情報処理学会論文誌 Vol.60 No.3 880–889 (Mar. 2019)

単語重みを用いたアソシエーション分析に基づく文書分類のための自動的クエリ拡張

安永 翼[1,a)] 山田 雄基[1] 濱上 知樹[2,b)]

受付日 2018年6月7日, 採録日 2018年12月4日

概要：文書分類問題では単語数の少ない文書を分類することが難しい．情報検索分野ではクエリの単語数の少なさに対して，アソシエーション分析に基づきクエリに含まれる単語と関連度の高い単語を追加することで検索性能を改善するクエリ拡張手法がある．しかし文書分類問題においては，クエリに含まれる単語と関連度の高い単語が正しいクラスを特徴付けるとは限らないため，分類性能を改善することはできない．また，従来のアソシエーション分析では文書に対する単語の重要度（単語重み）が考慮されていないため，単語間の関連度が適切でない可能性がある．これらの課題をふまえて，本論文では文書分類性能改善のために 2 つの提案を行う．(1) クラスごとに分割した文書集合に対してアソシエーション分析を行い，クラスに対する単語の重要度に基づいて推定されたクラスにおいてクエリに含まれる単語との関連度が高い単語を追加する．(2) アソシエーション分析における単語間の関連度計算に単語重みを利用する．実験では，クエリ拡張を用いた単語数の少ない文書の分類タスクを複数のデータセットで実施し，提案手法による拡張後クエリの分類性能改善を確認した．また関連度に設定する閾値に対する評価指標の感度分析により，多くの閾値設定で単語重みを考慮する方が分類性能を改善できることを明らかにした．

キーワード：アソシエーション分析，文書分類，自動的クエリ拡張，ファジィ集合

Automatic Query Expansion for Document Classification Based on Association Analysis with Term Weights

Tsubasa Yasunaga[1,a)] Yuki Yamada[1] Tomoki Hamagami[2,b)]

Received: June 7, 2018, Accepted: December 4, 2018

Abstract: We propose automatic query expansion for document classification based on association analysis with term weights. In document classification, it is difficult to classify document with a few terms. Automatic query expansion based on association analysis improves document retrieval performance by adding terms with high relevance to the query terms. However, terms with high relevance to the query terms does not always characterize correct class. Moreover, relevance between terms can be inappropriate in the conventional association analysis because term weights are not considered. For each of these problem, we propose two approaches. (1) we apply association analysis to each document set divided by class and add terms with high relevance to the query terms in the estimated class. (2) we use term weights in calculation of relevance between terms. The experimental result shows that the proposed method improves classification performance in some datasets and use of term weights improves classification performance in many settings of threshold of degree of relevance.

Keywords: association analysis, document classification, automatic query expansion, fuzzy set

1. はじめに

情報検索では検索システムに入力されるクエリの単語数

[1] 横浜国立大学大学院工学府
Graduate School of Engineering, Yokohama National University, Yokohama, Kanagawa 240–8501, Japan
[2] 横浜国立大学大学院工学研究院
Fuculty of Engineering, Yokohama National University, Yokohamai, Kanagawa 240–8501, Japan
[a)] yasunaga-tsubasa-gv@ynu.jp
[b)] hamagami@ynu.ac.jp

の少なさからユーザの求める文書とクエリとのミスマッチが大きいため，ユーザの求める文書を得ることは難しい．このミスマッチとは単語の違いであり，クエリに含まれる単語がユーザの求める文書において出現しない，または重要度が低いということである．このような場合にはユーザの求める文書とクエリとのミスマッチを軽減させるような単語を追加することで検索性能を改善する自動的クエリ拡張（Automatic Query Expansion, AQE）が有効である [1]．追加すべき単語の選択方法として，アソシエーション分析に基づく単語間の関連度を用いて条件を満たす単語を追加する方法がある [2]．アソシエーション分析はアイテム集合からなるトランザクションデータにおいて共起関係に基づいてアイテム間の関連性を分析する手法である [3]．Wei らは単語をアイテムと見なすことでアソシエーション分析を適用し，単語間の関連性を AQE に利用した [2]．

文書分類においても情報検索と同様にクエリの単語数が少ないときには訓練データとなる文書とのミスマッチが大きくなり，正しく分類することが難しい．しかしながら単純に AQE を適用するだけでは分類性能を改善することはできない．これは文書集合全体においてクエリに含まれる単語との関連度が高い単語が，正しいクラスを特徴付けるとは限らないからである．また，もう 1 つの課題として，従来のアソシエーション分析ではある 2 単語の関連度を計算するとき，その 2 単語が同時に出現する文書数によって共起性を定義しており，単語重みを考慮していない．単語重みは文書への出現頻度などによって計算される単語の重要度である [4]．すなわち単語重みを考慮しない従来のアソシエーション分析では，単語が文書に出現するかしないかのみが考慮される．これにより，ある単語が出現する文書について，その単語が重要である文書と重要でない文書を同等に扱ってしまい，このようにして計算された単語間の関連度は適切ではないと考えられる．

以上 2 つの課題のそれぞれに対して，本論文では文書分類性能改善のために以下の個別に適用可能な 2 つの手法を提案する．

(1) クラスごとに分割した文書集合に対してアソシエーション分析を行い，事前に推定されたクラスにおいてクエリに含まれる単語との関連度が高い単語を追加する [5]．
(2) 文書を語彙集合のファジィ部分集合として扱うことで，アソシエーション分析における単語間の関連度計算に単語重みを利用する [6]．

実験では，複数のデータセットに対する分類性能の評価や，単語を追加する条件となる関連度の閾値に対する評価指標の感度分析を行い，本手法の性質を明らかにする．

以下，2 章ではアソシエーション分析に基づく AQE の既存手法について述べる．3 章では提案手法として AQE を文書分類タスクに適用する方法とアソシエーション分析の関連度計算に単語重みを利用する方法を述べる．4 章では単語数の少ない文書を分類するタスクに AQE を利用する実験と結果について述べる．5 章では本研究の関連研究について述べる．

