6N-04

類似貢献度に基づく文書検索結果の可視化分析法

尾崎　了祐†　　斉藤　和巳†
†神奈川大学大学院 理学研究科

1　はじめに

ビッグデータ技術の進展により, 膨大な文書データから, 有用な情報を利活用するニーズは益々高まっている [1]. 本研究では, 与えられた文書クエリの類似検索結果文書集合を可視化する課題に焦点を当てる. ところが, 著者らの知る限り, この課題に関する先行研究は少数で [2], あまり研究が進んでない状況である. 本稿では, 検索結果の各文書をノードと見なし, それぞれを類似貢献度ベクトルに変換し, k-最近傍グラフを構築した可視化において, 検索結果文章がクエリに対し, どのような説明語で類似するかを付与するとともに, グループ分けして配色する新手法を提案する. livedoor ニュースコーパスを用いた評価実験では, 可視化例に基づく定性評価とともに, 説明語となる単語週数の定量評価により, 提案法の有効性を検証する.

2　提案手法

総数 N の文書集合を $\mathcal{N} = \{1, \cdots, n, \cdots, N\}$ とし, 語彙数 D の出現単語集合を $\mathcal{D} = \{1, \cdots, d, \cdots, D\}$ とする. また, 各文書 n とクエリ文書の D-次元特徴ベクトルを $\mathbf{x}_n$ と $\mathbf{q}$ とし, 文書間でのコサイン類似度を求めるため, 特徴ベクトルはノルムが 1 に正規化されているとする.

提案可視化分析法の手順を以下に述べる. まず, クエリ $\mathbf{q}$ に対し, 文書集合 $\mathcal{N}$ から h-最類似文書集合 $\mathcal{R}(h) \subset \mathcal{N}$ を求める. 次に, 類似文書 $n \in \mathcal{R}(h)$ に対し, 特徴ベクトル $\mathbf{x}_n$ から単語 d の類似貢献度を

$$y_{n,d} \quad = \quad \frac{q_d \times x_{n,d}}{\sum_{d' \in \mathcal{D}} q_{d'} \times x_{n,d'}} \tag{1}$$

で求め, D-次元類似貢献度ベクトル $\mathbf{y}_n$ を構築する. ここで, $y_{n,1} + \cdots + y_{n,D} = 1$ となる. そして, 類似貢献度ベクトル $\mathbf{y}_n$ 間の距離に基づき, $\mathcal{R}(h)$ をノード集合として k-最近傍 (NN: Nearest Neighnour) グラフ $G(\mathcal{R}(h), \mathcal{E})$ を構築する. ここで, $\mathcal{E}$ はノード間に張られたリンク集合を表す. 最後に, 単語 d で類似貢献度が最大となるノード (文書) 集合を

$$C(d) \quad = \quad \{n \in \mathcal{R}(h) \mid d = \arg\max_{d' \in \mathcal{D}}(y_{n,d'})\} \tag{2}$$

として求め, そこに属すノード群を同色にし, 単語 d を説明語として表示可能にした k-最近傍 $G(\mathcal{R}(h), \mathcal{E})$ の可視化結果を出力する.

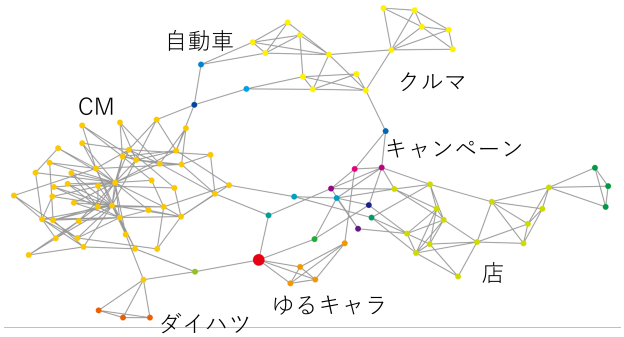
比較評価のため, 式 (1) でクエリとの類似貢献度を考慮せず, 類似文書 $n \in \mathcal{R}(h)$ それぞれの類似度として, $z_{n,d} = x_{n,d}^2$ を求め, D-次元ベクトル $\mathbf{z}_n$ を構築し, 説明語を $\arg\max(z_{n,d'})$ で求める方法を比較法と呼ぶ. なお, 先行研究 [2] では, クエリを考慮した類似貢献度ベクトルを用いず, 検索結果文書間の類似度を求める点で比較法の範疇と見なせる. また, 提案法において, グラフ構築に最小全域木 (MST: Minimum Spanning Tree) を用いる方法を MST 法と呼び比較評価する.

3　実験による評価

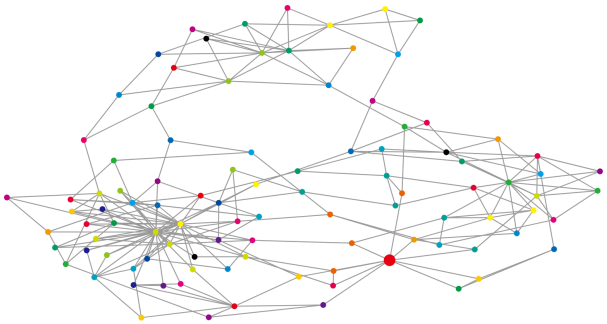
本実験では，文書数 $N = 7,367$ の livedoor ニュースコーパスを文書集合 $\mathcal{N}$ とし, MeCab [3] で出力される形態素をすべて単語として扱い, 語彙数は $D = 89,993$ とした. なお, グラフ可視化には, ばねモデル [4] を採用した. まず図 1 に，提案法, 比較法, 及び, MST 法での可視化結果の例を示す. ここで比較的大きな赤丸で描いたクエリは, 文書集合 $\mathcal{N}$ から随意に選択した, ゆるキャラ「カクカクシカジカ」によるダイハツ車 CM に関する文書である. そして, 類似文書数を $h = 100$ に設定し $\mathcal{R}(h)$ を求めた. ノード配色については, 類似文書 $\mathcal{R}(h)$ を類似度で降順に並べ, 同じ説明語となるノード群は同色となるよう, CMYK カラーシステム 24 色をサイクリックに割り当てた.

