情報を抽出することで，ユーザの視野を広げる材料になる．
② メディア毎の情報を抽出し提示することにより，ユーザの情報を分析する能力の向上に繋がる．
③ ユーザの情報リテラシーの向上に繋がる．

本研究では，差異情報抽出の第一歩として新聞の社説と Twitter を対象とする．

尚，本論文では，ユーザが入力した単語をテーマとし，そのテーマのニュースに含まれている単語をトピックと定義する．

以下に提案手法の生成手順を示す．

(1) ユーザがニュースのテーマを入力する．
(2) 入力したテーマを含む社説とツイートを取得する．
(3) 取得した社説とツイートからそのニュースのトピックを抽出する．
(4) 取得した社説とツイートから各トピックに関連しない文書を取り除く．
(5) (4)で取得した社説とツイートからトピック毎の文書間の類似度を求める．
(6) (5)から差異情報を抽出する．

2. 差異情報抽出手法

2.1. 社説とツイートの取得

本研究では，社説は新聞社説まとめサイト[2]からユーザが入力したニュースのテーマに関する社説の本文を取得する．ツイートはユーザの入力したニュースのテーマをクエリとし，Twitter API を用いてそのテーマに関するツイートを取得する．また，社説とツイートでは情報量が異なるため，社説を取得する際は，社説毎ではなく，社説内の段落毎に取得する．

2.2. 社説とツイートからトピックの抽出

取得した社説とツイートには複数のトピックが含まれている為，そのトピックを抽出する必要がある．そこで，トピックモデルの 1 つである Repeated Bisection[3]法を用いて社説，ツイート各々のクラスタリングを行い，トピックを抽出する．

2.3. 各トピックに関連しない文書の除去

Repeated Bisection 法はハードクラスタリングの為，各クラスタ内の社説とツイートには，ニュースのテーマに関連しない文書を含むクラスタが存在する場合がある．このようなクラスタは事前に取り除く必要がある．そこで，秋山ら[4]の提案しているクラスタ内の文書の密度に着目し，文書の密度が低いクラスタはガーベージクラスタと見なし，削除する．

2.4. 差異情報の抽出手法

社説とツイートのクラスタ同士を総当たりでクラスタ間のコサイン類似度を求める．そしてどのクラスタに対しても閾値 α 以下の類似度を持つクラスタをそのメディアの差異情報として抽出する．ここで，本論文では予備実験により，閾値 α を 0.3 とする．

---

[1] Twitter:https://twitter.com/
[2] LINE:https://line.me/ja/
[3] Google:https://www.google.co.jp/
[4] Yahoo!:https://www.yahoo.co.jp/

例えば，「加計学園」をニュースのテーマとした社説とツイート例を表１に示す．また，表１におけるメディア毎の差異情報を表２に示す．

表１：社説とツイート例

社説	野党は疑惑の徹底追及をすべき
	安倍総理は説明責任を果たせ
ツイート	安倍総理は説明責任を果たせてない
	詐欺学園、補助金返せ！

表２：差異情報抽出例

社説	野党は疑惑の徹底追及をすべき
ツイート	詐欺学園、補助金返せ！

3. 差異情報の提示

差異情報の提示のユーザインタフェースを図１に示す．赤字はユーザの入力したテーマであり，左ウィンドウは社説の差異情報である．また，右ウィンドウはツイートの差異情報である．各々のトピックを社説は緑，ツイートは青で示している．図１の例は大阪万博を入力テーマとした例である．

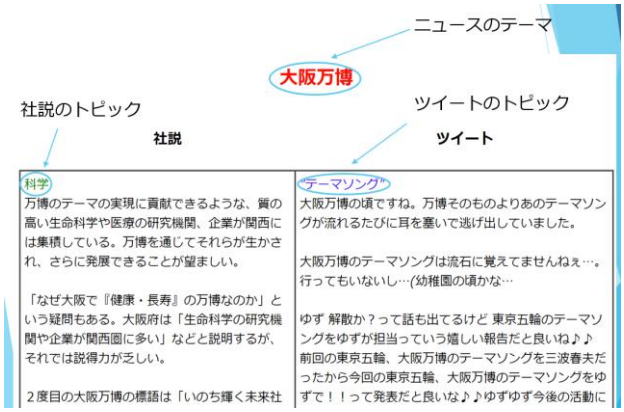

図 1:提示例(「大阪万博」)

4. 実験

提案手法の有効性を示すために実験を行った．

4.1. 実験条件

本研究では，社説を用いるためニュースのテーマを政治や経済などの時事関連を対象とした．そこで，現在，話題性のあるニュースのテーマで実験を行った．今回の実験では「大阪万博」と「カルロスゴーン」の２つのニュースのテーマで実験を行った．正解データは目視で判定した結果を用いた．

4.2. 実験結果

実験結果の適合率は，大阪万博は社説が 0.48，ツイートが 0.38 で，カルロスゴーンは社説が 0.21，ツイートが 0.56 であった．また，２つのテーマに関する社説とツイートをトピック毎にクラスタリングした結果の一例を表３に示す．

表３：トピックの抽出例

テーマ	トピック	
	社説	ツイート
大阪万博	松井 ギャンブル テーマ	２兆 活性 夢洲
カルロスゴーン	虚偽 責任 拘留	特捜 支配 組織

4.4. 考察

社説では，ニュースのテーマに直接関する内容が多く見受けられ，Twitter では，ニュースのテーマと関連性がある内容が多く見受けられた．また，どちらのメディアにもニュースのテーマに関連しない内容が含まれていた．これは，文書間のコサイン類似度の閾値を決定する際，各クラスタ内の文書内容を目視で判断し閾値を決定したため，設定した値では，ニュースのテーマに関連しない文書の除去が不十分であったと考えられる．

5. まとめと今後の課題

本研究では，同じニュースのテーマに関する社説とツイートを収集し，差異情報を抽出する手法の提案を行った．さらに，提案した抽出手法の実験と抽出した結果の提示も行った．今後の課題として，目視によって決定した閾値の改善を行う必要がある．時事関連以外のニュースのテーマでの実験を行う必要があることも今後の課題である．

参考文献

[1] 吉岡 真治，”NSContrast:世界ニュース比較分析システムの実験的評価”，言語処理学会 第15回年次大会，pp.494-497，2009.

[2] 新聞社説まとめサイト http://shasetsu.seesaa.net/

[3] Ying Zhao and George Karypis. Comparison of agglomerative and partitional document clustering algorithms. Technical report, Department of Computer Science,University of Minnesota, Minneapolis, MN 55455,2002.

[4] 秋山和寛，熊本忠彦，灘本明代，”ツイートからの多次元感情抽出手法の考察”，第８回ソーシャルコンピューティングシンポジウム(SoC2017)，信学技報，Vol. 117，No. 108，DE2017-3，pp. 11-16，2017.