2. アソシエーション分析に基づく自動的クエリ拡張

2.1 単語間の関連度

単語 t_1 が出現するとき単語 t_2 も出現するというルールを $t_1 \Rightarrow t_2$ と表し，t_1 を条件部，t_2 を結論部と呼ぶ．アソシエーション分析ではルールの強さを表す指標として支持度 *support*，確信度 *confidence*，リフト値 *lift* がそれぞれ式 (1)，(2)，(3) で定義される．

$$support(t_1 \Rightarrow t_2) = \frac{\sigma(t_1, t_2)}{|D|} \tag{1}$$

$$confidence(t_1 \Rightarrow t_2) = \frac{support(t_1 \Rightarrow t_2)}{support(t_1)} \tag{2}$$

$$lift(t_1 \Rightarrow t_2) = \frac{confidence(t_1 \Rightarrow t_2)}{support(t_2)} \tag{3}$$

D はアソシエーション分析を行う文書集合，$\sigma(t_1, t_2)$ は単語 t_1 と単語 t_2 が共起する文書数を表す．ただし単語 t が出現する文書数を $\sigma(t)$ とし，$support(t) = \sigma(t)/|D|$ である．本論文では支持度，確信度，リフト値を総称して関連度と呼ぶ．

2.2 クエリ拡張

クエリに単語 t_1 が含まれるときルール $r : t_1 \Rightarrow t_2$ が有効であると見なされれば単語 t_2 をクエリに追加する．有効であると見なされるルールをアソシエーションルールと呼び，アソシエーションルール集合 R は式 (4) で表される．

$$\begin{aligned} R = \{r | & support(r) > minsupp, \\ & confidence(r) > minconf, \\ & lift(r) > minlift\} \end{aligned} \tag{4}$$

minsupp，*minconf*，*minlift* はそれぞれ支持度，確信度，リフト値に設定する閾値を表す．

2.3 課題

AQE を文書分類問題に適用するとき，分類性能を改善するためには正しいクラスを特徴付ける単語を追加することが重要となる．しかし文書集合全体においてクエリに含まれる単語との関連度が高い単語が，正しいクラスを特徴付けるとは限らない．したがって AQE を単純に適用するだけでは分類性能は改善しないと考えられる．また，式 (1)〜(3) よりアソシエーション分析を行う文書集合が異なれば関連度も異なるため，単語間の関連度はクラスごとに異なる．たとえば “行く” という単語に対して，スポーツ記事では “スタジアム”，音楽記事では “ライブ” という単

語との関連度が高くなると考えられる．しかし文書集合全体に対してはこれらの単語の共起頻度は小さくなるため，文書集合全体に対してアソシエーション分析を行うと関連度は小さくなる．

本論文では，クラスごとに分割した文書集合に対してアソシエーション分析を行うことで，クラスに特有のアソシエーションルール集合を生成する．そしてクラスに対する単語の重要度に基づいてクエリのクラスを推定し，推定されたクラスのアソシエーションルール集合を用いてクエリ拡張を行うことで，クエリと推定されたクラスに属する訓練データとのミスマッチを小さくし，分類性能を改善する．

もう 1 つの課題として，アソシエーション分析ではある 2 単語間の関連度がその 2 単語の共起頻度を表す関数 σ を用いて計算されるが，このとき単語重みは考慮されていない．そのため，たとえば単語 t_1 が出現する文書について，単語 t_2 の単語重みが異なる文書があるとすると，どちらの文書も単語 t_1 と単語 t_2 が共起する文書として同等に扱われる．しかし結論部である単語 t_2 の単語重みが大きい文書は共起する文書として重視すべきであり，単語 t_2 の単語重みが小さい文書は共起する文書として軽視すべきである．この問題のために従来のアソシエーション分析で生成されるアソシエーションルールは，情報検索や文書分類の性能改善を妨げる単語を追加してしまう可能性がある．そのような単語を本論文では妨害単語と呼ぶ．

従来のアソシエーション分析で単語重みが考慮されないのは，文書が語彙集合のクリスプ部分集合として扱われているからである．すなわち単語は文書に属するか，属さないかのみが考慮される．そこで本論文ではファジィ集合の概念 [7] を導入し，単語が文書にどれくらいの度合いで属するかまで考慮する．これにより結論部の単語重みが大きい文書は重視，小さい文書は軽視し，生成されるアソシエーションルールによる妨害単語の追加を抑制することで，拡張後クエリの分類性能を改善する．

3. 提案手法

ここでは 2 章で述べた既存手法の改善策として，2 つの手法を提案する．3.1 節ではクラスごとのアソシエーションルール集合の生成方法と，生成された各アソシエーションルール集合をクエリ拡張に利用する方法を述べる．3.2 節では単語間の関連度計算に単語重みを利用する方法を述べる．

3.1 文書分類のための自動的クエリ拡張

本論文では 2 章で述べた全文書集合におけるアソシエーションルールを用いる情報検索のための AQE を AQE-R，ここで述べる文書分類のための AQE を AQE-C と呼ぶ．AQE-R と AQE-C の違いを**図 1** に示す．AQE-R では文書集合全体 D からアソシエーションルール集合 R を生成し，R を用いてクエリ Q を拡張することで拡張後クエリ Q' を得る．

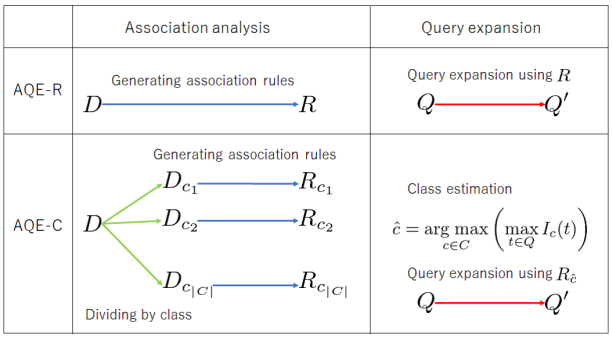

図 1 AQE-R と AQE-C の違い
Fig. 1 Difference between AQE-R and AQE-C.

