1U-1

Webからの意見情報抽出とその集約

鷲田 基
東京大学 工学部
電気工学科*

岡崎 直観
東京大学 情報理工学系研究科
電子情報学専攻†

石塚 満
東京大学 情報理工学系研究科
電子情報学専攻†

1 はじめに

ウェブを用いた情報発信が一般に広く浸透する一方，その情報を探すことは一般に難しい．情報検索技術の進歩により，クエリを用いて自分の知らない単語を探したり，その意味を調べることはある程度可能になってきている．しかし，ある対象に関する意見を集約することは，現在の検索エンジンの枠組みでは困難である．本研究は，Web を情報源として，ある対象の意見を述べていると思われる文書を検索するのに適切なクエリを自動生成し，その適合文書を取得する．そして，その適合文書の中から，意見を述べている箇所を機械学習により抽出する．最後に，その抽出された箇所を集約して，ユーザに提示する．

2 関連研究

Web 上にある適合文書から意見を表す文を抽出する手法として，Yu ら [1] による手法があげられる．Hu らはナイーブベイズ分類器を使った機械学習でそれぞれの文を「事実」「ポジティブな意見」「ネガティブな意見」「両方の意味を持つ意見」「どちらでもない意見」に分類している．

その手法として，Yu らは単語，N-gram とその品詞，連続した単語の正負を素性として用いている．

*Electrical Engineering Department, University of Tokyo
†Department of Information and Communication Engineering, School of Information Science and Technology, University of Tokyo

3 手法

本研究では，キーワードを使って抽出した Web 上の適合文書から，意見を述べている文を含むと思われるパラグラフを順に表示した．このとき，類似している意見を述べていると思われる文は一つに集約した．大まかなプロセスを図 1 に示す．

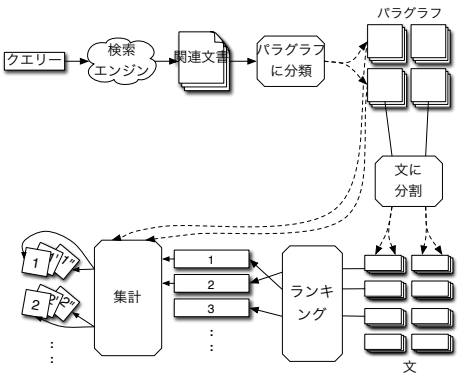

図 1: 手法の概要

適合文書の抽出　検索時は，まずキーワードを Google に渡し，検索結果の上位にある HTML の文書を適合文書とした．その中から，それぞれのテキストを解析してパラグラフを抜き出した．さらにパラグラフを文に分解し，それぞれの文を SVM にかけ，意見文であると思われる分の順に並べ替えた．最後に，それぞれの意見文が属するパラグラフを表示した．

パラグラフへの分解　適合文書からパラグラフに分解するために，まず< p >などのタグで分けられた部分は違うパラグラフであるとした．次に，空行によって分けられた部分は違うパラグラフであるとした．また，行頭にある「>」のような，ある種の特別な意味を持っていると思われる表記が現れた場合，それが連続して現れて

いる部分を一つのパラグラフとした．最後に，どうしてもパラグラフが長くなってしまうケースがあったため，一定以上パラグラフが長くなった時は，分割することとした．

文への分解　意見を述べているかの判定や，類似している意見の検索などは文を基準に行っているため，パラグラフを文に分解する必要がある．分解には句読点などの，文の間を分割している可能性が高い単語で分割した．

文が意見を表しているかどうかの振り分け　それぞれの文が意見を示しているかどうかの判定には，単語単位の N-gram を素性として Support Vector Machine（SVM）を使った機械学習にもとづいて行った．このプロセスの流れは図 2 の通りである．学習には「意見を述べている文」と

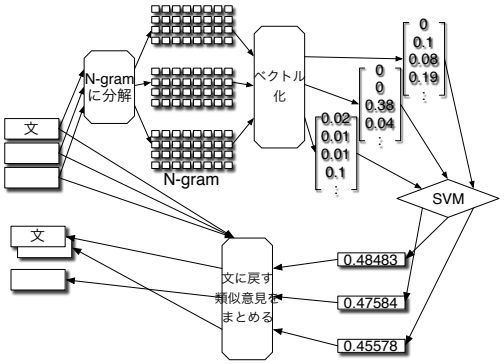

図 2: 意見文の判定

「そうでない文」にあらかじめタグ付けされている文から単語単位の N-gram を取り出し，SVM を使って学習させた．タグ付けされた文は，毎日新聞の 98 年から 99 年までの社説に掲載された文を，手動でタグ付けすることによって得た．

判定には，それぞれの文から N-gram を取り出し，それを素性として SVM にかけた．得られた出力から，それぞれの文を「意見文」らしさで並べ替えることができるので，その中の上位の文を意見文とした．上位のどのくらいかを意見文と判断するかによって，再現率と精度を調節することができる．

類似した意見を述べている文の判定　類似している意見を述べているかの判定には，それぞれの文を単位ベクトルとして表し，内積を使って計算した．文を表すベクトルは，N-gram を基底として，対象の文に含まれる N-gram の出現の仕方からベクトルを構成し，それを長さで割ることによって得た．二つの文の類似度は，それらを表す 2 つのベクトルの内積とした．

4　実験

意見文の振り分けの性能を見るため，3000 程度の学習に使われていない文を分類し，再現率を調節しつつ，それに対する精度を調べた．ここで，再現率とは実際に意見を述べている文のうち，分類器が意見を述べていると判断した文の割合である．一方，精度とは分類器が意見を述べていると判断した文のうち，実際に意見を述べている文であった割合である．

5　結果・考察

精度と再現率の関係は図 3 のようになった．再

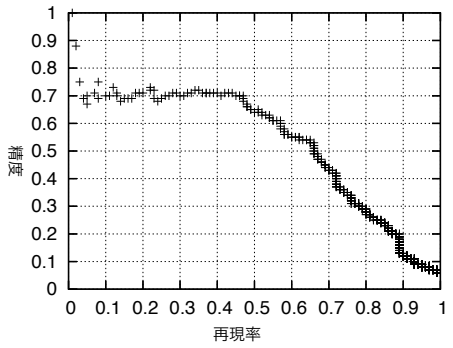

図 3: 精度と再現率の関係

現率が 0.4 くらいまでの所では精度は 0.7 程度であり、それ以上の所では再現率が落ちている。実際にストレスないレベルは 9 割近くまでいくことが重要と思われ、改善することが重要であると考える。改善策として、主に Web の様々な文書を使って学習データを増やすなどが考えられる。

参考文献

[1] Hong Yu, Vasileios Hatzivassiloglou. *Towards Answering Opinion Questions: Separating Facts from Opinion and Identifying the Polarity of Opinion Sentences*