RJ10421 (A0710-036) October 25, 2007
Computer Science

IBM Research Report

Artist Ranking through Analysis of Online Community Comments

Julia Grace[1], Daniel Gruhl[1], Kevin Haas[1], Meena Nagarajan[2], Christine Robson[1], Nachiketa Sahoo[3]

[1]IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

[2]Wright State University
3640 Colonel Glenn Highway
Dayton, Ohio

[3]Carnegie Mellon University
4800 Forbes Avenue
Pittsburgh, PA

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Artist Ranking Through Analysis of On-line Community Comments

"U R SO BAD!" and other compliments. . .

Julia Grace
IBM Almaden Research
650 Harry Road, San Jose CA
jhgrace@us.ibm.com

Daniel Gruhl
IBM Almaden Research
650 Harry Road, San Jose CA
dgruhl@us.ibm.com

Kevin Haas
IBM Almaden Research
650 Harry Road, San Jose CA
khaas@us.ibm.com

Meenakshi Nagarajan *
Wright State University
3640 Colonel Glenn Highway
Dayton, Ohio
nagarajan.5@wright.edu

Christine Robson
IBM Almaden Research
650 Harry Road, San Jose CA
crobson@us.ibm.com

Nachiketa Sahoo †
Carnegie Mellon University
4800 Forbes ave
Pittsburgh, PA
nsahoo@cmu.edu

ABSTRACT

We describe an approach to measure the popularity of music tracks, albums and artists by analyzing the comments of music listeners in social networking online communities such as MySpace. This measure of popularity appears to be more accurate than the traditional measure based on album sales figures, as demonstrated by our focus group study. We faced many challenges in our attempt to generate a popularity ranking from the user comments on social networking sites, e.g., broken English sentences, comment spam, etc. We discuss the steps we took to overcome these challenges and describe an end to end system for generating a new popularity measure based on online comments, and the experiments performed to evaluate its success.

Categories and Subject Descriptors

I.7.5 [**Document Capture**]; I.2.7 [**Natural-Language Processing**]

General Terms

Algorithms, Measurement

Keywords

Comment Analysis, Popularity Ranking, Social Networks

1. INTRODUCTION

Top-N lists have been a fascination of people since at least the fifth century BC when Herodotus published his "Seven Wonders of the World" [26]. From the superlatives in a high school yearbook to political polling, a community defines itself in part by ranking interests and preferences. In areas such as music, ranking also serves as a means of providing recommendations. For instance, a new artist appearing on a "Top Artists" Techno chart may be popular with fans of other Techno artists that appear on the list. Of course, this has non-trivial sales implications, so using and manipulating chart position has long been a controversial part of marketing [19].

The challenge of determining the Top-N list has lead to a host of innovative approaches to popularity rankings. Some of the hardest domains are those where tastes change quickly, such as popular music. Music often suffers from "over play" fatigue, where popular songs are played so frequently that they cease being as popular. As a result, attempts have been made to identify reasonable objective, observable proxies for interest.

The advent of records and radio as the primary means of distribution of popular music presented a reasonable "choke point" for such measurement. Since both recording music onto gramophone records and broadcasting music over radio waves required specialized machinery, it was safe to presume that any recorded music being listened to came from one of these sources. Simply counting the number of records sold and songs played acted as a reasonable proxy for what people listened to (e.g. `Billboard.com`).

While this may have been true in the 50's and 60's, record sales and radio plays have become increasingly poor predictors of what is popular in the face of rapidly increasing online music trading and downloads (both legal and illegal), device to device music sharing, on-line discussion forums targeting music (e.g., MySpace), Internet radio stations, etc. With the rise of new ways in which communities are exposed to music, comes the need to rethink how popularity is measured.

Another approach to measure popularity is conducting polls, i.e., asking people for their opinion. Challenges in polling are well known (Pliny the Younger wrote about them in 105 A.D.[2]). For this domain, one of the largest problems is that polling large samples is problematic and expensive. Rather than directly asking people what they think, however, we can make use of the wealth of information in online

*Research completed at IBM Almaden Research Center
†Research completed at IBM Almaden Research Center

WWW2008, April 21–25, 2008, Beijing, China.

communities. Popularity can now be determined by monitoring on-line public discussions, examining the volume and content of messages left for artists on their pages by fans and looking at what music is being requested, traded and sold in digital environments.