AQE-C では訓練データである文書集合全体 D をクラスごとに分割し，各文書集合に対してアソシエーション分析を行う．すなわちクラス $c \in C$ に属する文書集合 D_c からは式 (1)〜(4) を用いてアソシエーションルール集合 R_c が生成される．クエリ拡張時にはクラスに対する単語の重要度に基づいてクエリのクラスを推定し，推定されたクラス $\hat{c} \in C$ におけるアソシエーションルール集合 $R_{\hat{c}}$ を用いてクエリ拡張を行う．クラス c に対する単語 t の重要度 $I_c(t)$ は単語重みを用いて式 (5) で定義される．

$$I_c(t) = \frac{\sum_{d \in D_c} w_d(t)}{\sum_{c' \in C} \sum_{d' \in D_{c'}} w_{d'}(t)} \tag{5}$$

$w_d(t)$ は文書 d における単語 t の単語重み，D_c はクラス c に属する文書集合，C はクラス集合である．単語重み $w_d(t)$ は文書 d における単語 t の出現頻度などによって与えられる．式 (5) を用いてクエリの推定クラス $\hat{c}$ は式 (6) で決定される．

$$\hat{c} = \arg\max_{c \in C} \left(\max_{t \in Q} I_c(t) \right) \tag{6}$$

Q はクエリに含まれる単語の語彙集合である．

ここで，クエリのクラスを推定する段階で分類は可能であるが，この推定結果はクエリと訓練データ文書の類似度に基づいていない．本研究では分類結果の解釈性の観点から，クエリとの類似度による訓練データ文書の順位付けに基づいて分類することを想定している．したがって，情報検索問題においてクエリとユーザの求める文書とのミスマッチを軽減させるためにクエリ拡張が必要であったのと同様に，本研究における文書分類ではクラス推定だけでなくクエリ拡張を行う必要がある．

3.2 関連度計算における単語重みの利用

文書を語彙集合のファジィ部分集合として扱う．すなわち語彙集合 V 上の文書 d を特徴付けるメンバシップ関数 μ_d は式 (7) で表される．

$$\mu_d : V \to [0, 1] \tag{7}$$

このとき $\mu_d(t)$ は単語 t の文書 d におけるメンバシップ度を表す．本論文では $\mu_d(t)$ を単語 t の文書 d における単語重み $w_d(t)$ を用いて式 (8) で定義する．

$$\mu_d(t) = \frac{w_d(t)}{\sum_{t' \in V_d} w_d(t')} \tag{8}$$

V_d は文書 d に含まれる単語の語彙集合である．式 (8) を用いて，単語 t_1 と単語 t_2 の共起性を表す関数 $\sigma(t_1, t_2)$ を式 (9) で表す．

$$\sigma(t_1, t_2) = \sum_{d \in D_{t_1}} \mu_d(t_2) \tag{9}$$

D_t は単語 t を含む文書集合である．ただし $\sigma(t) = \sum_{d \in D_t} \mu_d(t)$ とする．ここで定義した σ を式 (1)〜(4) に適用することで単語重みを考慮したアソシエーションルールが生成される．

従来のアソシエーション分析におけるメンバシップ関数は式 (10) で表される．

$$\mu_d(t) = \begin{cases} 1 & (t \in V_d) \\ 0 & (t \notin V_d) \end{cases} \tag{10}$$

すなわち $\mu_d(t)$ は単語 t が文書 d に含まれれば 1，含まれなければ 0 をとり，このとき式 (9) は従来のアソシエーション分析における σ と等しくなる．

4. 実験

4.1 データセット

データセットとして英語の 20 Newsgroups [8]，日本語の livedoor ニュースコーパス [9] を用いた．各文書はそれぞれ 1 つのクラスに属している．20 Newsgroups データセットではクラス数が 4 の場合とさらに細分化されたクラス数が 20 の場合で実験を行った．livedoor ニュースコーパスデータセットのクラス数は 9 である．本研究では単語数の少ない文書の分類性能を改善することが目的のため，20 Newsgroups データセットではテストデータとして単語数が 30 語以下のものを用いた．livedoor ニュースコーパスデータセットでは単語数が 30 語以下のデータが 11 件と少数であったため，単語数が少ない順に 1 割のデータをテストデータとして用いた．訓練データの平均単語数は，20 Newsgroups データセットではクラス数 4 のとき 194.69 語，クラス数 20 のとき 194.48 語，livedoor ニュースコーパスデータセットでは 374.55 語であった．訓練データ数は，20 Newsgroups データセットではクラス数 4 のとき 9,608 件，クラス数 20 のとき 11,269 件，livedoor ニュースコーパスデータセットでは 6,639 件であった．テストデータ数は，20 Newsgroups データセットではクラス数 4 のとき 426 件，クラス数 20 のとき 553 件，livedoor ニュースコーパスデータセットでは 737 件であった．

4.2 分類方法

クエリとの類似度による訓練データ文書の順位付けに基づいた分類手法として k 近傍法を用いた．クエリの近傍 $k = 5$ 個の訓練データ文書が属するクラスのうち最も多いクラスを予測クラスとした．クエリ q と文書 d の類似度 $sim(q, d)$ はコサイン類似度を用いて式 (11) で与えられる．

$$sim(q, d) = \frac{\overrightarrow{q} \cdot \overrightarrow{d}}{|\overrightarrow{q}||\overrightarrow{d}|} \tag{11}$$

$\overrightarrow{q}$，$\overrightarrow{d}$ はそれぞれクエリ q，文書 d のベクトル表現である．文書ベクトル $\overrightarrow{d}$ は単語重みを用いて式 (12) で表される．

$$\overrightarrow{d} = \left(w_d(t_1), w_d(t_2), \cdots, w_d(t_{|V|})\right) \in \mathbb{R}^{|V|} \tag{12}$$

$w_d(t)$ は単語 t の文書 d における単語重み，V は語彙集合である．本実験では単語重みについて式 (13)，(14) の 2 通りで実験した．

$$w_d(t) = n_d(t) \tag{13}$$

$$w_d(t) = \frac{n_d(t)}{\sum_{t' \in V} n_d(t')} \left(1 + \log \frac{|D|}{|\{d' | t \in V_{d'}\}|}\right) \tag{14}$$

$n_d(t)$ は文書 d における単語 t の出現数，D は全文書集合である．$V_{d'}$ は文書 d' に含まれる語彙集合であり，$\{d' | t \in V_{d'}\}$ は単語 t を含む文書集合を表す．本論文では式 (13)，(14) を用いて表現される文書モデルをそれぞれ Bag of Words（BOW）モデル，Term Frequency and Inverse Document Frequency（TFIDF）モデルと呼ぶ．

