

**MINING AND ANALYZING CUSTOMER
OPINIONS/SENTIMENTS OF WEB 2.0 FOR
BUSINESS APPLICATIONS**

KAIQUAN XU

**DOCTOR OF PHILOSOPHY
CITY UNIVERSITY OF HONG KONG
AUGUST 2011**

CITY UNIVERSITY OF HONG KONG

香港城市大學

**Mining and Analyzing Customer
Opinions/Sentiments of Web 2.0 for Business
Applications**

挖掘和分析 Web 2.0 之客戶意見和情感來實現商
務应用

**Submitted to the
Department of Information Systems
資訊系統學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位**

by

Kaiquan Xu
许开全

AUGUST 2011
二零一一年八月

ABSTRACT

With the explosive growth in the amount of user-generated review data in the era of Web 2.0 (social networking sites, blogs, mini-blogs, discussion forums, online shopping websites), there is a pressing need to develop effective methods and tools to automatically extract valuable business intelligence from these user opinion data. In addition, the large numbers of user reviews often contain information about competitors and have become a new source for mining competitive intelligence. Many companies have begun to use social networking sites (SNS) as an important channel and platform to do online marketing and reputation management. Thus, analyzing users' sentiments on SNS has become key for these business applications.

Semantically annotating opinion data is an effective way to mine valuable information from the large number of customer opinions. Although supervised machine learning approaches have been explored for semantically annotating user opinions to facilitate market intelligence generation, such approaches often require numerous manually labelled training examples to produce accurate semantic annotations. In this study, we propose an active learning approach that can train a state-of-the-art, large-margin classifier with substantially fewer labelled training examples, and yet produce accurate semantic annotations of user opinions. In particular, our active learning method is underpinned by a novel query function that can efficiently locate the most informative unlabeled examples such that a large-margin classifier can learn the optimal parameter values based on them. Rigorous evaluation involving a benchmark test and an empirical test with real-world opinion data extracted from Amazon.com reveals that the proposed active learning method can train effective classifiers with far fewer

training examples and yet achieve similar performance to a typical state-of-the-art classifier without active learning.

For effectively mining competitive intelligence from customer opinions of Web 2.0, we propose a novel graphical model to extract and visualize comparative relations between products from customer reviews, with the interdependencies among relations taken into consideration, to help companies discover potential risks and further design new products and marketing strategies. Our experiments on a corpus of Amazon customer reviews show that our proposed method can extract comparative relations more accurately than the benchmark methods.

To analyze users' sentiments on SNS, a "sentiment community" is proposed as a tool. The sentiment communities with different polarities on SNS usually represent groups of users with certain preferences in common. Thus, discovering sentiment communities is very useful for enterprises to do customer segmentation and target marketing. A novel method based on an optimization technique is proposed for discovering users' sentiment communities, and comprehensive experimental evaluations are executed to demonstrate the method's effectiveness.

In summary, this dissertation covers the topics of semantically annotating opinion data, extracting comparative opinions, and analyzing users' sentiments on SNS. This work opens the door to analyzing the rich consumer-generated data of Web 2.0 and SNS for enterprises to use in business applications.

CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	iii
CONTENTS	iv
LIST OF FIGURES	viii
LIST OF TABLES	x
Chapter 1 Introduction	1
1.1 Background.....	1
1.1.1 User-Generated Opinions in Web 2.0 and Their Business Value	1
1.1.2 Existing Issues and Opportunities for Mining User Opinions.....	4
1.2 Research Problems and Objectives.....	6
1.3 Research Methodology	7
1.4 Overview of the Dissertation.....	8
Chapter 2 Literature review.....	12
2.1 Text Mining for Business Intelligence	12
2.2 Information Extraction	13
2.3 User Opinion Mining and Sentiment Analysis.....	15
2.4 Social Community Discovery.....	17
2.5 Summary.....	17
Chapter 3 Active Learning Approach for Semantically Analyzing opinions ..	19
3.1 Introduction	19
3.2 Related Work.....	25
3.2.1 Supervised Learning Approach	25