図 1(a) に示す提案法での可視化結果からは, CM, 店, クルマを説明語とし, 比較的密結合する多数のノード群とともに, 比較的大きな赤丸のクエリ近くに, 少数ながらダイハツ, ゆるキャラを説明語とするノード群も見て取れる. その他の単一説明語としては, 自動車, キャンペーンなどが出現する. これら結果より, ある観点で類似する文書数がどの程度かなど視覚的に把握可能と言える. これに対し, 図 1(b) に示す比較法での可視化結果では, 殆どのノード結合は異なる説明語 (配色) となり, 全体的な類似構造把握の点などでも限界が見受けられる. 一方, 図 1(c) に示す MST 法での可視化結果と比較すれば, ノード配色は提案法と同一であるが, 密結合する部分構造把握の点などで, 提案法に望ましい性質が見て取れる.

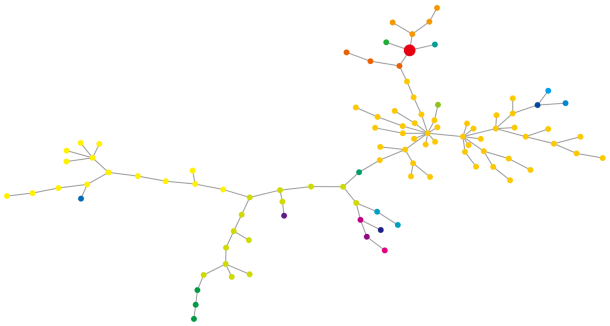
Visualizing Document Search Results Based on Similarity Contribution
†Ryosuke OZAKI　†Kazumi SAITO
†Kanagawa University



(a) 提案法



(b) 比較法



(c) MST 法

図 1: 可視化結果の例

次に, 類似貢献度ベクトルを用いる効果を定量評価する. まず, h-最類似文書 $\mathcal{R}(h)$ における各単語 d での類似貢献度の期待値を $u_d = \sum_{n\in\mathcal{R}(h)} y_{n,d}/h$ で求め, これら値を降順 $u_{d(r)} \geq u_{d(r+1)}$ に並び替え, その累積値を $v_s = \sum_{r=1}^{s} u_{d(r)}$ で求める. よって, $v_D = 1$ となる. 図 2 に, クエリ文書を文書集合 $\mathcal{D}$ のそれぞれに設定したときの累積値 v_s の期待値を示す. ここで, p050, p100, p200 は上述した提案法で検索文書数を $h = 50, 100, 200$ に設定した結果であり, c050, c100, c200 には, 類似貢献度期待値を特徴量での期待値 $u'_d = \sum_{n\in\mathcal{R}(h)} z_{n,d}/h$ に置き換えた比較法において $h = 50, 100, 200$ に設定した結果である. 図 2 より, 検索文書数 h に依らず, 提案法では上位 $r = 10$ 単語程度で累積貢献値が 0.5 程度になるのに対して, 比較法では, そのために上位数百単語程度が必要

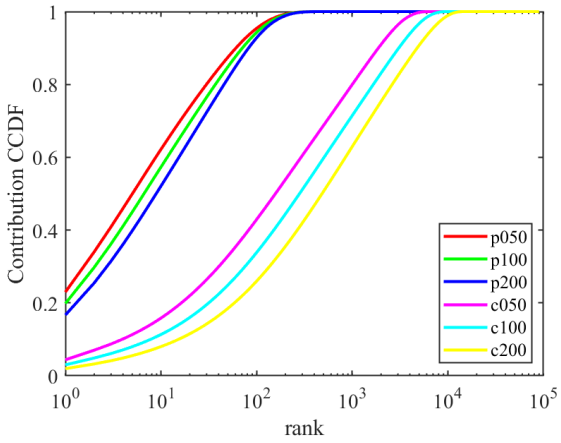


図 2: 貢献度期待値の累積分布

になる. すなわち, 提案法では, 比較的少数の単語に類似貢献度が集中するので, 妥当なノード配色や説明語選択の実現が期待できる.

4 おわりに

本稿では, 与えられた文書クエリの類似検索結果文書集合を可視化する手法を提案した. すなわち, 検索結果の各文書をノードと見なし, それぞれを類似貢献度ベクトルに変換し, k-最近傍グラフを構築した可視化において, 検索結果文章がクエリに対し, どのような説明語で類似するかを付与するとともに, グループ分けして配色する新手法を提案した. livedoor ニュースコーパスを用いた評価実験では, 可視化例に基づく定性評価とともに, 説明語となる単語数の定量評価により, 提案法の有効性を検証した. 今後の研究では, 提案法の有効性を幅広いデータで検証する.

参考文献

[1] 池田 弘行. 機械学習を用いた 文章データの解析・可視化技術. 東芝レビュー, No.5, pp.68-69, 2017.

[2] 小田 良治, 土井 章男. 大規模文書検索結果のクラスタリングと可視化. 情報処理学会第 66 回全国大会, 2004.

[3] T. Kudo, K. Yamamoto, and Y. Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

[4] T. Kamada and S. Kawai. An algorithm fordrawing general undirected graphs. Information Processing Letters. No.31, pp.7-15, 1989.