情報検索のためのトピック適応型単語クラスタリング

綱川 隆司 † 二宮 崇 ‡† 宮尾 祐介 § 辻井 潤一 §‡
† 東京大学 大学院情報理工学系研究科 コンピュータ科学専攻
‡ 科学技術振興機構 § 東京大学 大学院情報学環
{tuna, ninomi, yusuke, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

情報検索の分野において，クエリ（問い合わせ）に適合する文書を検索する際にキーワードマッチを用いると，クエリ中の単語を直接含まない文書を探し出すことができない．このため，クエリの各単語やその内容に関連する単語を新たにクエリに付け加えるクエリ拡張と呼ばれる手法が研究されてきた [1]．クエリ拡張手法の一つとして単語の共起情報を用いて自動的に構築されたシソーラスを用いる手法が挙げられる [2]．クエリに含まれる単語からシソーラスを引いて関連する単語をクエリに追加する．

このようなシソーラスは検索対象とする文書集合，あるいはその他の大量のテキストなどから自動的に構築されるが，その際にテキストがどのようなトピックを持っているかを考慮しないと，シソーラスに細かいトピックや分野といった情報が含まれず，クエリが求めるトピックに合致した単語が選ばれるとは限らなくなる．

本論文では文書クラスタリングを用いていくつかの文書クラスタを形成し，各文書クラスタごとに単語クラスタリングを行うことにより各文書クラスタのトピックに特化した単語クラスタを生成することを目指す．この際疎データ問題に対処するために，文書がクラスタに帰属する度合いを求めるファジィクラスタリング法を適用する．

さらに，単語クラスタを生成するための単語間の類似度を用いて，情報検索のクエリ拡張手法の一つである概念ベースクエリ拡張 [4] に文書クラスタの考えを導入して，その有効性を考察する．

2 背景

2.1 Fuzzy c-means 法を用いた文書クラスタリング

Fuzzy c-means 法 [3] は，クラスタリングの対象 x_j が単一のクラスタに属するようにするのではなく，各クラスタ C_k への帰属度 $\delta_{jk} \in [0,1]$ を求めるファジィクラスタリングの一種である．

文書集合を $\{d_1, \ldots, d_N\}$，文書に含まれる単語集合を $\{t_1, \ldots, t_M\}$ とすると，文書 d_j の特徴を表す文書ベクトル $\vec{d_j} = (d_{j1}, \ldots, d_{jM})$ は以下のように定義される:

$$\begin{aligned} d_{ji} &= (1 + \log \mathrm{tf}_{ji}) \cdot \mathrm{idf}_i, \\ \mathrm{idf}_i &= \log(N/N_i). \end{aligned}$$

ただし，tf_{ji} は文書 d_j に単語 t_i が出現した回数[1]，N_i は単語 t_i が出現する文書の数とする．文書間の類似度 $\mathrm{sim}(\vec{d}, \vec{d'})$ は 2 つの文書ベクトル間の余弦で定義される:

$$\mathrm{sim}(\vec{d}, \vec{d'}) = \frac{\vec{d} \cdot \vec{d'}}{|\vec{d}||\vec{d'}|}.$$

この文書ベクトルと文書間の類似度を用いて，Fuzzy c-means 法により各文書 d_j のクラスタ C_k への帰属度 δ_{jk} を求める．Fuzzy c-means 法のアルゴリズムは以下の通りである．

入力: クラスタ数 c，ファジィ化パラメータ $m(>1)$[2]，文書ベクトル $\vec{d_j}(j = 1, \ldots, n)$.
出力: 帰属度 $\delta_{jk}(j = 1, \ldots, N;\ k = 1, \ldots, c)$，クラスタ中心 $\vec{v}_k(k = 1, \ldots, c)$.

1. 文書 d_j のクラスタ C_k への帰属度 $\delta_{jk}(j = 1, \ldots, N;\ k = 1, \ldots, c)$ の初期値を適当に定める．

2. 各クラスタの中心 $\vec{v}_k$ を以下の式で更新する．

$$\vec{v}_k = \frac{\sum_{j=1}^{N} {\delta_{jk}}^m \vec{d_j}}{\sum_{j=1}^{N} {\delta_{jk}}^m}.$$

[1] $\mathrm{tf}_{ji} = 0$ のときは $d_{ji} = 0$ とする．
[2] 本論文の実験では $m = 1.5$ とした．

3. 帰属度 $\delta_{jk}(j=1,\ldots,N;\ k=1,\ldots,c)$ を以下の式で更新する．

$$\delta_{jk} = \begin{cases} \left[\sum_{k'=1}^{c}\left(\frac{\mathrm{sim}(\vec{d_j},\vec{v}_{k'})}{\mathrm{sim}(\vec{d_j},\vec{v}_k)}\right)^{\frac{1}{m-1}}\right]^{-1} & (\mathrm{sim}(\vec{d_j},\vec{v}_k)>0) \\ 0 & (\mathrm{sim}(\vec{d_j},\vec{v}_k)=0) \end{cases}$$