This work measures music popularity by mining music-enthusiasts' comments on artist pages on MySpace – a popular online music community [1]. To comprehend the meaning of the comments, we first had to overcome challenges due to broken sentences, unconventional writing, slang and spam. We transliterate, de-spam, and mine comments for music related information, then aggregate the results to create a Top-N list of popular artists. To test the effectiveness of our ranking system, we compare our top artists ranking to the Top-N list from `Billboard.com` [2]. Overall, many of the same artists appear on both lists, however, the ranking differs. In an informal poll of 74 students, the list generated by our system was found to better reflect the tastes of the students by more than 2 to 1.

Thus, our system creates a closer connection between the popular lists and the listeners by examining what the listeners are saying today. Since data can be gathered from online sources in near real time, we eliminate not only the traditional wait time for poll results, but also the dependency on other aggregated data such as sales figures. The data can also be filtered by demographics to create customized rankings. The net effect is more accurate and specific data, predicting today's top and upcoming artists, rather than reporting on last week's sales and airplay.

The remainder of the paper is organized as follows: In Section 2 we discuss the background to our research. Section 3 introduces the corpora we selected. The system which mines online communities, parses the content, and analyzes popularity is described in Section 4. Our experiment to validate the system and the results of our tests are described in Section 5 and discussed in Section 6. We look to future work in Section 7 and conclude in Section 8.

2. MOTIVATION AND BACKGROUND

Rank ordering and text mining are the two fields most relevant to our work. Text mining, also termed text data mining, generally refers to the process of adding meta-data (often with positional information) to text. In our work, comments on artist pages are processed to spot music related comments, remove spam and identify positive sentiments. We build upon two well known text mining tasks: opinion mining and spam identification, assisted by named entity spotting techniques. Comment annotation volumes are aggregated to order popular artists. Here, we present a brief literature review of these areas.

2.1 Rank Ordering

The role of social choice and mass popular culture in affecting Top-N lists and its influence on production and consumerism has been extensively studied in politics, arts and economics [1][18]. While these studies have been motivated by the communication industry's attempt at grasping popularity trends and consumer behavior, results have also been used to understand how a popular culture's operation shapes 'hit lists'[20].

Social science teaches us that fundamental to popularity is the presence of a vocal, popular minority compelled to share their opinions with a larger audience... in other words, when the cool kids find a new fad, the silent majority will follow. While the social structure that enables opinion sharing ranges from star ratings, reviews, blogs, and chats, trends have largely been driven by what is considered popular within a peer group. Harold Lasswell's comment in 1948 on communication models[15] – "Who (says) What (to) Whom (in) What Channel (with) What Effect"– summarizes the role that a community, their opinions and the popular culture materials they use, can have on trends and popularity lists. Communications within a society were shown to often be an inextricable part of outcome of functions like voting preferences.

Our work builds upon this literature. Since popularity often develops in and spreads through social communication channels, we hypothesize that one can arrive at a list of what is popular music today by measuring positive activity in these channels. By observing trends over time and patterns that stand out among the communications in these channels, we might also be able to forecast what is going to be popular tomorrow.

2.2 Unstructured Opinion Mining

Our approach for quantifying crowd preferences is directly related to work in opinion mining (OM) from public boards such as blogs, reviews, forums etc. Challenges in OM from casual text thick with comparative and sarcastic opinions have been addressed by a decade's worth of work in this area [6]. Our mining of teenager sentiments about artists or their music varies a bit from past work in OM because of the nature of the corpus and end goals of popularity ranking. Specifically, we limit OM to coarse assignments of positive and negative comments on an artist's page. In this respect, the goal of our work is similar to [11], [24] [7] and [13].

