中国科学院文献情报中心主办

现代图书情报技术
NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE

基于文体学的中文UGC作者身份识别研究

吕英杰[1], 范静[2], 刘景方[3]

1. 北京化工大学经济管理学院 北京 100029;
2. 北京外国语大学国际商学院 北京 100089;
3. 上海交通大学安泰经济与管理学院 上海 200052

Lv Yingjie[1], Fan Jing[2], Liu Jingfang[3]

1. School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China;
2. International Business School, Beijing Foreign Studies University, Beijing 100089, China;
3. Antai College of Economics and Management, Shanghai Jiaotong University, Shanghai 200052, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (398KB) HTML (1KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 网络的开放性和虚拟性给发布信息的作者身份识别造成很大困难,因此探索性地提出通过对网上的用户生成内容(UGC)的写作特点进行分析来识别其作者身份的方法。在传统的文体学研究基础上,结合中文UGC的特点,提取出词汇特征、句法特征、结构特征和内容特征等4类能有效识别不同作者写作风格的特征,然后运用文本分类算法对作者身份进行有效识别。通过实验表明在BBS论坛文本和博客文本这两种典型的中文UGC环境中,本研究采用的方法均得到很好的识别效果。

关键词： 文体学 用户生成内容 作者识别

Abstract： The characteristics of information network such as openness and virtuality make it difficult for authorship identification. Therefore, this paper proposes the approach of authorship identification of Chinese UGC based on stylistics. The authors integrate four types of features including lexical, syntactic, structural and content-specific features to compose writing-style features, and then use text classification technologies for authorship identification. The experimental results demonstrate that the proposed approach can be used for authorship identification of Chinese UGC efficiently.

Keywords： Stylistics, UGC, Authorship identification

收稿日期: 2013-04-18;

基金资助:

本文系国家自然科学基金项目“我国电子政务标准的产生机制及采纳扩散研究”(项目编号:71103021)和北京市哲学社会科学规划项目“北京市G2G电子政务业务协同的动力机制、推进方法与实证研究”(项目编号:13JGC085)的研究成果之一。

通讯作者 吕英杰 Email: luyingjie982@163.com

引用本文:

吕英杰, 范静, 刘景方 .基于文体学的中文UGC作者身份识别研究[J] 现代图书情报技术, 2013,V29(9): 48-53

Lv Yingjie, Fan Jing, Liu Jingfang .Authorship Identification of Chinese UGC Based on Stylistics[J] , 2013,V29(9): 48-53

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2013/V29/I9/48

Service

- 把本文推荐给朋友
- 加入我的书架
- 加入引用管理器
- Email Alert
- RSS

作者相关文章

- 吕英杰
- 范静
- 刘景方

[1] 孙晓明,马少平.基于写作风格的作者识别[C]. 见: 中国中文信息学会二十周年学术会议论文集. 北京:清华大学出版社,2001:198-204.(Sun Xiaoming,Ma Shaoping. Author Identification Based on Stylometric Approach[C].In: *Proceedings of the 20th Anniversary Chinese Information Processing Society of China*. Beijing: Tsinghua University Press, 2001: 198-204.)

[2] Efron B, Thisted R.Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?[J].*Biometrika*,1976,63(3):435-447.

[3] 张运良,朱礼军,乔晓东,等.基于句类特征的作者写作风格分类研究[J]. 计算机工程与应用,2009,45(22):129-131,223.(Zhang Yunliang,Zhu Lijun,Qiao Xiaodong,et al.Research on Text Authorship Categorization Based on Sentence Category Features[J]. *Computer Engineering and Applications*,2009,45(22):129-131,223.)

[4] 张凯,张明允.基于SVM的《红楼梦》写作风格研究[J]. 贵阳学院学报:自然科学版,2011,6(1):55-57.(Zhang Kai,Zhang Mingyun.Research on the Writing Style of “Dream of the Red Chamber” Based on SVM[J].*Journal of Guiyang College: Natural Sciences*,2011,6(1):55-57.)

[5] 年洪东,陈小荷,王东波.现当代文学作品的作者身份识别研究[J]. 计算机工程与应用,2010,46(4):226-229.(Nian Hongdong,Chen Xiaohe,Wang Dongbo. Research on Authorship Attribution of Contemporary Literature[J].*Computer Engineering and Applications*,2010,46(4):226-229.)

[6] 武晓春,黄萱菁,吴立德.基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 2006,20(6):61-68.(Wu Xiaochun,Huang Xuanjing,Wu Lide. Authorship Identification Based on Semantic Analysis[J]. *Journal of Chinese Information Processing*,2006,20(6):61-68.)

[7] De Vel O,Anderson A,Corney M,et al.Mining E-mail Content for Author Identification Forensics[J]. *ACM SIGMOD Record*,2001,30(4):55-64.

[8] Zheng R,Li J,Huang Z, et al.A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques[J].*Journal of the American Society for Information Science and Technology*,2006,57(3):378-393.

[9] Abbasi A,Chen H.Identification and Comparison of Extremist-group Web Forum Messages Using Authorship Analysis [J]. *IEEE Intelligent Systems*,2005,20(5):67-75.

[10] Holmes D I,Forsyth R S.The Federalist Revisited:New Directions in Authorship Attribution[J].*Literary and Linguistic Computing*,1995,10 (2):111-127.

[11] Juola P,Baayen H.A Controlled Corpus Experiment in Authorship Identification by Cross-entropy[J]. *Literary and Linguistic Computing*,2005,20 (S):59-67.

[12] Abbasi A,Chen H. Writeprints:A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace[J]. *ACM Transactions on Information Systems*,2008,26(2):1-29.

[13] Salton G,Buckley C.Term-weighting Approaches in Automatic Text Retrieval [J]. *Information Processing and Management*,1988,24 (5):513-523.

[14] Battiti R.Using Mutual Information for Selecting Features in Supervised Neural Net Learning [J]. *IEEE Transactions on Neural Networks*,1994,5(4): 537-550.

[15] Yang Y,Pederson J O.A Comparative Study on Feature Selection in Text Categorization [C].In: *Proceedings of the 14th International Conference on Machine Learning*.1997:412-420.

[16] Friedman N,Geiger D,Goldszmidt M. Bayesian Network Classifiers[J].*Machine Learning*,1997,29 (2-3):131-163.

[17] Quinlan J R.C4.5:Programs for Machine Learning [M]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.,1993.

[18] Cortes C,Vapnik V.Support-Vector Network[J].*Machine Learning*,1995,20 (3):273-297.

[1] 赵辉, 刘怀亮.面向用户生成内容的短文本聚类算法研究[J]. 现代图书情报技术, 2013,29(9): 88-92