4. クラスタ中心が変化しなくなるまで Step 2, 3 を繰り返す．

2.2 Similarity Thesaurus

Similarity thesaurus [5] は単語クラスタリングにおける単語類似度を定義する一手法であり，ある 2 つの単語 t,t' に対し，t の各文書における出現頻度のベクトルと t' の各文書における出現頻度のベクトルの余弦を t,t' 間の類似度とする．

単語 t_i の特徴を表すベクトル $\vec{t_i} = (w_{i1},\ldots,w_{iN})$ を次の式で定める．

$$\begin{aligned} w_{ij} &= (1+\log \mathrm{df}_{ij})\cdot \mathrm{itf}_j, \\ \mathrm{itf}_j &= \log(M/M_j). \end{aligned}$$

ただし，df_{ij} は単語 t_i が文書 d_j に出現した回数[3]，M_j は文書 d_j が含む異なり単語数とする．

単語間の類似度 $\mathrm{sim}(\vec{t},\vec{t'})$ は，文書の場合と同様に以下の式で定義される．

$$\mathrm{sim}(\vec{t},\vec{t'}) = \frac{\vec{t}\cdot\vec{t'}}{|\vec{t}||\vec{t'}|}.$$

2.3 概念ベースクエリ拡張

本論文では概念ベースクエリ拡張 [4] と呼ばれる手法を用いて，前節の文書クラスタ別の単語間類似度による精度向上への有効性を確かめる．概念ベースクエリ拡張では，クエリをベクトル空間モデルを用いてベクトル化し，それに最も近い単語ベクトルを持つ単語をクエリに追加する．

クエリは文書で与えられることを想定しており，クエリ q に含まれる単語の重みベクトルを $\vec{q}=(q_1,\ldots,q_M)$ で表す．各 q_i は 2.1 節の文書ベクトルと同様に以下の式で定義される．

$$q_i = (1+\log \mathrm{tf}_i)\cdot \mathrm{idf}_i.$$

このクエリに対して，クエリの概念を表す仮想的な単語ベクトル $\vec{q_c}$ を 2.2 節の $\vec{t_i}$ を用いて以下のように定義する．

$$\vec{q_c} = \sum_{i=1}^{M} q_i \cdot \frac{\vec{t_i}}{|\vec{t_i}|}.$$

[3] $\mathrm{df}_{ij}=0$ のときは $w_{ij}=0$ とする．

クエリと単語 t の類似度を

$$\mathrm{simqt}(q,t) = \vec{q_c}\cdot\frac{\vec{t}}{|\vec{t}|}$$

と定義すると，

$$\begin{aligned} \mathrm{simqt}(q,t) &= \left(\sum_{i=1}^{M} q_i\cdot\frac{\vec{t_i}}{|\vec{t_i}|}\right)\cdot\frac{\vec{t}}{|\vec{t}|} \\ &= \sum_{i=1}^{M} q_i\cdot\frac{\vec{t_i}\cdot\vec{t}}{|\vec{t_i}||\vec{t}|} = \sum_{i=1}^{M} q_i\cdot\mathrm{sim}(\vec{t_i},\vec{t}). \end{aligned}$$

単語 t をクエリ q に追加する場合は，$\mathrm{simqt}(q,t)$ の値を重みとして用いることができる．クエリ拡張に用いる単語を制限するために，閾値 T を定めて以下のように拡張クエリを求める．

$$\begin{aligned} \vec{q_e} &= (q_{e1},\ldots,q_{eM}), \\ q_{ei} &= \begin{cases} \mathrm{simqt}(q,t_i) & \left(\frac{\mathrm{simqt}(q,t_i)}{\sum_{i=1}^{M} q_i} \geq T\right) \\ 0 & (\text{それ以外}) \end{cases}. \end{aligned}$$

拡張したクエリ $\vec{q}+\vec{q_e}$ を用いて，各文書 d_j に対し，$s(d_j)=\mathrm{sim}(\vec{d_j},\vec{q}+\vec{q_e})$ の値をスコアとしてランク付けすることができる．