3.2.2	Existing Active Learning Approaches.....	27
3.3	The Proposed Active Learning Method.....	30
3.3.1	The Intuition Behind the Proposed Query Function.....	32
3.3.2	The Large-Margin Classifier	34
3.3.3	Candidate Parameter Value Set—Version Space.....	36
3.3.4	Eliminating Poor Parameter Values	39
3.4	The Proposed Query Functions	43
3.4.1	Minimum Sentence Margin (MSM) Query Function.....	43
3.4.2	Minimum Entity Margin (MEM) Query Function	44
3.4.3	Unexplored Dimensions (UD) Query Function.....	45
3.4.4	Combined Query Function	46
3.5	Empirical Evaluation	46
3.5.1	Evaluation Based On Benchmark Data Set	47
3.5.2	Evaluation Based on Opinion Data Extracted From Amazon	56
3.5.3	Limitations.....	61
3.6	Summary.....	62
Chapter 4	Mining Comparative Opinions for Competitive Intelligence.....	64
4.1	Introduction	64
4.2	Related Work.....	66
4.2.1	Differences with existing work.....	66
4.2.2	Graphical Model	68
4.2.3	Summary.....	70
4.3	Overall Process of Mining Comparative Opinions.....	71
4.3.1	Problem Formulation.....	71

4.3.2	General Process	72
4.4	Two-Level CRF with Unfixed Interdependencies.....	76
4.4.1	Comparative Relations	76
4.4.2	A Graphical Model for Comparative Relation Extraction.....	80
4.4.3	Illustrative Examples	84
4.5	Experimental Evaluation	85
4.5.1	Data Collection and Annotation	86
4.5.2	Evaluation Settings and Criteria	87
4.5.3	Experiment Results and Discussions.....	90
4.5.4	A Comparative Relation Map Scenario.....	96
4.6	Summary.....	98
Chapter 5	Sentiment Community: Exploring Users' Sentiments on SNS	99
5.1	Introduction	99
5.2	Difference From Social Community Discovery	101
5.3	Sentiment Community.....	102
5.3.1	Definition.....	102
5.3.2	Discovering the Sentiment Community	103
5.4	Experimental Evaluation	107
5.4.1	Synthetic Data.....	107
5.4.2	Real Data	114
5.5	Conclusion.....	118
Chapter 6	Summary.....	119
6.1	Contributions of the Dissertation.....	119
6.2	Future Directions	120

Appendix	122
References	125

LIST OF FIGURES

FIGURE 1.1 RELATIONSHIPS AMONG CHAPTERS.	11
FIGURE 3.1 THE PROCESS OF SUPERVISED LEARNING APPROACHES FOR SEMANTIC ANNOTATION.....	25
FIGURE 3.2 THE ACTIVE LEARNING APPROACH FOR SEMANTIC ANNOTATION.	31
FIGURE 3.3 CANDIDATE PARAMETER VALUE SET AND OPTIMAL PARAMETER VALUE.	33
FIGURE 3.4 INEFFECTIVE TRAINING EXAMPLES AND EFFECTIVE TRAINING EXAMPLES.	34
FIGURE 3.5 VERSION SPACE V^{struct} WITH TWO HYPERPLANES AS THE BOUNDARIES.	37
FIGURE 3.6 ω^* IN VERSION SPACE.	38
FIGURE 3.7 THE LARGEST RADIUS HYPERSPHERE WITH ω^* AS THE CENTER.	39
FIGURE 3.8 THE HYPERPLANES ARE FAR FROM ω_i^*	41
FIGURE 3.9 THE HYPERPLANES ARE CLOSE TO ω_i^*	42
FIGURE 3.10 F_1 SCORES OF THE RANDOM, LC, MSM, AND MEM QUERY FUNCTIONS.....	50
FIGURE 3.11 NO. OF TRAINING EXAMPLES USED BY THE RANDOM, LC, MSM, AND MEM QUERY FUNCTIONS.	50
FIGURE 3.12 F_1 SCORE OF THE RANDOM, LC, MEM, AND UD QUERY FUNCTIONS.....	53
FIGURE 3.13 NO. OF TRAINING EXAMPLES USED BY THE RANDOM, LC, MEM, AND UD QUERY FUNCTIONS.	53
FIGURE 3.14 F_1 SCORE OF THE COMBINED QUERY FUNCTION.	56
FIGURE 3.15 NO. OF TRAINING EXAMPLES USED BY THE COMBINED QUERY FUNCTION.....	56
FIGURE 3.16 A SCREEN SHOT OF THE CUSTOMER OPINIONS ANALYSIS SYSTEM.....	59
FIGURE 3.17 F_1 SCORES ON NOKIA N95 DATA SET.....	59
FIGURE 3.18 F_1 SCORES ON NOKIA E71 DATA SET.	60
FIGURE 3.19 F_1 SCORES ON BLACKBERRY 8320 DATA SET.	60
FIGURE 4.1 GRAPHICAL REPRESENTATION OF LINEAR-CHAIN CRF	70
FIGURE 4.2 GENERAL PROCESS OF BUILDING COMPARATIVE RELATION MAP.	72
FIGURE 4.3 THE DEPENDENCIES GRAPH FOR EXAMPLE 6.	78