4.3 評価指標

評価指標として以下の 3 つを用いた．

(1) クエリ 1 件あたりの平均単語数
(2) 正解率
(3) F 値（F1-score）

平均単語数はクエリ拡張により追加された単語数の目安となる．正解率は全テストデータのうち正解したテストデータの割合である．F 値は適合率と再現率それぞれのマクロ平均 P_m，R_m の調和平均として式 (15) で与えられる．

$$F = \frac{2P_m R_m}{P_m + R_m} \tag{15}$$

P_m，R_m は式 (16) で与えられる．

$$P_m = \frac{1}{|C|} \sum_{c \in C} P(c),\ R_m = \frac{1}{|C|} \sum_{c \in C} R(c) \tag{16}$$

C はクラス集合，$P(c)$，$R(c)$ はそれぞれクラス c についての適合率と再現率である．すなわち $P(c)$ はクラス c と予測したテストデータのうち正解したテストデータの割合であり，$R(c)$ はクラス c に属するテストデータのうち正解したテストデータの割合である．

拡張後クエリの平均単語数が少なく，正解率と F 値が大きければ，妨害単語の追加が抑制されているといえる．

4.4 比較手法

2 章で述べた AQE-R と 3.1 節で述べた AQE-C の違いに加え，3.2 節で述べた単語重みの考慮をするかどうかで場合分けし，各評価指標を以下の 4 つの手法に関して比較する．

(1) AQE-R
(2) 単語重みを用いた AQE-R
(3) AQE-C
(4) 単語重みを用いた AQE-C

初期クエリに関しても同様の評価を行う．また，正解率と F 値については式 (6) のクラス推定による分類とも比較する．

4.5 使用する単語の選択と関連度の閾値設定

使用する単語は出現する文書数が 10 以上かつ出現する文書数の総文書数に対する割合が 0.8 以下のものとし，支持度，確信度，リフト値の閾値はそれぞれ $minsupp = 0$，$minconf = 0.5$，$minlift = 1$ とした．支持度の閾値を $minsupp = 0$ としているのは使用する単語の選択によりあらかじめ出現数が低く共起する可能性の低い単語を除外しているためである．

4.6 実験結果

クラス数 4 の 20 Newsgroups データセット [8] で BOW モデル，TFIDF モデルを用いたときの実験結果をそれぞれ**図 2**，**図 3** に示す．AQE-C は AQE-R に比べて正解率と F 値が改善された．単語重みの使用により拡張後のクエリの平均単語数は減少し，BOW モデルを用いた AQE-C での F 値を除いて正解率と F 値は改善された．TFIDF モデルでは，単語重みを用いた AQE-C のみ正解率と F 値が初期クエリよりも高くなった．

クラス数 20 の 20 Newsgroups データセットで BOW モデル，TFIDF モデルを用いたときの実験結果をそれぞれ**図 4**，**図 5** に示す．クラス数が増加しても同様に AQE-C は AQE-R に比べて正解率と F 値が改善された．単語重みの使用により拡張後のクエリの平均単語数は減少し，BOW モデルを用いた AQE-R での F 値を除いて正解率と F 値は改善された．

クラス数 9 の livedoor ニュースコーパスデータセット [9] で BOW モデル，TFIDF モデルを用いたときの実験結果をそれぞれ**図 6**，**図 7** に示す．異なるデータセットでも同様に AQE-C は AQE-R に比べて正解率と F 値が改善され

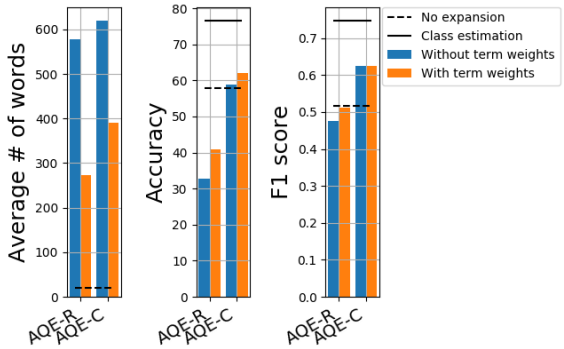

図 2 クラス数 4 の 20 Newsgroups データセットで BOW モデルを用いたときの AQE の評価

Fig. 2 Evaluation metrics of AQE with BOW in 20 Newsgroups (4 classes).

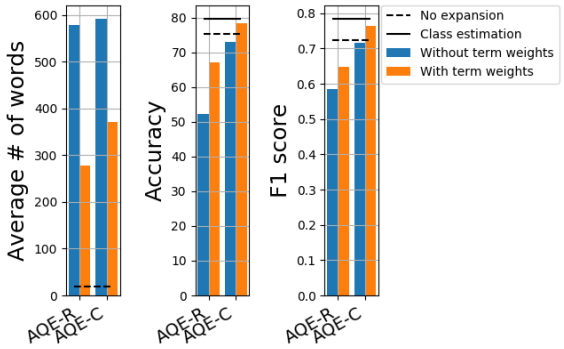

図 3 クラス数 4 の 20 Newsgroups データセットで TFIDF モデルを用いたときの AQE の評価

Fig. 3 Evaluation metrics of AQE with TFIDF in 20 Newsgroups (4 classes).

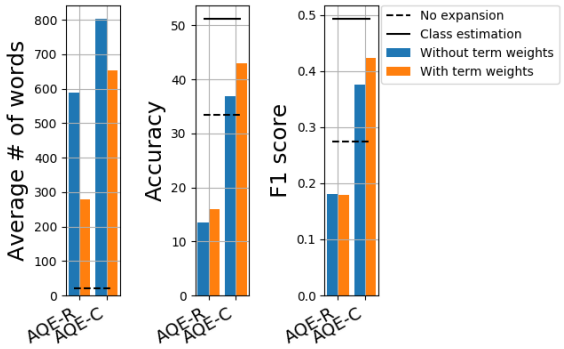

図 4 クラス数 20 の 20 Newsgroups データセットで BOW モデルを用いたときの AQE の評価

Fig. 4 Evaluation metrics of AQE with BOW in 20 Newsgroups (20 classes).

