5L-6

ウェブからの分析対象文書抽出手法の検討

高山　泰博†　　今村　誠†　　鍛治　伸裕‡　　豊田　正史‡　　喜連川　優‡

三菱電機（株）情報技術総合研究所†　　東京大学生産技術研究所‡

1　はじめに

ウェブ文書には、実社会における関心が反映されており、文書中に含まれる製品や企業活動に対する評判などの情報を分析してマーケティングや風評監視等に活用することが期待されている。この分析の対象とするウェブ文書(分析対象文書)の抽出に全文検索を用いると、検索式のキーワードに多義語が含まれる場合には検索結果中に分析目的に合致しない文書が混在してしまう。そこで、本稿では、語の曖昧性を解消して分析対象文書を精度良く抽出する方式を検討する。

2　従来方式の課題

個々のウェブ文書中のテキストを事例文章データと呼ぶ。また、分析対象文書中の事例文章データの集合を求める際に、緩い検索条件(抽出対象を特徴付ける多義語自身をキーワードとする検索式)でウェブ全体から検索した結果の事例文章データの集合を抽出対象の「上限」と呼ぶ。ここでは、多義語の曖昧性を解消して、上限から分析目的に合致した事例文章データの集合を高精度に抽出することを目標とする。

多義語の曖昧性解消には、従来から Naive Bayes や SVM 等の機械学習の分類器による手法[(i),(ii)]が用いられている。これらの従来研究は、辞書に列挙された語義について新聞記事等で訓練データ、評価データを作成し、分類器のアルゴリズムを評価するものである。

一方、ウェブの場合には、語が製品名や会社名等に種々の使われ方をするため、どのような多義性があるかが未知であり、訓練データの作成方法が明確ではない。ここで、抽出対象である事例文章データの集合を「正例」、抽出対象でない事例文章データの集合を「負例」と呼ぶ。

多義性が未知な場合に訓練データを作成して文書を抽出する方式として、**図 1**に示す方式がある。**図 1**の方式を用いた予備実験では抽出精度(再現率)が低いという問題があった。これは、分類器に与えた訓練データの負例に多数の正例が含まれており、素性(分類に用いる情報)に影響したためと考えられる。したがって、訓練データを洗練化していく手続きを明確化することが課題となる。

(1) 緩い検索条件の全文検索で上限を作成する。
(2) 上限に対して、AND 条件を追加した厳しい検索条件で検索して、正例を作成する。この事例文章データの集合を「正例の下限」と呼ぶ。
(3) 上限から正例の下限を除いた残りの事例文章データの集合を負例とする。
(4) (2)の正例の下限、(3)の負例を分類器に訓練データとして与え、2 値分類を行う。

図1 分析対象文書抽出の従来方式

3　提案方式

3.1 提案方式の概略と用語の定義

提案方式の概略を図 2に示す。ここでは、訓練データの作成と自動分類(分類器への適用)を合わせて学習と呼ぶ。提案方式では、厳しい検索条件によるゴミの少ない全文検索結果で訓練データを作成して自動分類し、その分類結果を人手でサンプリングして分類に有効なキーワードを見つけ、訓練データに追加することを繰り返す。

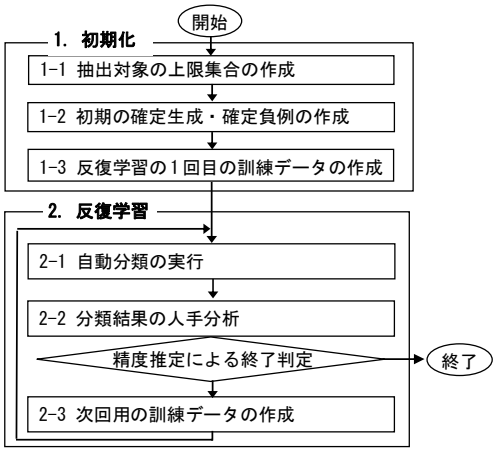

図2 分析対象文書抽出の提案方式の概略

ここでは、抽出対象であるという確信度が高い事例文章データの集合を「確定正例」$PosD_i$ と呼び、確定正例を全文検索する検索式を pos_i で表す。また、抽出対象でないという確信度が高い事例文章データの集合を「確定負例」$NegD_i$ と呼び、その検索式を neg_i で表す。添字 $_i$ は、反復学習の i 回目を表す。以下、添字を同様な意味で用いる。

事例文章データの集合の全文検索を search(検索式, 検索範囲)で表す。検索範囲は、検索式中の各キーワードの共起範囲であり、前後 n 文内の場合を $s_{\pm n}$（n≧0，文章全体の場合は n=∞）で表す。

A consideration on an extraction method of the documents for analysis from Web

†Yasuhiro TAKAYAMA, Makoto IMAMURA. Information Technology R&D Center, Mitsubishi Electric Corporation.

‡Nobuhiro KAJI, Masashi TOYODA, Masaru KITSUREGAWA. Institute of Industrial Science, The University of Tokyo.