Among others, Esuli and Sebastian's [7] approach to determine the semantic orientation of opinion terms through gloss analysis is closest in spirit to what we do. By using a vectorial representation of online dictionary definitions of a set of seed positive and negative adjectives, they train a binary text classifier on the seed definitions and then apply it to the test set. In this work, we use natural language processing of comments to identify sentiments. A youth slang dictionary [3] along with the statistical significance of the sentiment interpretation is used to transliterate unconventional expressions and identify opinion polarities. While the task of identifying semantic orientation of opinions is not new, our contribution in mining transliterations to identify opinion polarities in casual and broken English is novel.

2.3 Comment Spam Identification

Online public sources tend to be infested with bot-generated spam content [16]. There are two broad classes of work in countering spam in such forums: preventing and identifying

[1] www.myspace.com

[2] From Wikipedia: The Billboard Hot 100 is the United States music industry standard singles popularity chart issued weekly. Chart rankings are based on airplay and sales; the tracking-week for sales begins on Monday and ends on Sunday; while the airplay tracking-week runs from Wednesday to Tuesday. A new chart is compiled and officially released to the public by Billboard on Thursday. Each chart is dated with the "week-ending" date of the following Saturday.[25]

[3] www.urbandictionary.com

spam. Our work falls in the second class of content-based identification of spam but differs from past work in terms of the end goal and because of the nature of our dataset. Typical content-based techniques work by testing content on patterns or regular expressions [17] that are indicative of spam or by learning Bayesian models over spam and non-spam content [3]. Recent investigations on removing spam in blogs use similar statistical techniques with good results [23].

These techniques were largely ineffective on our corpus because comments are rather short (1 or 2 sentences), share similar buzz words with non-spam content, are poorly formed and contain frequent variations of word/slang usage. Our approach of filtering spam is an aggregate function that uses a finite set of mined spam patterns from the corpus and other non-spam content such as artist names, and sentiments that are spotted in a comment.

3. CORPUS DETAILS

Music popularity or opinions on music are the subject of many heated discussions in online communities. Our choice of online data corpora is motivated by two main factors: a target audience of teenagers, and a desire for music-centric content.

We choose teenagers because of their considerable effect on the overall popularity of music. A trio of industry reports around the effect of communities on music consumption identifies the growing population of online teenagers as the biggest music influencers; 53% of which is also spreading word about trends and acting as primary decision-makers for music sales [12]. We exploit this majority's appreciation of music in online music communities to complement traditional metrics for ranking popular music. While our extended work in this area has used six online communities, we limited ourselves to MySpace for generating data in this paper to avoid the issues around multi-site ranking fusion.

3.1 MySpace

MySpace is a popular social networking site that has a section dedicated to music artists and fans. Both major, independent, and unsigned music artists have taken advantage of this popular and free-for-all social networking site to manage online relationships with fans. Members of this community, over half of whom are our target demographic, learn about latest artists, albums, and events, and express their opinions on the comments section of an artist's page. We crawl and mine comments from such artist pages to determine popularity numbers. In addition, we also gather user demographic information like comment poster's age, gender and location to derive demographic trends. Table 1 shows the crawled structured and unstructured data.

Type	Crawled Data
Structured	Artist Name, Albums, Tracks, Genre, Country, User, Age, Sex, Location
Unstructured	Posted comments mentioning Artist, Album, Tracks, Sentiments and Spam

Table 1: Description of Crawled Data

3.2 Corpus Characteristics

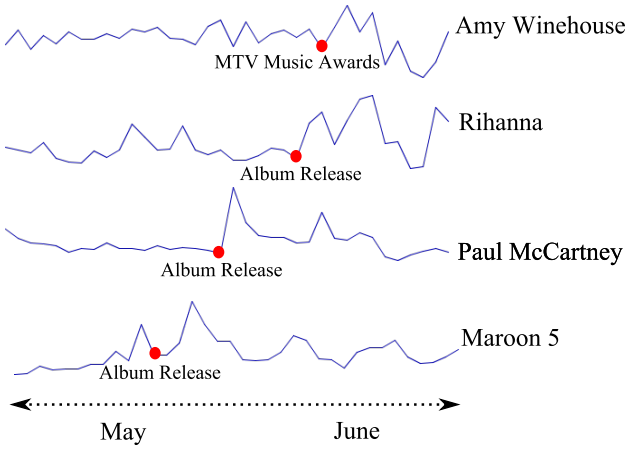

Figure 1: Spikes in comment volumes and rises in popularity occur after newsworthy events.