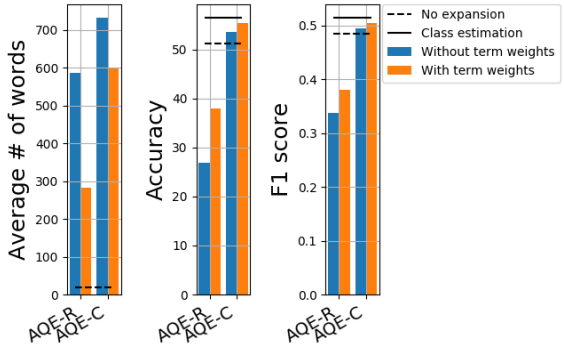

図 5 クラス数 20 の 20 Newsgroups データセットで TFIDF モデルを用いたときの AQE の評価

Fig. 5 Evaluation metrics of AQE with TFIDF in 20 Newsgroups (20 classes).

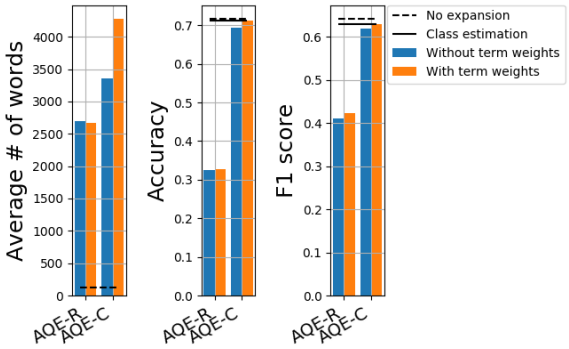

図 6 クラス数 9 の livedoor ニュースコーパスデータセットで BOW モデルを用いたときの AQE の評価

Fig. 6 Evaluation metrics of AQE with BOW in livedoor news corpus (9 classes).

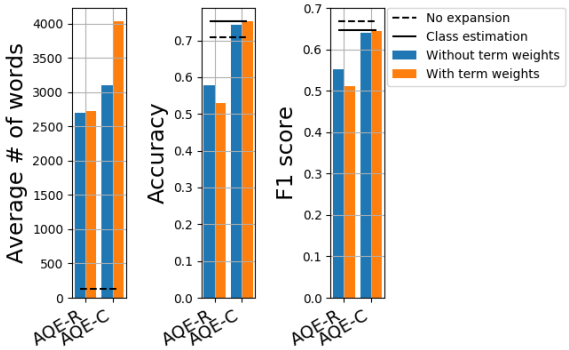

図 7 クラス数 9 の livedoor ニュースコーパスデータセットで TFIDF モデルを用いたときの AQE の評価

Fig. 7 Evaluation metrics of AQE with TFIDF in livedoor news corpus (9 classes).

た．20 Newsgroups データセットでの結果と異なる傾向として，単語重みの使用により拡張後のクエリの平均単語数が BOW モデルを用いた AQE-R を除いて増加し，TFIDF モデルを用いた AQE-R では提案手法によって正解率と F 値が低下した．また，正解率と F 値に関して TFIDF モデルでの正解率を除きクラス推定による分類が初期クエリの分類より低い．クラス推定はクラスに対する単語の重要度に基づいて行われるため，初期クエリに正しいクラスを特徴付ける単語が存在しない場合はクラス推定を誤る．すなわち livedoor ニュースコーパスデータセットでは初期クエリに正しいクラスを特徴付ける単語が存在しないテストデータが多いと考えられる．

4.7 関連度の閾値に対する評価指標の感度

生成されるアソシエーションルールは関連度の閾値に影響を受けるため，関連度の閾値に対する評価指標の感度を確認する．クラス数 4 の 20 Newsgroups データセットで TFIDF モデルを用いた場合について，確信度の閾値 *minconf* とリフト値の閾値 *minlift* の片方を固定し，もう片方を変化させて同様のクエリ拡張を用いた分類実験を行った．閾値を変化させる範囲は単語重みを用いた AQE-C における関連度に対するルール数の分布から実験的に決定した．

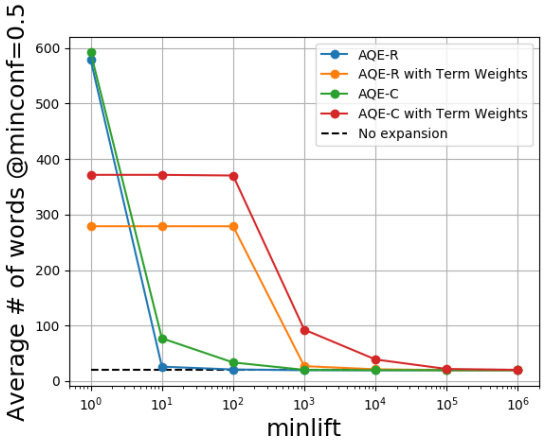

図 8 リフト値の閾値に対するクエリ 1 件あたりの平均単語数の感度

Fig. 8 Sensitivity of average number of words in one query to *minlift*.

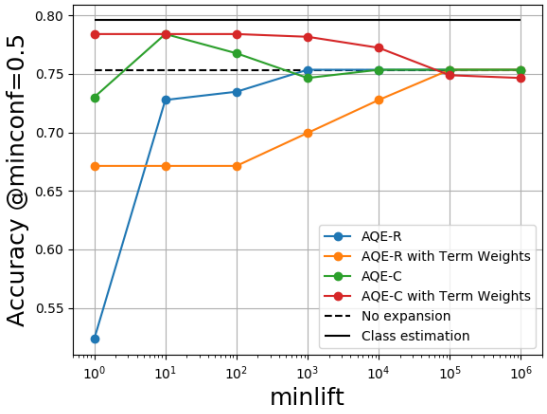

図 9 リフト値の閾値に対する正解率の感度

Fig. 9 Sensitivity of accuracy to *minlift*.

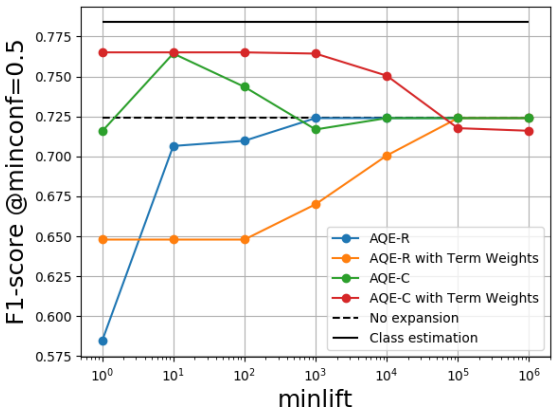

図 10 リフト値の閾値に対する F 値の感度

Fig. 10 Sensitivity of F1 score to *minlift*.

