Review Mining for Music Digital Libraries: Phase II

J. Stephen Downie
University of Illinois at Urbana-Champaign
501 E. Daniel St. Champaign, IL, 61820

jdownie@uiuc.edu

Xiao Hu
University of Illinois at Urbana-Champaign
501 E. Daniel St. Champaign, IL, 61820

xiaohu@uiuc.edu

ABSTRACT

We continue our work on the automatic mining of user-created music reviews towards the goal of connecting user opinions to music objects in Music Digital Libraries (MDL). We demonstrate an experimental system which automatically discovered the key descriptive patterns that differentiated positive from negative reviews which helps us to better understand our successful Phase I results. Comparison to an earlier study indicates an important consistency across projects that warrants further investigation.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *User issues.* H.3.1 [**Information Storage and Retrieval]:** Content Analysis and Indexing – Linguistic Processing

General Terms

Design, Experimentation, Human Factors

Keywords

music reviews, descriptive patterns, frequent pattern mining

1. INTRODUCTION

A successful Music Digital Library (MDL) system should be able to connect users' opinions on music items to the music objects contained in the system. User-generated reviews found on online store sites, review websites, and various forums including blogs, mailing lists and wikis represent a rich resource for creating new forms of descriptive and qualitative MDL metadata. Data mining and natural language processing techniques can be applied on such review data to analyze how users construct and use music information. Phase I of our work on review mining [3] was highly successful in creating a Naïve Bayesian classifier to automatically classify positive and negative music reviews (86.25%) as well as automatically identifying the music genre being discussed in each review (81.25%). In Phase II, we now dig into a finer level of analysis to investigate the reasons for the success of the Phase I tests with regard to the differentiation of positive and negative reviews. We thus focus on the *descriptive patterns* found in the positive or negative reviews. Descriptive patterns are defined as the term sets used by users to express their feelings and opinions in their reviews. Descriptive patterns are strong indicators of information selection and usage, and thus are good candidates to

connect user-generated opinions to music objects in MDL systems. To extract these descriptive patterns for analysis, we apply part-of-speech (POS) tagging and frequent pattern mining [2] techniques. The results are analyzed in relation to our Phase I findings and also compared to an earlier related study [1].

2. PHASE II EXPERIMENTS

2.1 Dataset

The dataset is the same as the Phase I dataset, namely customer reviews of music CDs published on www.epinions.com, a website devoted to online customer reviews. Each review is associated with a numerical rating expressed as a number of stars (from 1 to 5), with higher ratings indicating more positive opinions. Here, positive and negative refer to those reviews with 5 star and 1 star ratings, respectively. Each review is segmented into sentences with each sentence representing a "transaction" in frequent pattern mining [2]. Table 1 presents the dataset statistics.

Table 1. Statistics of the dataset

Reviews	Positive	Negative
Total reviews	400	400
Total sentences	63118	30053
Total words	1027713	447603
Avg. (*STD*) sentences per review	157.80 (*75.49*)	75.13 (*41.62*)
Avg. *(STD)* words per sentence	16.28 (*14.43*)	14.89 (*12.24*)

2.2 POS Tagging

As nouns are usually neutral in terms of semantic orientation (i.e. positive or negative), we identify adjectives, adverbs and verbs as candidate descriptive terms, using the POS tagging tool in the T2K toolkit [4]. After POS tagging, stop words (except for negatives) are stripped out, terms are stemmed, and a transaction file is created for generating frequent patterns using frequent pattern mining technique. In the transaction file, each line contains adjectives, adverbs and verbs from a review sentence.

2.3 Frequent Pattern Mining

Frequent pattern mining finds patterns consisting of items that frequently co-occur in individual transactions. In this case, items are descriptive terms while transactions are review sentences. Also using T2K, we conducted the experiments on the sentence level because sentences better contextualize the usage of concurrent terms than their usage at a broader document level.

3. RESULTS AND DISCUSSION

First, we examined the most frequent patterns consisting of single terms, and found the single-term patterns in positive and negative reviews to be **surprisingly** similar. Table 2 lists the top two single-term patterns and the numbers of sentences containing them. It is interesting to see that *"good"*, a semantically positive

word is on the top of the list of negative reviews! To understand these counterintuitive results, we manually analyzed the contexts of *"good"* in negative reviews, and found that there are roughly five categories of *"good"* used in a negative context:

1. Negation: e.g. "Nothing is good."
2. Song titles: e.g. "*Good Charlotte*, you make me so mad."
3. Rhetoric: e.g. " And this is a good ruiner: … "
4. Faint praise: e.g. "the only good thing… is the packaging."
5. Expressions: e.g. "You all have heard … the good old cliché."

