

Blog Mining for Business Intelligence

Michael Chau
School of Business
The University of Hong Kong

Inaugural Workshop on Knowledge Management & E-Learning
October 28, 2009

Presentation Outline

- Background of Blog Mining
- Research Questions
- Study on iPod
- Preliminary Findings
- Future Work
- Questions and Comments

Introduction

- Blog - a web-based publication that allows users to add content, as on an Internet forum, periodically.
- Blogs have become one of the fastest growing Web-based media.
- Become increasingly popular because of the availability of easy-to-use blogging tools and free blog hosting sites
 - www.blogger.com
 - www.xanga.com
 - www.livejournal.com

Introduction

- Blogs contain a lot of information that was previously difficult to obtain (e.g., particular comments of individual customers, views on a topic).
- However, it is not easy to extract such information
 - Blog data is updated very frequently
 - Large amount of data
 - Written in casual language

Web Mining

- Many studies on data mining, text mining, and Web mining have been reported in recent years (Etzioni 1996; Hearst, 1999; Kosala & Blockeel, 2000; Chau & Chen, 2003).
 - Web crawling
 - Natural language processing/entity extraction
 - Co-occurrence analysis
 - Document classification
 - Document clustering
 - Web structure mining
 - Community analysis/network analysis
- These techniques can be applied to extract knowledge from blogs.

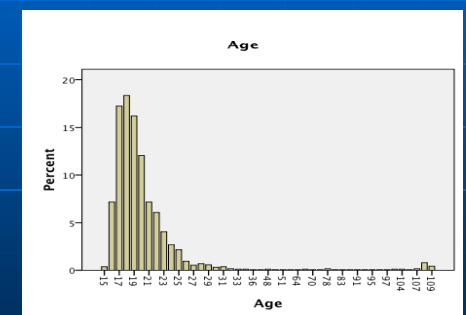
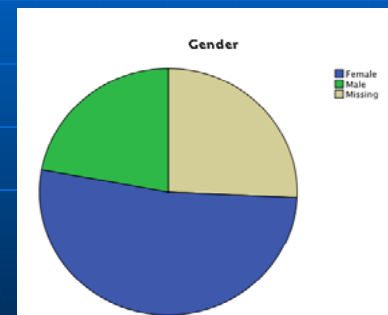
Research Questions

- How to extract important business information such as key bloggers about certain products using blog mining techniques? Are bloggers who frequently blog about a product popular in attracting other bloggers' attention?
- What can topical and stylistic analyses tell us about bloggers?
- Do bloggers form communities based on their attitudes toward a product? Do bloggers interact with others holding different or even opposite attitudes toward a product?

Study on iPod

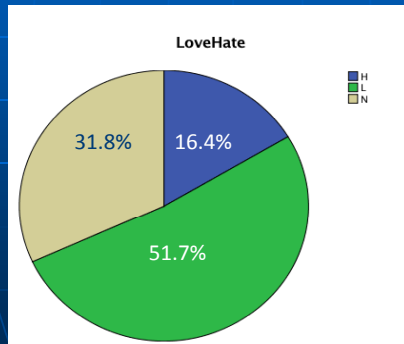
- We chose Apple's iPod music player as an example to illustrate how blog mining can be done to extract useful business information from blogs.
- We collected our blog data on a popular blog hosting site Xanga (www.xanga.com)
- All the bloggings (groups) on Xanga that contained the word "iPod" in their titles or descriptions were identified
- After removing the irrelevant and invalid ones, we have 204 groups with 3426 bloggers.
- We manually classified these groups as "positive", "negative" or "neutral"

Bloggers



Groups

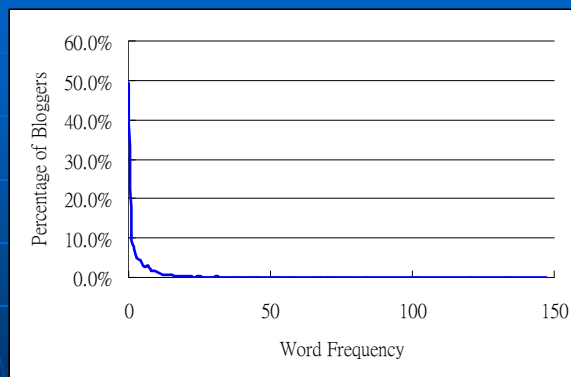
- 201 iPod groups
- Average 17.9 members (2 to 668)



