学校编码：10384　　分类号＿＿＿密级＿＿＿
学号：23020081153263　　UDC＿＿＿

厦门大学

硕 士 学 位 论 文

基于条件随机场的中文文本情感分析研究

Research on Sentiment Analysis Based on Conditional Random Fields

徐 超

指导教师姓名：张东站　副教授

专 业 名 称：计算机软件与理论

论文提交日期：2012 年　　月

论文答辩时间：2012 年　　月

学位授予日期：2012 年　　月

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿＿＿

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

摘 要

伴随着互联网的发展以及新的网络应用的出现，互联网用户由单纯的“读”网页，向“读、写”网页，共同建设互联网发展，由此网上产生了大量带有用户主观感情的数据，从这些带有主观感情的海量数据中挖掘出网络用户的观点在现实生活中具有很大的意义，在这种背景下文本情感分析(sentiment analysis)技术应运而生。

文本情感分析近年来成为一个新的研究热点。根据研究的粒度大小不同，文本情感分析可分为词汇级、句子级以及文档级，然而句子级以及文档级的情感分析往往以词汇级的为基础。目前对于词汇级的情感分析研究有基于规则的方法和基于机器学习方法，针对中文文本词汇级情感分析大部分使用基于规则的方法。基于规则的方法，通常具有较高的准确率，但召回率很差，为此本文对此进行了深入的研究。

首先，针对基于规则的情感词汇自动识别算法的不足，提出了基于条件随机场的情感词汇识别算法，该方法将情感词汇识别当成序列标注任务，并给出了标注集、特征模板以及特征验证与特征筛选的方法。实验结果验证了基于条件随机场的情感词汇自动识别算法的有效性。

其次，针对传统基于词汇相似度的词汇褒贬性自动识别算法的不足，本文提出了一种改进的基于词汇相似度的褒贬性自动识别算法。该算法以基于聚类的褒贬基准词选择算法为基础，基于词群的概念克服了词汇间相似性与词汇间褒贬倾向不一致对词汇褒贬性计算的影响。实验结果表明，该方法在准确性方面优于传统基于词汇相似度的词汇褒贬性计算算法。

最后，给出了一种评论文摘自动生成的方法，该方法基于评价搭配抽取，评价短语极性判断以及评价对象聚类。实验结果验证了该方法的可行性。

关键字：情感分析；条件随机场；评论文摘

Abstract

With the rapid development of Internet and the emergency of new network application, Users of Internet begin to "writing" page and join in the construction of the Internet from only "reading" page. And more and more subjective data with users' sentiment is created by Internet users, and It's more useful of the summarized sentiment retrieved from the subjective data. Under this context, text sentiment analysis technology came into being.

Text sentiment analysis in recent years become a new hot research field. The text sentiment analysis can be divided into the vocabulary level, sentence level and document level based on the studying granularity. Recently studying of text sentiment analysis of vocabulary level has rule-based method and machine learning based methods. And the most used method for Chinese text sentiment analysis is rule-based method. Rule-based methods, usually with a higher accuracy rate, but recall rate is poor. To this point, a thorough research for Chinese text sentiment analysis is made in this paper.

First, for the inadequate of rule-based method, a new sentiment word recognition method based on CRF is given. The sentiment word recognition is trade as Sequence tagging in the method and related tagging set, feature templates and the methods of feature validation and feature selection are given. The experimental results verify the validity of the automatic sentiment word recognition based on CRF.

Secondly , for the inadequate of traditional sentiment trend computing base on word similarity, an improved sentiment trend computing method based on word similarity is presented in this paper. The method is based on the benchmark sentiment word selection algorithm using clustering algorithm, which overcomes shortcoming of the inconsistent between the words similarity and sentiment trend between the two word. The experimental results show that the method is better than the traditional word sentiment trend computing method in accuracy.

Finally, a method used for automatically retrieving sentiment abstracts is introduced. This method is based on evaluation relation extracting, sentiment trend

computing of evaluation phase and evaluation objects clustering. The experimental results shows the feasibility of this method.

Key words: sentiment analysis; conditional random fields; sentiment abstract

目 录

Contents