4.7.1 リフト値の閾値に対する評価指標の感度

確信度の閾値を $minconf = 0.5$ に固定し，リフト値の閾値 *minlift* を 1，10，10^2，10^3，10^4，10^5，10^6 と変化させたときのクエリ 1 件あたりの平均単語数，正解率，F 値をそれぞれ**図 8**，**図 9**，**図 10** に示す．

単語重みを用いない AQE-R では *minlift* を高くすると拡張後クエリの平均単語数は減少，正解率と F 値は向上し，次第に初期クエリの評価指標値に等しくなっている．このことから単語重みを用いない AQE-R で *minlift* が小さいときに追加されていた単語は妨害単語であったことが分かる．

単語重みを用いない AQE-C では同様に拡張後クエリの平均単語数は減少していくが，$minlift = 10, 10^2$ では初期クエリよりも正解率と F 値が改善されており，分類に有用な単語を選択的に追加できている．

単語重みを用いた場合，$minlift = 1, 10, 10^2$ では AQE-R，AQE-C ともにどの評価指標も変化していない．すなわち単語重みを用いた場合は，この範囲でリフト値が閾値以下になるルールがあるとしても，そのルールはすでに確信度が 0.5 以下であり，生成されるアソシエーションルールに影響を与えない．

単語重みを用いた AQE-C は正解率と F 値に関して $minlift = 10^2$ まで比較手法中最大であるが，$minlift = 10^5, 10^6$ では初期クエリでの値を下回っている．このことから単語重みを用いた AQE-C においてきわめて高いリフト値を持つルールで追加される単語は妨害単語であることが分かり，*minlift* が小さいときは分類に有用ないくつかの追加単語が，妨害単語が誤ったクラスを特徴付けるよりも強く正しいクラスを特徴付けていると考えられる．

また $minlift = 10$ では単語重みを用いない AQE-C と単語重みを用いた AQE-C の正解率と F 値がほぼ等しくなっているが，拡張後クエリの平均単語数は単語重みを用いた AQE-C の方が多い．したがってこのとき単語重みを用いた AQE-C では，妨害単語ではないが不要な単語が多く追加されていることが分かる．

4.7.2 確信度の閾値に対する評価指標の感度

リフト値の閾値を $minlift = 1$ に固定し，確信度の閾値 *minconf* を 0.1〜1 まで 0.1 刻みで変化させたときのクエリ 1 件あたりの平均単語数，正解率，F 値をそれぞれ**図 11**，**図 12**，**図 13** に示す．

単語重みを用いない AQE-R ではリフト値の閾値に対する感度と同様の傾向であり，妨害単語が追加されることが分かる．

単語重みを用いた AQE-R は単語重みを用いない AQE-R と同様の傾向を示すものの，正解率と F 値に関してほとんどの *minconf* で単語重みを用いない AQE-R より高く，妨害単語の追加を軽減させている．

正解率と F 値に関して単語重みを用いた AQE-C のみすべての *minconf* で初期クエリより改善されており，単語重みを用いない AQE-C は $minconf = 0.1$ のときを除き単語重みを用いた AQE-C を超えなかった．

また，AQE-C で拡張後クエリを正しく分類するにはクラスを正しく推定することが前提であるが，$minconf = 0.1$ における単語重みを用いない AQE-C と $minconf = 0.2$ における単語重みを用いた AQE-C では正解率と F 値がクラス推定よりも高い．これはクラス推定を誤ったが拡張後クエリを正しく分類した場合があるということであり，誤ったクラスにおけるアソシエーションルールで追加される単語に，正しいクラスを特徴付ける単語が含まれていたことを意味する．このような単語は，*minconf* が大きくなると

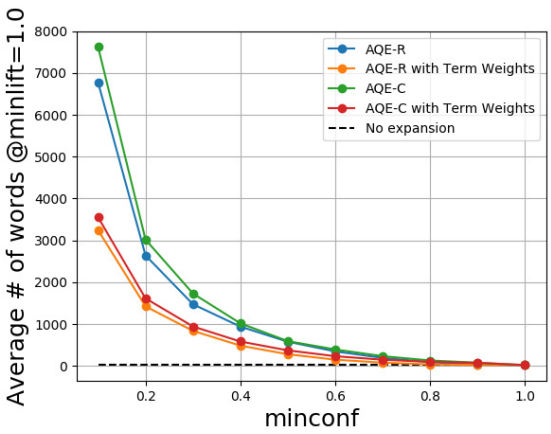

図 11 確信度の閾値に対するクエリ 1 件あたりの平均単語数の感度
Fig. 11 Sensitivity of average number of words in one query to *minconf*.

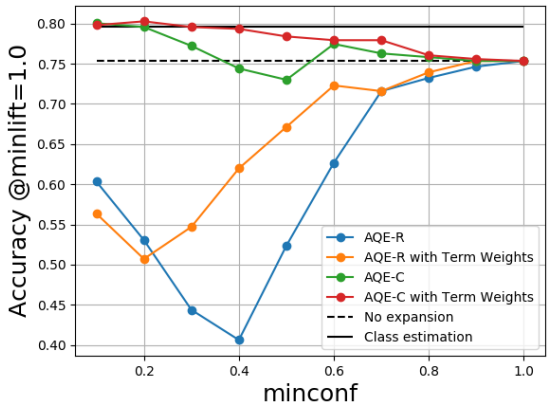

図 12 確信度の閾値に対する正解率の感度
Fig. 12 Sensitivity of accuracy to *minconf*.

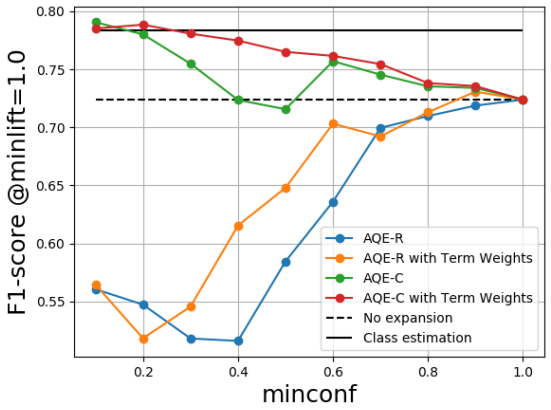

図 13 確信度の閾値に対する F 値の感度
Fig. 13 Sensitivity of F1 score to *minconf*.

