6P-02

辞書を用いたクラスタリングとその多重ラベル付け

横沢 薫† 三浦 孝夫†

法政大学理工学部創生科学科

東京都小金井市梶野町 3-7-2

1. 前書き

近年,インターネットの普及やスマートフォンの出現により,Twitter,Blog などを介して誰でも情報や意見の発信が可能になった。その結果,膨大な情報量となり,検索が困難になった。

本研究では, 大量のデータを素早く理解するため,データにラベル(文書集合の概括的な性格)付けをし,欲しい情報の検索を容易にする手法を提案する。クラスタリングを用いて大量の文書を類似集合に分け,それらの特徴を表すキーワードを抽出する。本提案のアイデアは、語義曖昧性による特徴の解釈の違いを避ける為,キーワードを,辞書を用いて語義を特定することにある。辞書には語義ごとに SynsetID が付与されていると仮定し,それらを使って文書集合の特徴を表現する。上記の過程を経て文書集合に対してラベルを選定する。

2. 密度クラスタリング

クラスタリングは類似したデータを同じ集合に集約する手法である。このため距離概念が必要となる。代表的に k-means 法では文書ベクトルの余弦類似度を用いるため,比較的離れたデータや,類似していないデータを同じ集合に含める可能性がある。本稿ではこの問題を避ける為,データ間の密度が一定以上となるクラスタを形成する *DBSCAN* を使いることにより,クラスタ内の分散が比較的小さくなり,より類似したデータの集合を形成しやすくなる[1]。

3. 概念辞書と語義曖昧性解消

語義の曖昧性を解決するため,本研究では Wordnet を用いる。Wordnet は語の意味、品詞や例文が格納された概念辞書である。Wordnet に含まれる各語(見出し語)はその語義ごとに意味番号(SynsetID)が割り当てられ,また説明文(語釈文)が対応する。同じ語が複数の意味を有する多義語(“paper”には紙と論文の意味)では,この SynsetID は複数個対応し,異なる語が同じ意味を有する同義語(“book”,”reserve”は共に予約するの意味)では,SynsetID を共有する。語義の数だけ SynsetID が存在している。

語義曖昧性を解消するため, *Lesk アルゴリズム*が知られる[1]。格納された語釈文と,語義を特定したい単語を含む文書との類似度を計算し,類似度が高い語釈文の SynsetID を選択する。この前提は「語の近くにある単語は,その単語と共通の話題を持つ傾向がある」という仮定に基づいており,語義を特定したい単語の語釈文とその単語付近を比較することでその単語の語義を特定する,語義曖昧性解消のためのアルゴリズムである。例えば”book”を含む文書と,”book”の”本”の意味の語釈文と”予約”の意味の語釈文との単語出現頻度(Term Frequency,TF)に基づいた類似度を用い,”book”がどの意味かを推定する。

4. 提案手法

ここではニュース記事集合を想定する。ラベル付けは、”キーワード抽出”と”ラベル付け”の 2 ステップからなる。キーワード抽出のため,形態素解析などで前処理した文書ベクトル集合に DBSCAN を施し,密度の高いクラスタを形成する。クラスタごとに高い出現頻度(TF)の語上位を抽出し,語義を特定する。頻出語が属するクラスタと,Wordnet が含むその語釈文との類似度を余弦類似度で求め,高い方の SynsetID をそのクラスタのキーワードとする。

キーワードとして設定された SynsetID を多く含むクラスタ内の複数文書をクラスタから選び,その見出し(Headline)をラベルとする。複数文書を選ぶので,1 つのクラスタに対し,複数のラベルが付与される。なお,見出しの存在しない記事については,記事の 1 文目の文章を代用する。

5. 実験

5.1 実験準備

本研究では Reuter News Courpus(Reuter-21758)の記事 12555 件(見出し語 : 136485 語,延べ語:2650805 語)を対象とする。このデータから特徴語を抽出するため,不要語(Cornell University)の除去,単語を語幹化するステミング(Porter Stemming)を施し,更に,Zipf の法則に基づき,中頻度語のみを抽出する(見出し語:43498 語,延べ語:1046393 語)。

Cluster Labelling with Dictionary

† Kaoru Yokosawa and Takao Miura

Dept.of Advanced Science, HOSEI University

DBSCAN パラメタを Eps=5, minPts=3 とする。実験結果の評価基準として Micro Precision を用いる。正解ラベルと推測ラベルを設定し, 正しく推測できた数で適合率を評価する。正解とみなすのは, 正しいラベルを正しいと判断した場合, または間違っているラベルを間違っていると判断した場合とする。複数ラベルが付与されている文書については, 部分的に正解している場合も正解とみなす。この正解数の総文書数との比率を適合率とする。すべての推測ラベルについて算出し, 合計したものをクラスタに付けられたラベル数(本研究では 5 個)で割ることで, クラスタ単位の適合率を算出する。全クラスタごとに算出して平均をとり適合率とする。

