5M-2　類似文書クラスタリング手法による新聞記事分野コード推定実験

鈴木 雅実‡　村松 茂樹†　松本 一則†　井ノ上 直己†　橋本 和夫†

通信・放送機構‡　　KDD研究所†

1. はじめに

本研究の目的は，教育利用な情報（コンテンツ）の発見を支援するような情報検索／フィルタリング手法を提案し，その有効性を評価することにある。インターネット等を通じてアクセス可能となった大量の情報資源の中から，教育目的に利用可能な文書を発見することは容易ではない。そこで，利用実績のある（または利用価値の高いとみなされた）文書に対して，コンテント・ワードの出現頻度分布において類似する文書を導く，類似文書検索（岩山・徳永[1]では，文書連想検索）手法に基づく検索方式を検討している。本稿では，新聞記事素材を対象に，この方式のベースとなる類似文書クラスタリング手法の分類精度を測定するための予備実験を行なった結果を報告する。

2．実験方法

2.1 対象文書

朝日新聞 1999 年の記事から選んだ，6 分野で各 100 件の記事を用いた。ここで 6 分野は，政治／経済／社会／外報／運動／学芸であり，正確にはその分野の記事作成を担当する部門を示す出稿部により，予めラベル付けされているものである（各分野内では原則として連続した期間中の記事）。

2.2 実験手順

本実験では，確率モデルに基づく文書類似度計算モデル（後述）を用いて，600 件の記事の相互類似度を計算する。結果として出力される階層木を分割することにより，n 個のクラスタを生成し，生成されたクラスタ内の記事のラベルの分布状況を調べる（クラスタ数を変化させた場合の精度も比較）。

Estimating Newspaper Genre Code using a Similarity-based Document Clustering Method

Masami Suzuki (TAO), Shigeki Muramatsu, Kazunori Matsumoto, Naomi Inoue and Kazuo Hashimoto (KDD R&D Laboratories, Inc.)

2.3 類似度計算モデル

今回，類似度計算に用いたモデルは，確率モデルである。種々の計算モデルが提案されているが，岩山・徳永[1]が示したように確率型モデルは，類似度(距離)の値に対する意味が明確であること，クラスタリング時の精度が 良さそうなことが知られていることから，文書間の類似度 $Sim(d_x, d_y)$を，次のように計算する。ここで，$P(c|d_x)$は，文書 d_x と文書集合 c が与えられた時の類似度である。

$$Sim(d_x, d_y) = \frac{P(\{d_x, d_y\}|d_x) \cdot P(\{d_x, d_y\}|d_y)}{P(\{d_x\}|d_x) \cdot P(\{d_y\}|d_y)}$$

なお，$P(c|d_x)$ の計算方法は文献[1]に示されている。

3．実験結果

分割数＝20 の場合のクラスタ分類結果を表 1 に示す。この分割は，木構造（2 分木）をボトムアップ的に各クラスタ内のリーフ（葉＝文書）数を 16 ～ 32 のオーダーとなるように，自然な切断を行なった場合に相当する。この結果を見ると，同一クラスタ内のラベル（記事コード）が分散しているクラスタがあり，さらに分割することにより，クラスタとしてのまとまりが向上すると考えられる。

そこで，クラスタ・セグメント内で最も多い分野ラベルが過半数に満たないクラスタをさらに 2 分して，より妥当と思われる分類結果を導いた（最終的に分割数＝30）。ただし，第 2 位のラベルが相対的に大きな割合を占める，クラスタ No. 10,17 については分割を行なった。なお例外として，クラスタ No. 8 については適切な分割が得られなかったため，そのままとした。新聞記事に特有な分野間の重なり等も影響していると考えられる。

分割数 20 の場合と 30 の場合を比較すると，各分野ラベル（記事コード）間の重なり率は，表 2 に示すように，すべての組み合わせにおいて分割数 30 の場合の方が低い数値となり，各クラスタの特徴(分布）がより鮮明になったことを示唆している。

実際，各クラスタ内で最も多い分野ラベルをそのクラスタ自身のラベルとみなした場合の正解率は，分割数 20 の場合では平均 53%，最高値が政治の 64%で，同率の 3 ラベルが存在するクラスタ No. 6 を考慮に入れた場合でも，社会(30+8)および学芸(38+8)が 50%に満たないのに対し，分割数 30 では平均 59%，すべての分野で 50%を越えた（最高値は政治 66%；社会 52%，学芸 61%）。