3 文書クラスタを用いた単語クラスタリングと情報検索

3.1 文書クラスタを用いた単語クラスタリング

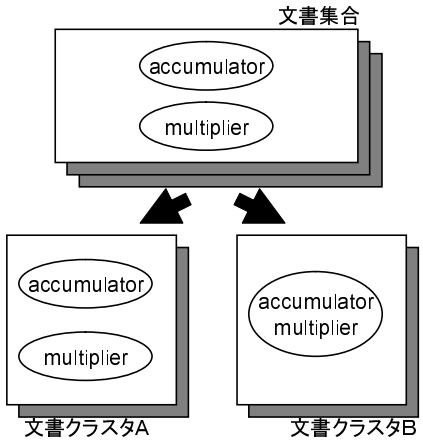
文書クラスタリングの結果を用いるときは，単語ベクトル $\vec{t_i}$ を次のように変形する．文書 d_j の文書クラスタ C_k への帰属度を δ_{jk} とすると，文書クラスタ C_k における単語ベクトル $\vec{t_i}^{(k)}$ は，

$$\begin{aligned} \vec{t_i}^{(k)} &= (w_{i1}^{(k)},\ldots,w_{iN}^{(k)}), \\ w_{ij}^{(k)} &= \delta_{jk}(1+\log \mathrm{df}_{ij})\cdot \mathrm{itf}_j. \end{aligned}$$

で表される．類似度は $\mathrm{sim}_k(t,t') = \frac{\vec{t}^{(k)}\cdot\vec{t'}^{(k)}}{|\vec{t}^{(k)}||\vec{t'}^{(k)}|}$ で求められる．

各文書クラスタ C_k について異なる類似度 sim_k が得られるため，それを用いた単語クラスタリングの結果もそれぞれ異なったものになり，文書クラスタの性質を反映した単語クラスタを生成できると考えられる（図 1 参照）．

各文書が単一のクラスタに属すハードクラスタリングを用いた場合，すなわち $\delta_{jk}\in\{0,1\}$ の場合は，各文書クラスタ内に出現する単語のみでそ



accumulator（加算器）と multiplier （乗算器）を類義語として扱える分野と扱えない分野が存在するとき，文書クラスタごとの単語クラスタリングによってその特徴をとらえることができる．

図 1: 文書クラスタ別の単語クラスタリング

れぞれ単語クラスタリングを行ったのと等価になる．この場合は全文書集合に対する単語クラスタリングに比べてデータ量が少なくなり生成される単語クラスタの質が落ちる可能性がある．ソフトクラスタリングを用いれば，データ量を保ったまま文書クラスタの性質を単語クラスタに反映させることが期待できる．

3.2 概念ベースクエリ拡張への適用

概念ベースクエリ拡張において拡張されるクエリ $\vec{q}_e$ は 文書クラスタごとに異なる．各文書クラスタ C_k に対する $\mathrm{sim}_k(t,t')$ を用いた拡張クエリを $\vec{q}_e^{(k)}$ とする．

クエリと各文書クラスタの関連度 $\mathrm{sim}(\vec{q},\vec{v}_k)$ を用いて拡張クエリの重み付けを行うと式 (1) が得られる．

$$s'(d_j) = \vec{d}_j \cdot \left(\vec{q} + \sum_{k=1}^{c} \mathrm{sim}(\vec{q},\vec{v}_k)\vec{q}_e^{(k)} \right). \qquad (1)$$

各文書のクラスタへの帰属度 δ_{jk} を考慮し，文書とクエリが同じトピックに属するときにスコアを大きくするよう修正すると，(2) が得られる．

$$s''(d_j) = s'(d_j) \sum_{k=1}^{c} \delta_{jk}\mathrm{sim}(\vec{q},\vec{v}_k). \qquad (2)$$

4 実験

前節で導いたスコアリング法の有効性を確認するため，表 1 に示したテストセットから以下の

表 1: テストセット

テストセット	内容	文書数	クエリ数
CACM	コンピュータ科学	3204	52
ISI	図書館学	1460	41
LISA	図書館学	6004	39
MED	医学	1033	30
NPL	電子工学	11429	93

表 2: 平均適合率
c は文書クラスタ数．ideal はセット B の各文書に対してもとのテスト集合に対応するクラスタを割り当てた場合．

条件	A	B
クエリ拡張なし	0.1903	0.0501
概念ベースクエリ拡張	0.2338	0.0649
スコア s', $c=1$	0.2375	0.0647
スコア s', $c=5$	0.2407	0.0654
スコア s', $c=10$	0.2380	0.0651
スコア s', ideal	—	0.0676
スコア s'', $c=1$	0.2375	0.0647
スコア s'', $c=5$	0.2267	0.0627
スコア s'', $c=10$	0.2129	0.0589
スコア s'', ideal	—	0.0808

2 つの実験用セットを作成して情報検索の実験を行った．

A. CACM

B. CACM + ISI + LISA + MED + NPL

テストセットに含まれる単語のうちストップワードはあらかじめ取り除いた．クエリのうちそのクエリに対する正解の文書数が 0 であるものは除いた[4]．拡張前のクエリによるランク付けと，s', s'' によるランク付けを文書クラスタ数を変化させて精度評価を行った．評価は再現率が $\{0, 0.1, 0.2, \ldots, 1\}$ の 11 点についての適合率で行い，その平均値を平均適合率として算出した．クエリ拡張の閾値 T は，各実験においてそれぞれ平均適合率が高くなる値を選んだ．