5.2 実験結果

DBSCAN の結果, 11 個のクラスタが形成された。尚, このうち 0 番目のクラスタは, 密度の高いところから離れ, どの集合にも属さないような例外データであり, ここに割り振られたデータは考慮しない。残りの 1 番目から 10 番目までのクラスタから, TF の多い上位 5 語を選択する。実際に得られたキーワードを表 1 に示す。

クラスタ番号	キーワード				
1	pct	said	market	real	cut
2	said	billion	percent	market	cut
3	said	mph	rate	one	cut
4	said	cut	percent	came	billion
5	said	loss	net	index	billion
6	said	billion	percent	race	million
7	said	loss	million	percent	one
8	said	percent	billion	loss	would
9	loss	million	mln	percent	said
10	loss	said	would	mln	percent

表 1 得られたキーワード

Lesk アルゴリズムに従い, 類似度が大きい SynsetID を, その語義と推測し, クラスタのキーワードとする。次に, それらの SynsetID を多く含む記事上位 5 つをクラスタ内から選ぶ。記事内の語を Lesk アルゴリズムに従い, すべて SynsetID を表し, キーワードとなる SynsetID を多く含む文書を選択する。実際に得た文書とクラスタ番号を表 2 に示す。上記の過程を経て得られた文書の見出しをクラスタの推測ラベルとする。実際に得られたラベルを図 1 に, 適合率を表 3 に示す。

クラスタ番号	記事番号				
1	6645	7421	8143	11028	11039
2	6645	8143	10398	10629	2984
3	6645	7421	8143	8563	11053
4	11053	3171	5175	8143	10398
5	12242	12416	11028	254	4452
6	6825	1930	4452	5175	5278
7	11053	12416	10629	11028	11393
8	12416	9682	10089	11039	603
9	1	254	396	477	903
10	11053	6825	10629	427	2984

表 2 得られた記事番号

```
1@The Bundesbank lived up to its reputation of surprising financial markets wit
ay a move which has thrown the EMU ball back to its European partners financial
Six months to June 30 (in millions of marks unless stated) Group net loss 19
Public Service Electric and Gas Co said on Friday it had proposed a pilot progr
mers to buy the fuel from suppliers other than itself
Results of soccer matches in the Danish super league over the weekend
South Africa's ruling African National Congress said on Monday it would hold a
```

図 1 実際に得られた推測ラベル

クラスタ番号	正解数(ラベル)	正解数(全体)	文書数	適合率
1	1186.321561	95.1387548	61	0.204416
2	815.0996747	68.459403	35	0.340181
3	653.5337516	46.2266251	32	0.4418
4	463.9170287	35.6579178	24	0.542092
5	461.393983	38.7919284	24	0.495586
6	324.8537466	30.4132833	18	0.593406
7	505.2106098	49.0248371	37	0.278519
8	599.0208849	52.2244215	30	0.382338
9	725.8726967	61.6454073	28	0.420535
10	345.7414761	35.6515812	25	0.387912

表 3 適合率

クラスタごとの適合率平均は 0.41 である。

ベースラインとした k-means では 20 クラスタが生成され平均適合率は 0.152 となる。これらの評価では, k-means 法は文書数を単位とし, DBSCAN の場合は推測ラベルの見出しを文書が有する確率として, 推測ラベルとなる見出しを有する文書と, クラスタに含まれる文書との TF を基にした類似度を計算し, それを単位としている。

5.3 考察

表 3 から, 提案手法のクラスタごとに適合率に差がある。表 4 に示した表から, クラスタに含まれる「キーワードの延べ語数の逆数」と比較してみると, クラスタごとの適合率はキーワードの延べ語数に依存していることが読み取れる。正しいラベルを抽出できるか否かはキーワードの延べ語数で決まっている可能性がある。

6. 結論

密度クラスタリングを施し, 辞書を用いて意味まで推測し, SynsetID を使ったマルチラベルの付与はベースラインと比較すると適合率は 2 倍以上であった。しかし, 適合率は 41%と, 正解は半分以下であり, 実用性にはやや欠ける。SynsetID に, TF を基にした重み付けや, 別のクラスタリングの手法の改善が考えられる。

クラスタ番号	適合率	1/key
1	0.542092	0.004762
2	0.340181	0.001577
3	0.593406	0.005263
4	0.420535	0.003534
5	0.382338	0.003521
6	0.387912	0.003378
7	0.495586	0.003175
8	0.4418	0.006098
9	0.278519	0.001587
10	0.204416	0.004115

表 4 適合率とキーワードの延べ語数

参考文献

[1] Han, J. et al: DataMining, Morgan-Kauffman, 2011