E-018

単語論理和での属性拡張による文書分類性能向上
Disjunction of Words as Generic Concept for Document Classification

廣川佐千男 ‡
Sachio Hirokawa

1. はじめに

文書中での単語の出現だけでなく、単語間の関連を利用することで文書検索や文書識別の性能向上が期待されている。例えば、単語の同義語や上位下位関係のデータベースである WordNet を文書分類に利用する研究がある[1,3, 6,7]がある。しかし、上位下位関係を用いること で、常に性能が向上するとは限らない。また、WordNet のような大規模コーパス は汎用ではあるが人手で構築されたものであり、新しいテーマや特定の文書群 について全ての単語の関連を網羅しているわけではない。一方、[2]や [4]では、文書や Web ページに現れるパターンに着目することで、人手によらず単語間の上位下位関係を抽出している。しかし、一つの単語、あるいは、 複数の単語の上位あるいは下位の概念として求まるのは、分析対象の文書群に出現する具体的な単語だけである。

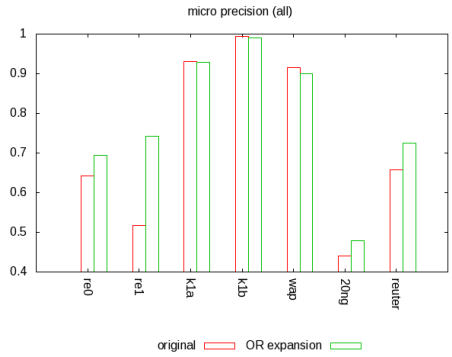
本稿では、二つの単語 u と v の形式的な論理和 u+v を上位概念とする定式化を提 案する。しかし、二つの単語の和 u+v はいくらでも考えられる。本稿では、論理和 u+v を仮想的な単語とすることで、文書分類の識別性能が向上するときに、 u+v は意義がある、と考える。つまり、u+v を単独の概念として対象文書をベ クトル化したとき、文書の識別性能が向上すれば、新たに追加した u+v が正例 と負例を識別する上位の意味ある概念になっている、と考えられる。それでも、任意の 単語 u と v の組み合わせでは個数が爆発する。そこで、本稿では、与えられた正例と負例を識別するための特徴語の上位 20 個の単語群についてのみ、論理和を考える。また、u+v に対応する文書集合、つまり、u または v を含む文書集合が、 u を含む文書集合あるいは v を含む文書集合と一致していたら新に u+w を考えても 意味がない。そこで、本稿ではそのような組み合わせは除外する。

本稿では、このような単語の論理和を追加して文書をベクトル化したとき、識別性能が向上するかどうかを、7 種類の標準的データセットについて調べた。

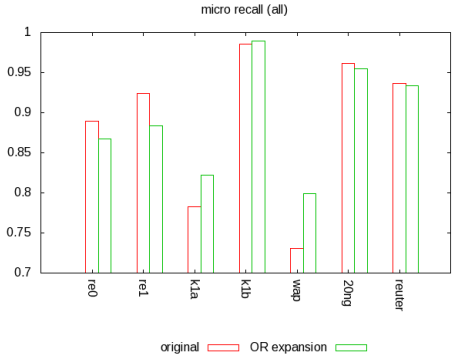
2. 実験結果

[5]では、正例の文書集合と負例の文書集合を線型カーネルの SVM で識別 するとき得られるモデルにおいて、各単語の重みの上位の単語に限定して文書をベクトル化すること（属性選択）で、識別性能が向上する場合があることが示されている。本稿では、この単語の重みを使い、正のスコアの単語上位 20 個 と負のスコアの単語上位 20 個について単語の論理和を新たな概念として追加す る。しかし、特徴語 ui と uj が交わらないという条件だけだと、ui+uj という論理 和での文書集合と単独の単語 ui での文書集合が同じになる場合がある。これだ と、ui+uj という属性を追加しても意味がない。そこで、ui+uj が ui あるいは uj とは同じにならない、という条件も満たす場合に限定して論理和 ui+uj を新たな 属性として追加するようにした。具体的には、各カテゴリーc_kで SVM スコア上位 20 までの単語の組 ui,uj について、論理和 ui+uj を満たす文書数が単独の単語 ui, uj を含むものよりも多い場合についてだけ、論理和 ui+uj を新たな属性と して追加する。

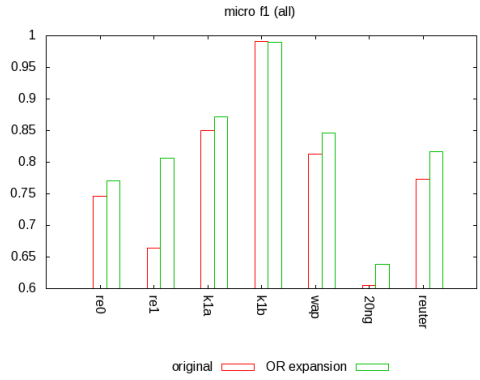
図 1 に 5 つのデータセット reuter21578, 20newsgroups, wap, k1a, k1b に対する precision、recall、F1-score、accuracy を示す。棒グラフで 左側(original)が単語ベクトル化したときの識別性能、右側(OR expansion)が 論理和を追加したベクトル化での識別性能(micro average)を示す。20 newsgroups の recall と wap の precision 以外の全ての場合について、提案手法により性能が向上し ていることが分る。