Content Analysis

- We first look at the number of times that the word iPod was mentioned in each of the collected blogs
- Highest frequency: 47
- Lowest frequency: 0

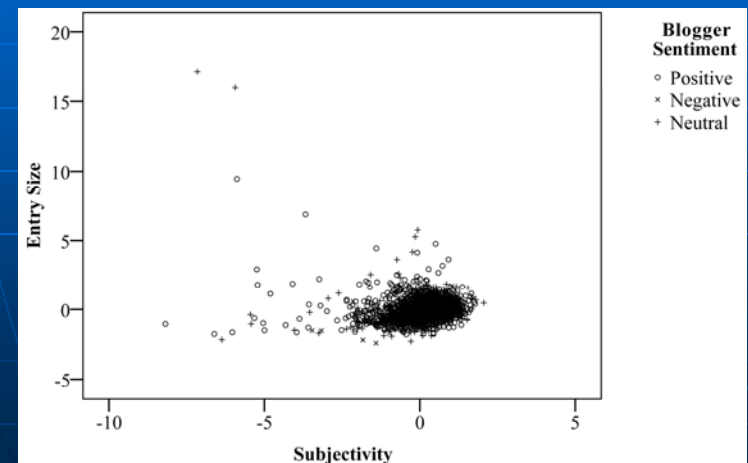
Content Analysis



- 49% do not mention iPod at all!

Content Sentiment Analysis

- Principal Component Analysis: explained 39.8% of the variance



Content Sentiment Analysis

- Wilks' Lambda value of 0.942 ($p < 0.05$) indicates that sentiment has a significant effect on stylistic features
- The neutral bloggers write longer
- The negative bloggers are more subjective

Structural Analysis

- Three types of relationship
 - Subscription
 - Commenting
 - Co-membership

Structural Analysis

- Find the key bloggers
 - Highest in-degree in subscription network: 9
 - Highest out-degree in subscription network: 9
 - Highest in-degree in commenting network: 20
 - Highest out-degree in commenting network: 16
- High correlation between subscription in-degrees and commenting in-degrees ($r = 0.43$, $p < 0.005$, $df = 475$)
- The correlation between subscription out-degrees and commenting out-degrees is nonsignificant

Structural Analysis

- Community Analysis
 - Hierarchical clustering analysis using all three types of links
 - The largest connected component (giant component) consists of 1914 bloggers connected by 2880 links.

Structural Analysis

- Minimum spanning tree of the network in which each branch of the tree represents a community



Summary of Preliminary Findings

- The popular bloggers who receive many subscriptions also tend to receive comments from many bloggers.
- The busy bloggers who subscribe to (or comment on) many other blogs may not necessarily comment on (or subscribe to) many others.
- The key bloggers with high degrees may not necessarily often blog about iPod.
- Different attitudes toward iPod do not keep bloggers from interacting with one another.

Future Work

- Explore different text classification and text clustering techniques and their applications to blog content analysis
- Combine the content analysis and structural analysis methods
- Apply to more blog sites and more products as well as other areas

Acknowledgement

- This research has been supported in part by the following grants:
 - HKU Seed Funding for Basic Research, "Social Mining for Web 2.0 Applications," May 2009 – April 2011.
 - HKU Seed Funding for Basic Research, "Analyzing Blogs for Competitive Advantages," January 2006 – December 2007.
- Collaborators:
 - Jennifer Xu, Bentley University, USA
 - Jason Li, T. Park, H. Zhao, Drexel University, USA

Further Information

- Dr. Michael Chau
 - School of Business
Faculty of Business and Economics
The University of Hong Kong
Pokfulam, Hong Kong
 - E-mail: mchau@business.hku.hk