All the above examples contain *"good"*, but none express positive opinions. This observation prompted us to consider deeper contexts. Therefore, we moved our analytic focus to double-term and triple-term frequent patterns. A pattern with multiple terms means these terms are frequently concurrent within individual sentences. Table 2 presents the top-ranked extracted patterns. It is apparent that the double-term patterns in both lists are still quite similar: 3 out of 5 most frequent patterns are the same for positive and negative reviews. However, obvious differences finally show up in the comparison of the triple-term patterns: 9 out of 10 most frequent patterns in negative reviews contain negatives (*"not", "don't"*), while only 1 pattern in the positive reviews contains a negative. Furthermore, note how 9 of the most frequent triple-term patterns in the positive reviews describe "pleasant" music experiences with 7 of these focusing on things "melodic." Thus, it is only at the triple-term level that we are able to arrive at a meaningful explication of the high classification precision in distinguishing positive and negative reviews in our Phase I work.

Table 2. Most frequent patterns ranked by frequency

LEVEL	Positive reviews	Negative reviews
Single term	not (3417 sentences)	not (1915 sentences)
	good (1621 sentences)	good (1025 sentences)
Double terms (stemmed)	not good	not good
	not realli	not bad
	realli good	not realli
	not listen	not sound
	not great	realli good
Triple terms (stemmed)	sing open melod	not realli good
	sing smooth melod	not realli listen
	sing fill melod	bad not good
	sing smooth open	bad not sound
	not realli good	pretti tight spit
	sing lead melod	bad not don't
	sound realli good	realli not don't
	sing plai melod	realli bad not
	accompani sing melod	pretti bad not
	sing soft melod	not sing sound

Table 3 is a comparison to a recent study of a user survey designed to elicit information about songs people most hate (i.e., what they think is the "worst song ever") [1]. Column 2 presents the original set of key terms (in rank order) from [1]. Our current study also discovered the frequent use of these terms in our dataset. Column 1 presents these terms ranked according to their occurrence frequencies in our negative reviews. It is not hard to see that the orders of the two lists are quite similar. This indicates that the two studies concur on the nature of negative user opinions, despite the differences in research methods applied and the populations studied. While not conclusive, this finding suggests to us that we are on the right investigative track.

Table 3. Comparison of negative descriptive term rankings

This Study	Cunningham et al '05
bad	really
annoying	bad
hate	worst
really	annoying
inane	boring
horrible	horrible
stupid	awful
worst	hate
awful	stupid
crap	crap
bore	inane

4. CONCLUSION AND FUTURE WORK
Customer reviews provide user-generated data in a great quantity and in great detail, and thus represent an excellent resource to exploit as a source of user-generated metadata. This work pioneers new approaches to connecting user-generated input to music objects in MDLs. The application of part-of-speech tagging and frequent pattern mining techniques has deepened our understanding of our Phase I review mining results. Our Phase II findings indicate that single-term and double-term patterns are not rich enough to capture the meaning/intention of user-generated reviews. Similarity of findings with earlier work suggests ongoing investigation is warranted. For the upcoming Phase III, we hope to generalize our findings by examining non-music cases to see if other DL types can benefit from our review mining techniques.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Cunningham, S.J., Downie, J.S., and Bainbridge, D. "The Pain, The Pain": Modeling music information behavior and the songs we hate. In *Proceedings of the 6th International Symposium on Music Information Retrieval.* 2005, 474-477.

[2] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD 2000.* 1-12.

[3] Hu, X., Downie J.S., West K., and Ehmann A. Mining Music Reviews: Promising Preliminary Results. In *Proceedings of the 6th International Symposium on Music Information Retrieval.* 2005, 536-539.

[4] Welge, M., et al. *Data to Knowledge (D2K) An Automated Learning Group Report*. NCSA, University of Illinois at Urbana-Champaign, 2003. (http://alg.ncsa.uiuc.edu)