正解率と F 値が低下していることから，推定されたクラスにおける確信度が小さいことが分かる．すなわち *minconf* が小さいとき確信度が低いルールで追加される推定されたクラスにとっての妨害単語が，クラス推定の誤りを修正する場合がある．

4.8 実験結果の考察

関連度の閾値に対する評価指標の感度分析から閾値によっては単語重みの考慮が有効でない場合があることが分かった．しかしながら図 9，10，12，13 に示すとおり，単語重みを用いた AQE-C は単語重みを用いない AQE-C よりも多くの閾値設定で分類性能を改善することができた．したがって，アソシエーション分析ではデータセットごとに適切な閾値を検討する必要があるが，AQE-C を用いるとき，単語重みを考慮することは単語重みを考慮しない場合より少ない閾値の検討で分類性能を改善できるという利点がある．

図 9，10 より，単語重みを用いた AQE-C ではきわめて高いリフト値を持つルールで追加される単語は妨害単語であることが分かった．したがってリフト値については閾値として下限 *minlift* だけではなく上限 *maxlift* も設定し，式 (3) におけるリフト値に関するアソシエーションルールの条件を $minlift < lift(r) < maxlift$ とすると，さらに有効なアソシエーションルール集合が生成されると考えられる．

本論文では文書分類タスクを扱ったが，比較対象として AQE-R についても拡張後クエリの分類性能を評価した．図 2，3，5 に示すとおり，AQE-R は AQE-C に劣るものの単語重みの考慮によって分類性能の改善が可能であった．このことから異なるクラスに属する文書を含む文書集合全体に対してアソシエーション分析を行う場合でも，単語重みを考慮することで妨害単語を追加するようなアソシエーションルールの生成を抑制可能であることが明らかとなった．

5. 関連研究

5.1 単語数の少ない文書の分類

単語数の少ない文書の分類を扱う主要な既存研究との違いを述べる．

Man の方法では本研究と同じく支持度と確信度が用いられている [10]．このうち支持度は単語の出現文書数に基づく関連度という点で本研究と同じであるが，本研究はここに文書に対する単語の重要度まで考慮するために単語重みを利用している点に相違がある．また，Man の方法では確信度を単語–クラス間の関連度として使用している．これに対し本研究では，クラスごとに分割した文書集合に基づくアソシエーションルールを用いることでクラスの特徴を獲得しようとする点に相違がある．

Dai らの方法は，クラスタリングによって得られた各クラスタとの類似度を新たな特徴として利用する [11]．外部知識を利用することなく，データセットを分析することでクエリを拡張（新たな特徴を獲得）するための知識を抽出するという点では本研究と同じであるが，その知識がクラスタ集合かアソシエーションルール集合かという違いがある．また，Dai らの研究ではデータセットの文書すべてが単語数の少ない文書であることを想定しており，アソシエーション分析では共起関係をとらえることが難しいと考えられる．これに対して本研究ではデータセットとしては単語数の多い文書が主であり，分類対象としてのクエリが単語数の少ない文書であることを想定している．たとえば，電子カルテデータをデータセットとして利用でき，問診情報のみをクエリとする場合などがあげられる．このような場合にはアソシエーション分析で共起関係をとらえることができるため，本研究ではアソシエーション分析に着目している．

Wei らは本研究と同じくアソシエーション分析を用いている [12]．Wei らの方法では関連度計算に単語重みを利用していないのに対し，本研究の提案手法では関連度計算に単語重みを利用する．また，クラスごとにアソシエーション分析を行うことでクラスごとの特徴をとらえようとする点も Wei らの方法にはない要素であり，本研究の新規性である．

5.2 単語重みを用いたアソシエーション分析

単語重みを用いたアソシエーション分析に関する研究として，Latiri らは文書に対する単語のメンバシップ関数を単語重みを用いて定義し，単語間の関連度計算において各単語のメンバシップ度が特定の条件を満たす文書を各単語が共起する文書と見なしている [13]．Latiri らの手法ではアソシエーション分析における各関連度の閾値とは別に，関連度を計算する各単語がある文書において共起するかどうかを判断する条件のために導入されるパラメータを設定する必要がある．本研究ではメンバシップ度を直接的に関連度計算に利用しており，設定する必要があるパラメータは各関連度の閾値のみである．また，Latiri らは情報検索タスクでクエリ拡張実験を行ったのに対し，本論文では文書分類タスクでのクエリ拡張実験によって，単語重みを考慮して生成されるアソシエーションルール集合が分類性能改善に対して有効であることを明らかにした．

5.3 アソシエーション分析を用いた文書分類

アソシエーション分析を用いた文書分類に関する研究として，Antonie らは単語があるクラスに属する文書に出現するときその単語とクラスが共起していると考えることで，単語間の関連度ではなく単語集合からクラスへの関連度を計算し，分類を行っている [14], [15]．ただしこのとき 3.1 節で述べたクラス推定と同様に，分類結果はクエリと

各訓練データ文書との類似度に基づいていない．これに対し本研究では，クエリ拡張によってクエリと訓練データ文書とのミスマッチを軽減し，クエリと各訓練データ文書との類似度に基づく分類を行う．これにより，クエリとの類似度が高い訓練データ文書を分類結果の根拠となる関連文書として参照できる．

5.4 知識抽出と外部知識の利用

アソシエーション分析はクエリ拡張および文書分類のためにデータセットから知識を抽出する方法であるといえる．しかしながら，既存の辞書や Wikipedia などの外部知識を利用する方法もある．ここでは本研究と外部知識を利用する既存研究との目的の違いについて述べる．

河合らの研究では，分類体系に対する汎用性の確保と分類精度の向上を目的として，意味属性体系を利用した文書分類手法を提案している [16]．また，中山らの研究では，幅広い分野で利用可能な基盤リソースの構築を目的として，Wikipedia を利用した概念間の関連度計算手法を提案している [17]．意味属性体系や Wikipedia などの外部知識を利用するこれらの手法は，文書分類およびクエリ拡張のモデルをデータセットから独立させ，汎用的に利用することを想定しており，そのうえで分類性能を改善することを目的としている．

これに対して，外部知識を利用せずにデータセットから知識抽出を行う方法は，外部知識がデータセットに対して適切でない場合にも利用することができる [11]．すなわち，本研究ではアソシエーション分析によってデータセットに特化した知識を抽出することを想定しており，そのうえで分類性能を改善することを目的としている．