テストセット A, B に対して，クエリをそのまま検索に用いる手法，概念ベースクエリ拡張に基づく手法，スコアとして s', s''，文書クラスタ数として $c = 1, 5, 10$ を用いる手法の各々の平均適合率を表 2 に示した．また各テスト集合をそれぞれ文書クラスタとみなした場合の平均適合率も示した．再現率の各点における適合率のグラフは図 2, 3 に示した．

[4]表 1 のクエリ数は取り除いた後の数を表す．

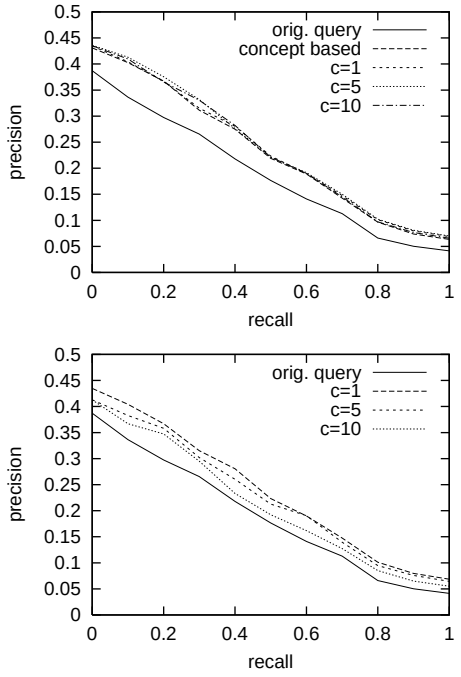


図 2: セット A の適合率-再現率カーブ（上段：スコア s'，下段：スコア s''）

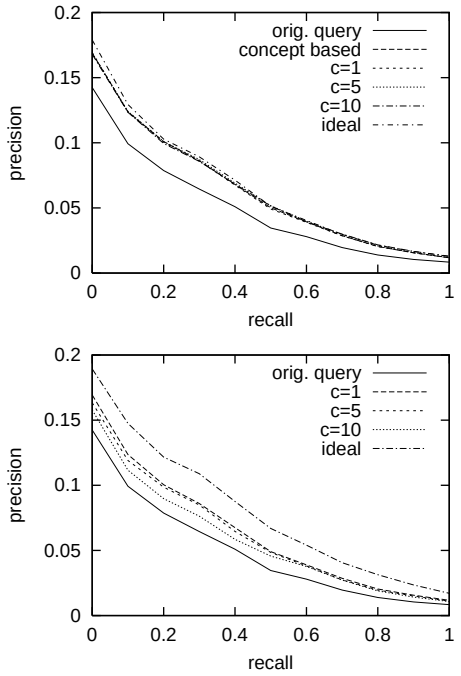


図 3: セット B の適合率-再現率カーブ（上段：スコア s'，下段：スコア s''）

テストセット A および B において，拡張前のクエリの適合率に対して概念ベースクエリ拡張を用いた方が適合率が高くなった．文書クラスタリングを適用した例では，スコア s' を用いた場合は平均適合率に大きな差は見られず，スコア s'' を用いた場合は適合率の低下が観測された．

一方で，文書クラスタリングの代わりに各テストセットを文書クラスタとみなした場合では，スコア s'' について比較的改善がみられた．これは文書クラスタリングをより精密に行うことで精度を改善できる可能性を示していると考えられる．クエリ拡張の最適な閾値 T はスコアリング手法やクラスタ数，テスト集合によって大きく変化するため，これを効率よく発見するのは今後の課題である．

5 まとめ

本論文では文書のファジィクラスタリングを用いて文書集合に対して各トピック（文書クラスタ）への帰属度を求め，それを用いたトピックに特化した単語クラスタの生成，および概念ベースクエリ拡張を用いた情報検索への応用を試みた．情報検索での実験では文書クラスタリングによる効果ははっきりと現れなかったが，文書クラスタリングの精密化により効果が出る可能性もある．今後の課題として，式 (1), (2) によるスコアリングでは全ての文書に対して同じクエリの拡張を行っているが，クエリの拡張を文書ごとに行うことによって，各文書の性質を反映したクエリ拡張による精度の向上を目指したい．

参考文献

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[2] Carolyn J. Crouch and Bokyung Yang. Experiments in automatic statistical thesaurus construction. In *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88, 1992.

[3] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1974.

[4] Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.

[5] P. Schauble and D. Knaus. The various roles of information structures. In *Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e.V."*, pages 282–290, 1992.