One of the salient features of such corpora is the availability of near realtime data. We show that it is possible to assess popularity trends, correlate chatter with external events (like artists winning awards) and identify the beginning and persistence of trends to enable marketing focus on early-adopter segments without the lag from sales data (which may take weeks to collect and aggregate). Over a period of 26 weeks (Jan through Jun 2007) 788,384 unique comments were observed for the top 100 artists in this time frame. The volume of comments highlights the importance of a scalable crawling and mining system. Figure 1 illustrates spikes in comment volumes on artist pages that coincide with real events. The ability to gauge buzz and popularity the day after an artist releases an album or appears on television is invaluable to record labels as they attempt to sway the buying decisions and loyalties of fans.

We conducted some experiments on a random sample of 600,000 of these comments and observed the following characteristics of the unstructured component of the corpus:

- More than 60% of terms used to indicate sentiment contained slang that required special treatment.
- Less than 4% of the comments expressed negative sentiments about artists; comparative or sarcastic comments were rare occurrences. Detection of sentiments proved to be an important step in the process of spam detection.
- Almost 40% of comments on an artist's page were self-promotional or advertisement related spam. Spam comments were often less than 300 words long, and appreciative comments less than 100 words long.
- The natural language construction of over 75% of the non-spam comments was non-conventional, often resulting in inaccurate or failed linguistic parses.

Our annotator system, which is responsible for gleaning structure out of this unstructured content, effectively deals with these limitations by using a combination of statistical and linguistic techniques.

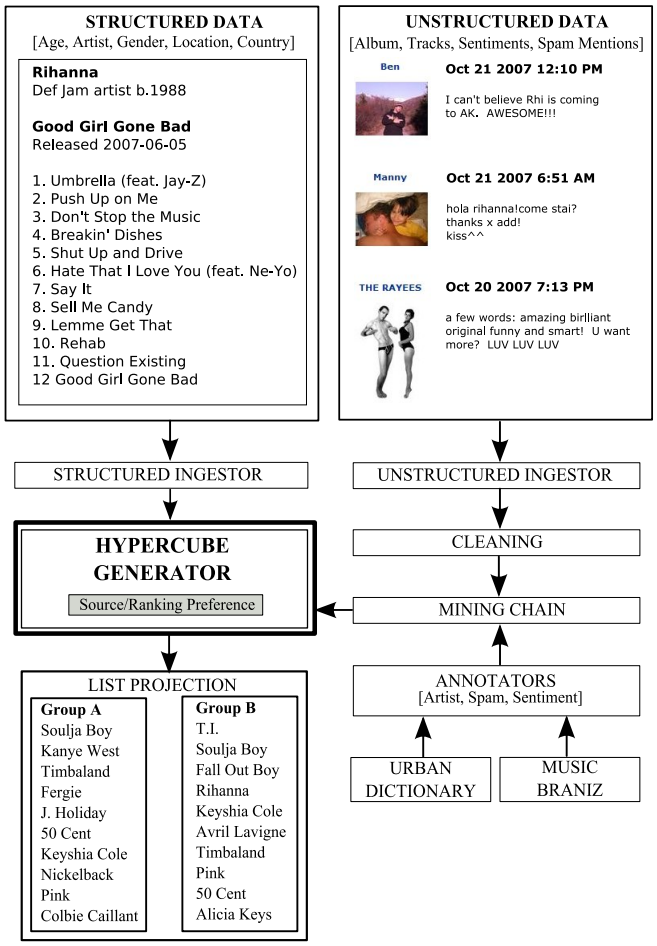

Figure 2: System Component Architecture.

4. SYSTEM DETAILS

The overall architecture of our system consists of four components (see Figure 2):

1. **Crawling**: Fetching the data from the source site, transforming the pages and comments into common formats, and ingesting the data into the database.

2. **Annotating**: Ingested comments are passed through a UIMA[8] chain of annotators to adjudicate if the comment is related to the artist or music, if it has any associated positive or negative sentiments, and if it is a spam comment.