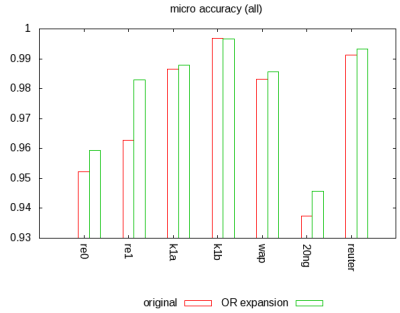
(a) precision



(b) recall

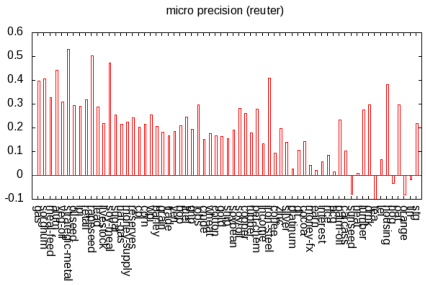


(c) F1-score

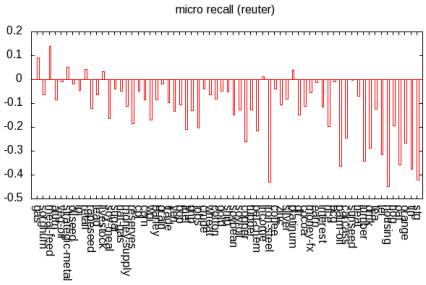


(d) accuracy

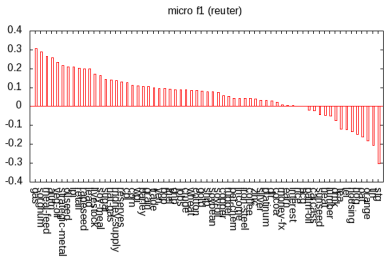
図 1　単語和による識別性能



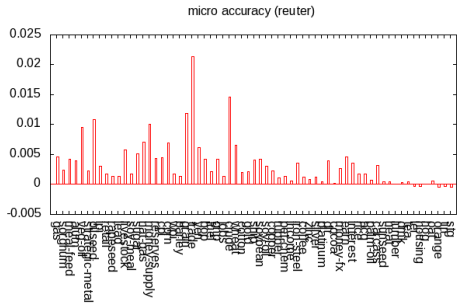
(a) precision



(b) recall



(c) F1-score



(d) accuracy

図 2　Reuter についての単語による識別性能向上

図 2 は reuter について、横軸にカテゴリをとり、縦軸に提案手 法による識別性能(micro average)の向上をプロットしたものである。殆どのカ テゴリーで precision も accuracy も向上している。いくつかのカテゴリーで recall が減少しているが、precision と recall の調和平均である F1-score は殆どのカテゴリ で増加しているので、precision の向上率が accuracy の低下率を上 回っているといえる。

3. まとめと今後の課題

文書識別のための特徴的な単語の二つの論理和として、文書中に明示的に単語 として出現していない上位概念を捉える方法を提案した。5 種類のベンチマークのデータセットについて提案手法による文書識別性能を、標準的なベクトル化 による識別性能と比較し、提案手法の有用性能を示した。本稿では、論理和を 考える単語を 40 個数の特徴語に限定したが、この個数を変化させたとき、性能 がどのように違ってくるかが今後の課題である。また、具体的な単語の論理和 u+v が、何らかの意味のある概念を表しているかどうかの評価が必要である。

参考文献

[1] 福本文代,鈴木良弥,2002, WordNet の同義語クラスとその上位関係を利用した文 書の自動分類, 情報処理学会論文誌, Vol.43, No.6, pp.1852-1865
[2] M. A. Hearst, 1992, Automatic acquisition of hyponyms. from large text corpora, Proceedings of the 14th International Conference on Computational Linguistics, pp. 539-545
[3] A. Hotho, S. Staab, G. Stumme,2003, Ontologies improve text document clustering, Proceedings - IEEE International Conference on Data Mining, ICDM , pp.541-544
[4] 新里圭司,鳥澤健太郎, 2003, HTML 文書からの単語間の上位下位関係の自動獲得, 情報処理学会研究報告自然言語処理, 108(2003-NL-158),pp.95-102
[5] T. Sakai, S. Hirokawa, 2012, Feature words that classify problem sentence in scientific article, Proceedings of 14th International Conference on Information Integration and Web-Based Applications and Services, pp. 360-367
[6] S. Scott, S. Matwin, Text Classification Using WordNet Hypernyms, 1998, in: COLING/ACL Workshop in Usage of WordNet on NLP Systems, pp.45-51
[7] 上嶋宏,三浦孝夫,塩谷勇,2004, 同義語,多義語の考慮による文書分類の精 度向上, 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol.2, pp.137-144