分類器に正例として与える事例文章データの集合を「分類用正例」$PosC_i$ と呼び、負例として与える事例文章データの集合を「分類用負例」$NegC_i$ と呼ぶ。分類用正例と分類用負例を合わせて分類器のための訓練データ TD と呼ぶ。分類器が推定フェーズで分類するデータを分類対象データ CD と呼ぶ。

3.2 分析対象文書の抽出手続き

提案方式の具体的な文書抽出手続きを述べる。

ステップ1：初期化

(1-1) 抽出対象の上限の作成

緩い検索式（初期検索式 sup）によるモレの少ない全文検索結果を上限 Sup とする。

【例】sup=抵抗(*)　Sup:=search(抵抗, $s_{\pm\infty}$)

(1-2) 初期の確定正例と確定負例の作成

(i) 初期の確定正例 $PosD_0$:= search(pos_0, $s_{\pm n}$)

初期検索式を詳細化した厳しい検索条件で上限を検索し、ゴミの少ない正例を作成する。

【例】pos_0 = 抵抗 AND 電気

(ii)初期の確定負例 $NegD_0$ の作成 $NegD_0 := \phi$

(1-3) 反復学習の1回目の訓練データの作成

(i) 1回目の分類用正例　$PosC_1 := PosD_0$

(ii) 1回目の分類用負例　$NegC_1$:= Sup − $PosD_0$

上限から初期の確定正例を除いたものを1回目の分類用負例とする。上記は**図 1**の従来方式の訓練データに相当する。

ステップ2：反復学習

(2-1) 自動分類

分類用正例、分類用負例を訓練データとし、上限中の確定正例、確定負以外のデータを分類対象データとして分類器に与え、推定スコアを得る。

・入力：

- 訓練データ　　TD = ($PosC_i$, $NegC_i$)
- 分類対象データ CD = Sup − ($PosD_i$ + $NegD_i$)

・出力：分類対象データ毎の推定スコア

(2-2) 分類結果の人手分析

(i)精度評価のための分析

①自動分類結果を無作為抽出し、抽出精度(再現率、適合率)、事前確率 $prbPre_i$ を推定する。

②目標の抽出精度が得られれば、終了する。

(ii) 訓練データ洗練化のための素性分析

分類に有効なキーワードを見つけ、確定正例、確定負例の検索式にキーワードを追加する。

【例】pos_2 = 抵抗 AND (電気 OR 回路 OR 浮遊)

【例】neg_2 = 抵抗 AND (勢力 OR 改革)

(iii) 訓練データ洗練化のための追加データ作成

(i)、(ii)の分析において、キーワードで特徴づけられない追加訓練データを作成する。i 回目の分析で得た「全文検索で特徴づけられない正例文章データの集合」を $PosH_i$、「全文検索で特徴づけられない負例文章データの集合」を $NegH_i$ とする。

(*) 電気部品の「抵抗」に関する文書を、社会に対する「抵抗」等との曖昧さを解消して抽出する例で説明する。

(2-3) 次回用の訓練データの作成

(i)分類用正例の作成

①確定正例 $PosD_{i+1}$:= $PosD_i$ + $PosH_i$ + search(pos_{i+1}, $s_{\pm n}$)

②分類用正例 $PosC_{i+1}$:= $PosD_{i+1}$

(ii)分類用負例の作成

①確定負例 $NegD_{i+1}$:= $NegD_i$ + $NegH_i$ + search(neg_{i+1}, $s_{\pm n}$)

②分類用負例 $NegC_{i+1}$:= $NegD_{i+1}$ + chooseNeg({$PosD_i$, (Sup − $PosD_i$ − $NegD_i$)}, {Sup − $PosD_i$ − $NegD_i$})

(2-3)(ii)②の chooseNeg は、正例との類似度が小さい文章データの集合（想定負例）を求める関数であり、chooseNeg({分類用正例,分類用負例},{分類対象データ})である。確定正例を分類用正例とし、上限から確定正例と確定負例を除いたデータの集合を分類用負例とし、分類用負例と同じデータを分類対象データとして自動分類した結果から事前確率 $prbPre_i$ または分類スコアの閾値により想定負例を選択することを表す。(2-3)(ii)②で、確定負例に関数 chooseNeg で選択した想定負例を加えることで、分類用負例を拡充できる。

4　実験

日用品の製品名(多義語)を題材に、分類器に Naive Bayes を用いて文書抽出実験を行った。実験データは、blog 記事 12,381 文書、対象の多義語の出現数 19,805 件、正解数 7,861 件(抽出対象の製品名の場合)である。

提案方式の抽出精度は、再現率と適合率の Break Even Point が従来手法に対して 3.5%向上した。再現率は 24.4%向上して 94.5%になり、適合率は 1.2%低下して 96.6%となった。上記から、提案方式では、適合率をほぼ維持したまま再現率を向上できることを確認した。

5　まとめと今後の課題

本稿では、語にどのような多義性があるかが未知なウェブから、系統的に訓練データを作成して、分析対象文書を抽出する方式を検討した。今後は、文書の種類ごとに正例、負例の追加にどのような効果があるかを明確にする予定である。

謝辞

本研究の一部は文部科学省リーディングプロジェクト e-Society 基盤ソフトウェアの総合開発「先進的なストレージ技術および Web 解析技術」による。

参考文献

(1)白井清明: SENSEVAL-2 日本語辞書タスク, Vol.10, No.3, 自然言語処理 (2003).

(2)Manning, C.D., Schuetze, H.:Foundations of Statistical Natural Language Processing, MIT Press (1999).