領域相關詞彙極性分析及文件情緒分類之研究

Domain Dependent Word Polarity Analysis for Sentiment Classification

游和正 Ho-Cheng Yu
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
p98922004@ntu.edu.tw

黃挺豪　Ting-Hao (Kenneth) Huang
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
r96944003@ntu.edu.tw

陳信希 Hsin-Hsi Chen
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
hhchen@ntu.edu.tw

摘要

情緒分析乃近年來發展迅速之一熱門研究領域[1,2]，旨在透過文本分析技術探討作者之意見傾向與情緒狀態。其中，以情緒詞與情緒詞典為基礎之各種方法尤為知名。然而，情緒詞之情感傾向及其行為於不同領域文本中之行為並不盡然相同。本研究聚焦於情緒詞彙於不同領域文本中之行為，對房地產、旅館、和餐廳等三種不同領域之文本進行分析，並發現部分情緒詞彙於不同領域文本中的情緒傾向非但有差異，甚至彼此衝突。此外，部分未收錄於情緒詞典中之「非情緒詞」，在特定領域中亦可能成為「領域相依」之詞彙，影響情緒分類。本研究繼而提出不同詞彙權重計算方式，將此資訊加入舊有情緒分類系統中。在使用 LIBSVM 的線性核函數方式，對房地產、旅館、和餐廳等三種語料使用 5 次交叉驗證方式進行分類。實驗結果顯示所提出之 TFSSIDF 分類方法，結合 TFIDF、臺灣大學情感詞典，及計算語料之領域極性情感傾向程度(SO)，強化領域相關及領域不相關之情緒詞之權重，通過 t 檢定有效提升各領域中文件分類之效能[3,4]。

Abstract

The researches of sentiment analysis aim at exploring the emotional state of writers. The analysis highly depends on the application domains. Analyzing sentiments of the articles in different domains may have different results. In this study, we focus on corpora from three different domains in Traditional and Simplified Chinese, then examine the polarity degrees of vocabularies in these three domains, and propose methods to capture sentiment differences. Finally, we apply the results to sentiment classification with supervised SVM learning. The experiments show that the proposed methods can effectively improve the sentiment classification performance.

關鍵詞：文件情緒分類、詞彙極性分析、機器學習

Keywords: Document Sentiment Classification, Word Polarity Analysis, Machine Learning

實驗結果

下表是採用單一詞彙不同詞彙權重、在三種不同領域文件的情緒分類結果，評估的標準是準確率。僅使用單一之 TFSO 或 TFIDF 的效果很接近，但是將 IDF 與 SO 相乘，也就是 TFSOIDF，其效果更好，TFSOIDF 優於其他兩種。若結合情感辭典，可將分類效果更進一步提昇。表中呈現 TFSSIDF 優於 TFSOIDF，TFSDIDF 優於 TFIDF。總結，Unigram 的結果以 TFSSIDF 為最佳，TFSOIDF 與 TFSDIDF 次之，接著是 TFIDF，與其他方法。(註：TF: 詞彙頻率，IDF: 逆向文件頻率，SO: 情感強烈程度，SD: 情緒詞典)

語料	TFIDF	TFRF	Delta	TFSO	**TFSOIDF**	TFSDIDF	**TFSSIDF**
房地產	0.848	0.849	0.853	0.847	0.854	0.852	**0.863**
旅館	0.916	0.906	0.914	0.915	**0.924**	0.918	0.923
餐廳	0.861	0.839	0.849	0.854	0.871	0.869	**0.875**

參考文獻

[1] Bo Pang and Lillian Lee, “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, issue 1-2, pp. 1-135, 2008.

[2] Lun-Wei Ku and Hsin-Hsi Chen, “Mining Opinions from the Web: Beyond Relevance Retrieval,” *Journal of American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1838-1850, 2007.

[3] Man Lan, Sam-Yuan Sung, Hwee-Boon Low, and Chew-Lim Tan, ”A Comparative Study on Term Weighting Schemes for Text Categorization,” In *Proceedings of 2005 IEEE International Joint Conference on Neural Networks*, pp. 546-551, 2005.

[4] Justin Martineau and Tim Finin, “Delta TFIDF: An Improved Feature Space for Sentiment Analysis,” In *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, pp. 258-261, 2009.