FIGURE 4.4 THE DEPENDENCIES GRAPH FOR EXAMPLE 7.	78
FIGURE 4.5 THE DEPENDENCIES GRAPH FOR EXAMPLE 8.	79
FIGURE 4.6 THE DEPENDENCIES GRAPH FOR EXAMPLE 9.	79
FIGURE 4.7 GRAPHICAL REPRESENTATION OF THE TWO-LEVEL CRF WITH UNFIXED INTERDEPENDENCIES.	81
FIGURE 4.8 PERFORMANCE COMPARISON OF THREE METHODS.....	91
FIGURE 4.9 PERFORMANCE COMPARISON WITH J&L’S METHOD.	93
FIGURE 4.10 PERFORMANCE COMPARISON WHEN USING DIFFERENT LINGUISTIC FEATURES.	95
FIGURE 4.11 SCREEN SHOT OF COMPARATIVE RELATION MAP.	97
FIGURE 5.1 EXAMPLE OF SENTIMENT GRAPH	102
FIGURE 5.2 TWO-CONNECTED STRUCTURE.....	109
FIGURE 5.3 THREE-CONNECTED STRUCTURE.....	110
FIGURE 5.4 PERFORMANCE OF PROPOSED METHOD UNDER TWO-CONNECTED STRUCTURE.	112
FIGURE 5.5 INFLUENCE OF SENTIMENT COMMUNITY STRUCTURE.	114
FIGURE 5.6 COMMUNITIES IN STRUCTURE OF FLIXTER.	116
FIGURE 5.7 SENTIMENT COMMUNITIES WITH SENTIMENT CONSTRAINT $c = 10$	117

LIST OF TABLES

TABLE 3.1 SUPERVISED LEARNING METHODS FOR SEMANTIC ANNOTATION	27
TABLE 3.2 CATEGORIES OF ACTIVE LEARNING METHODS.....	30
TABLE 4.1 SUMMARY OF THE DATASET.	86
TABLE 4.2 PERFORMANCES OF THREE METHODS.....	90
TABLE 4.3 PERFORMANCES OF THE PROPOSED METHOD AND J&L'S METHOD.	93
TABLE 4.4 PERFORMANCES WHEN USING DIFFERENT LINGUISTIC FEATURES.	94
TABLE 5.1 PARAMETER VALUES.....	110
TABLE 5.2 SUMMARY DESCRIPTION OF THE MOVIE RATINGS DATASET.....	115
TABLE 0.1 F_1 SCORES OF EACH QUERY FUNCTION.....	123
TABLE 0.2 PAIRED-SAMPLES T -TEST OF THE HYPOTHESIS.....	125