6. おわりに

文書分類問題において分類対象となるクエリの単語数が訓練データより少ない場合にクエリ拡張によって分類性能を改善することを目的として，2 つの個別に適用可能な手法を提案した．1 つはアソシエーション分析に基づく自動的クエリ拡張を文書分類タスクに適用するための AQE-C という手法である．クラスごとに分割した文書集合に対してアソシエーション分析を行い，クラスに対する単語の重要度に基づいて初期クエリから推定されたクラスにおけるアソシエーションルール集合を用いてクエリ拡張を行うことで，分類性能を改善できることを示した．もう 1 つはアソシエーション分析における単語間の関連度計算に単語重みを利用するための手法である．単語重みに基づいて計算される文書に対する単語のメンバシップ度を直接的に関連度計算に用いることによって，生成されたアソシエーションルールによる拡張後クエリの分類性能を改善できることを示した．また，AQE-C では単語重みの考慮が多くの関連度の閾値設定で有効であることを示した．

提案手法の AQE-C でクラス推定を行うのは，クラスごとに生成したアソシエーションルール集合のうちいずれをクエリ拡張に用いるかを決定するためである．AQE-C は AQE-R に比べて分類性能は優れているが，クラス推定の性能に依存するところが大きく，初期クエリに正しいクラスを特徴付ける単語が存在しない場合にはクエリ拡張によって分類を妨害する単語を追加してしまう．一方，本論文では AQE-R についても拡張後クエリの分類性能評価を行うことで，クラスを考慮せず文書集合全体に対してアソシエーション分析を行う場合でも，単語重みを考慮することで単語重みを考慮しない場合より拡張後クエリの分類性能を改善可能であることを明らかにした．

本論文では拡張後クエリの単語数や正解率，F 値に関する評価を行ったが，関連度の閾値に対する評価指標の感度分析は限定的な範囲での分析であった．今後は，次の検討または分析を行う必要がある．

(1) 関連度の閾値の組合せについてのさらなる網羅的な検討，
(2) 生成されるアソシエーションルールが提案手法によりどう変化するかの分析，
(3) 分類に有用な単語と分類を妨害する単語の特徴分析．

参考文献

[1] Carpineto, C. and Romano, G.: A survey of automatic query expansion in information retrieval, *ACM Computing Surveys* (*CSUR*), Vol.44, No.1, p.1 (2012).
[2] Wei, J., Bressan, S. and Ooi, B.C.: Mining term association rules for automatic global query expansion: Methodology and preliminary results, *Proc. 1st International Conference on Web Information Systems Engineering, 2000*, Vol.1, pp.366–373, IEEE (2000).
[3] Agarwal, R., Srikant, R., et al.: Fast algorithms for mining association rules, *Proc. 20th VLDB Conference*, pp.487–499 (1994).
[4] Schütze, H., Manning, C.D. and Raghavan, P.: *Introduction to information retrieval*, Vol.39, Cambridge University Press (2008).
[5] 山田雄基，伊藤　豪，中田雅也，濱津文哉，濱上知樹：アソシエーション分析に基づく単語補完を用いた電子カルテデータの分類，第 44 回知能システムシンポジウム講演資料（CD-ROM），講演番号：B5-2 (2017).
[6] 安永　翼，中田雅也，濱上知樹：ファジーアソシエーション分析に基づく文書分類のための自動的クエリ拡張，第 45 回知能システムシンポジウム講演資料（CD-ROM），講演番号：A1-4 (2018).
[7] Dubois, D.J.: *Fuzzy sets and systems: Theory and applications*, Vol.144, Academic press (1980).
[8] 20 Newsgroups, available from ⟨http://qwone.com/~jason/20Newsgroups/⟩ (accessed 2017-12-08).
[9] livedoor ニュースコーパス，入手先 ⟨https://www.rondhuit.com/download.html#ldcc⟩（参照 2018-04-18）.
[10] Man, Y.: Feature extension for short text categorization using frequent term sets, *Procedia Computer Science*, Vol.31, pp.663–670 (2014).
[11] Dai, Z., Sun, A. and Liu, X.-Y.: Crest: Cluster-based representation enrichment for short text classification,

Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.256–267, Springer (2013).
[12] Wei, H., Shan-fei, L., Yue-jin, T. and Bing, G.: Association rules based short text feature extension, *IJCSNS*, Vol.9, No.10, p.227 (2009).
[13] Latiri, C.C., Yahia, S.B., Chevallet, J. and Jaoua, A.: Query expansion using fuzzy association rules between terms, *Proc. JIM* (2003).
[14] Antonie, M.-L. and Zaiane, O.R.: Text document categorization by term association, *Proc. 2002 IEEE International Conference on Data Mining*, *ICDM 2002*, pp.19–26, IEEE (2002).
[15] Zaïane, O.R. and Antonie, M.-L.: Classifying text documents by associating terms with text categories, *Australian Computer Science Communications*, Vol.24, No.2, pp.215–222, Australian Computer Society, Inc. (2002).
[16] 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol.33, No.9, pp.1114–1122 (1992).
[17] 中山浩太郎，伊藤雅弘，白川真澄，道下智之，原　隆浩，西尾章治郎：Wikipedia マイニング，人工知能学会論文誌，Vol.24, No.6, pp.549–557 (2009).

安永 翼　（学生会員）

1994 年生．2017 年横浜国立大学理工学部数物電子情報系学科卒業．同年から同大学大学院工学府博士課程（前期）で機械学習，自然言語処理，クエリ拡張の研究に従事．

山田 雄基

1992 年生．2015 年横浜国立大学理工学部数物電子情報系学科卒業．2017 年同大学大学院工学府博士課程（前期）修了．現在，NTT コミュニケーションズアプリケーション&コンテンツサービス部 AI 推進室に勤務．

濱上 知樹　（正会員）

1966 年生．1999 年千葉大学大学院自然科学研究科後期課程修了．2001 年同研究科助手．2004 年横浜国立大学大学院工学研究院助教授．2008 年 10 月同大学院教授，知能システム，機械学習，強化学習，医療支援システム，社会システムへの応用研究に従事．博士（工学）．計測自動制御学会，電気学会，電子情報通信学会，IEEE 各会員．