3. **Hypercube construction**: The data is rolled up by a variety of dimensions (e.g., age, gender, locale) and a summary hypercube of comment and sentiment volume is constructed.

4. **Projection to a list**: Ultimately we want a Top-N list, so we need to project this hypercube to a single value which is used to order the list of artists, tracks, albums, etc.

We will explore each of these steps in turn, with particular attention to approach and lessons learned.

4.1 Crawling and Ingesting

The crawling and ingesting component gathers data from a potentially diverse set of sources and maps the data to a normalized format for further processing. It must do so in a way that is scalable to millions of comments and extensible to changes in the data sources and annotation schemes.

Given constrained data acquisition bandwidth, we need to prioritze how to examine the sub-parts of the site and assign a frequency with which to revisit each artist page. As an example, we might seed our set of artists to consider by looking at a "top artists" list from social networking sites such as MySpace, or from published top charts such as Billboard's "Top Singles" charts.

Given this seed list, we can then identify the artist pages for these candidates and begin to pull semantic information on fan preferences (i.e., comments) from these pages. For the sake of politeness we need to wait a few seconds between fetches to reduce the load on any given server and achieve sustainable crawling, but given a multiplicity of sites this does not impact overall crawl rate. Comment data, as described in Section 3, consists of a structured component such as artist name, a time stamp, the user demographics of the poster, plus an unstructured component (i.e., the comment text).

The list of artists can be quite extensive. There are nearly 50,000 artists in an initial set. With a politeness wait between requests, this means that one could only check a few hundred artists an hour and exhaustive rescans could take days. In the fast changing environment of a social network music community, rapidly emerging artists could be missed for extended periods. With a goal of obtaining a near real time pulse of the community the desiderata is a Top-N list once every 4 hours. Without allocating more bandwidth, this means new data representing only a couple thousand artists is possible. Fortunately, not all artists are commented on with the same frequency – thus crawling with a prioritization scheme is possible.

We use two data gathering schedules that arbitrarily split the available crawl bandwidth:

1. **Priority crawl**: A process that scans roughly one thousand artist pages in 4 hour cycles. These are the artists have the highest variance/uncertainty in their comment incidence rate[4].

2. **Exhaustive crawl**: A process that scans all the artists at the rate of about one thousand per hour. In each scan we collect all the comments generated since the last scan and generate new estimates for the comment rate and its uncertainty.

These techniques allow us to bring in the maximally useful comment stream, which is transformed via a site-specific remapping function into a normalized data format. This one step is by far the most brittle of the entire system, as it

[4]Since only a small subset of artists is examined in our Priority crawl we create a simple estimator of the number of comments an artist would have at any time. Over several scans we can then create an estimate of error. We can then look at how long it has been since we last obtained firm data on a source to generate expected error, and then sort the priority crawl list to maximally reduce uncertainty. This equation can also be back solved to define requisite crawl rate for a given error bound.

needs to deal with the site format and access pattern changes of the crawled sites.

Once the data is normalized it is stored in a relational database (DB2) using a data model that is easily extensible for future additions of data sources. Each comment is uniquely identified by a combination of user, data source, artist and time-stamp (best estimate or exact) values. We track artists across data sources, but we do not at this time link posters across data sources. Comment annotations are stored in an extensible schema of two tables: one storing the list of annotations and the other storing a comment, annotation pair per record. When a new annotation is generated for the existing comments only the new information can be added to the set of tables.

4.2 Annotators

The annotation component automatically processes comments to compute the total number of positive comments for each artist. We use the scalable, UIMA[8] based framework to host a short chain of three annotators:

1. Artist and Music Annotator: Spotting artist, album, track, and other music related (e.g. labels, tours, shows, concerts) mentions.

2. Sentiment Annotator: Spotting and transliterating sentiments in comments.

3. Spam Annotator: Identifying comments that are spam or do not directly contribute to artist/music popularity figures (e.g. comments about an artist's DUI charge).

Each annotator is an "analysis component" that processes one comment at a time to find the entity of interest independently. However, the output of each of the three annotators is made available to the other annotators to allow observations to be made incrementally. Such composibility helps deal with short comments or those containing spam and non-spam content in the same "sentence". Annotation results are then aggregated over time periods to characterize the volume of positive, negative and spam comments. Additionally, counts of tracks and album mentions on an artist's page are also tallied.