表 1 新聞記事のクラスター分類結果(分割数＝20)

Cluster No.	リーフ数	政治	経済	外報	社会	運動	学芸
1	32	0	0	3	6	0	**23**
2	32	2	0	3	11	1	15
3	24	0	0	0	4	**18**	2
4	19	4	3	0	7	1	4
5	31	2	3	6	12	0	8
6	32	3	5	0	8	8	8
7	21	6	1	2	4	6	2
8	32	3	**10**	4	4	7	4
9	32	13	5	6	4	2	2
10	32	1	0	1	3	16	11
11	27	**17**	0	2	2	3	3
12	32	**28**	0	3	1	0	0
13	30	0	**28**	2	0	0	0
14	31	8	6	**16**	0	1	0
15	33	6	5	**17**	2	3	0
16	34	2	6	3	11	4	8
17	32	0	17	0	11	0	4
18	32	2	8	**17**	3	0	2
19	30	2	3	2	1	**20**	2
20	32	1	0	13	6	10	2
合計	600	100	100	100	100	100	100

(注)ゴシック体のラベルのあるクラスタは再分割せず.

表 2 記事コード間の重なり率

	政治	経済	外報	社会	運動	学芸
政治	100	34	42	34	29	29
経済	**29**	100	38	43	27	32
外報	**37**	**37**	100	41	31	32
社会	**27**	**33**	**37**	100	42	69
運動	**22**	**23**	**24**	**32**	100	42
学芸	**22**	**23**	**31**	**48**	**33**	100

（注）対角線より右上が分割数 20，左下が分割数 30

さらに，クラスタ分割の妥当性については種々の評価方法があるが [4]，次のような対数尤度を用いると，分割数 20 の場合では $LL = -701$，分割数 30 では $LL = -566$ であった（すべて均等に分布したと仮定した場合の数値は -1075）。

＜対数尤度の計算式＞
クラスタリングで得られた m 個のクラスタ・セグメントを $S_1, S_2, \cdots S_m$，文書ラベルの種類を $L_1, L_2, \cdots L_n$ とし，セグメント S_i に含まれる L_j の個数を $N(i, j)$ とすると，
対数尤度 LL は

$$LL = \sum_{i=1}^{m} \sum_{j=1}^{n} N(i, j) \log P(i, j)$$

ただし，$P(i, j) = N(i, j) / \sum_{k=1}^{n} N(i, k)$

4．考察

前節の結果に見られる通り，表面的な基準でクラスタを生成したとしても，さらにクラスタを再分割する方が分野ラベル（記事コード）の推定をより正確に行なえる場合があることが分かるが，最適な分割を予測することは困難である。ただし，種々の基準を用いて，より妥当な分割を行なう試みもある[3]。さらに，クラスタ分割の適切性は，類似文書検索をクラスタ検索で実行した場合の精度にも当然影響するが，対象によっては網羅検索により良い精度が得られることが知られており，計算時間と精度の兼ね合いで最適手段を選ぶことになろう[1] [2]。今後は，この予備実験結果に対して新聞記事の特性等も考慮した分析をふまえた上で，教育情報の検索に適した手法提案を行なう予定である。

謝辞 本研究は，通信・放送機構(TAO)の直轄研究「学校における複合アクセス網活用型インターネットに関する研究開発」の一環として実施しているものである。指導頂く東工大・清水康敬先生をはじめ関係各位の支援と助言に感謝いたします。

参考文献

[1] 岩山真・徳永健伸："確率的クラスタリングを用いた文書連想検索"，自然言語処理，Vol.5 No.1，pp.101–118, 1998.
[2] 青木圭子，松本一則，橋本和夫："大量文書向けのクラスタリング手法の評価"，情報処理学会第 56 回全国大会，1998.
[3] 村松茂樹 他：MDL 基準を用いた文書集合の特徴化手法，電子情報通信学会総合大会，D-8-10, 2000.
[4] Anderberg："クラスター分析とその応用"，西田英郎監訳，内田老鶴圃，1991.