All of these annotators are driven off of basic entity spotting. We look to simple arbitrary window-based entity spotting techniques backed by domain dictionaries which have been used in the past with fairly reasonable success [9], [22]. While publicly available artist and track dictionaries provide the necessary dictionary support, the possible variations of the entity (misspelling, nick names, abbreviations, etc.) occurring in this often teen-authored corpus approaches the infinite.

Considering other techniques, there is good work on using natural language (NL) parses to spot nouns (for example) and/or a statistical strength to indicate an entity's importance in the corpus [10]. Unfortunately, the "broken English" and possible variations of entities in this corpus make simple NLP problematic.

As a result, we have gone with a hybrid of these two methods: a dictionary and window-based spotter complemented with a part-of-speech tag analysis and the corpus-wide distribution strength of an entity. The natural language parsing of sentences is obtained using the Stanford NL Parser[14] and the distribution strength of an entity in the corpus is found based on an implementation of the Bayardo pruning method[21].

To evaluate our annotators, we processed a corpus of 600,000 comments gathered over a period of 26 weeks. All precision and recall figures presented in this section are calculated over a random sample of 300 comments from 9 artists (the restricted set due to the need to hand tag the entire test corpus for recall numbers). Tunable cut-off thresholds for annotators were determined based on experiment.

4.2.1 Music related / Artist-Track Annotator

The goal of this annotator is to spot artist and track mentions in a comment. Empirical evaluation suggests that the number of occurrences of comments on an artist's page that mention some other artist or tracks of other artists is insignificant (and thus ignored at this point). This annotator is backed by an artist's tracks and albums list from MusicBrainz and a short dictionary of music related words like tour, concert, album, etc.

Spotting an artist/track in a comment proceeds as follows:

1. Window of words + Jaccard similarity of a dictionary entity and entity spotted in text. Variable window lengths are obtained from number of words in a dictionary entity that we are trying to spot.

2. A shallow NL parse of the comment to verify the spotted entity's part of speech tag; considered favorable if the tag is a noun or a noun phrase. This verification is done only for artists and for tracks that are one-word long, since parses failed to identify longer tracks as noun phrases due to often odd sentence constructions.

3. Look up the spotted entity's corpus-wide statistics.

4. If the combined score of the three steps is greater than a tunable threshold (e.g., 0.9 for artists and 0.8 for tracks), record the annotation with the dictionary value of the spotted entity. For example, 'Aiimmy' in the comment is annotated as 'Amy Winehouse' to facilitate aggregation of number of artist mentions.

Table 2 shows the results of the annotator on the base algorithm and excluding the NL parse technique. We contend that a combination of NLP and statistical techniques yields good results in such casual broken English corpora.

Annotator Type	Precision	Recall
Artist	1.0	0.86
Track	0.67	1.0
Artist excluding NLP component	1.0	0.64

Table 2: Artist-Track Annotator

Analysis of results indicated two main reasons for lowered precision of the track annotator. First, false positives such as one word track names such as 'Smile','Dare' etc. were used in free-speech in combination with poorly structured sentences. Secondly, common heuristics like capitalized first letter or tagged as a noun/noun phrase often failed due to misspellings and non-standard writing conventions.

We observed that the recall suffers due to arbitrary variations of names (e.g. 'Rihanna' is sometimes referred to in the corpus as 'Riri'), odd sentence constructions and incomplete artist dictionaries (often missing names of members of a band).

4.2.2 Sentiment Annotator

This annotator seeks to identify the sentiment expressed in a comment. One of unique challenges we faced compared to previous efforts in this area was the very large number of ways posters express sentiment. In order to identify sentiments and their polarities, this annotator translates a variety of slang expressions to a finite set of known bad and good sentiments using a popular slang dictionary – `UrbanDictionary.com` (UD). First, a seed of 60 positive and 45 negative sentiments is created manually to assist in this transliteration. UD provides a set of related tags and user-defined and voted definitions for a slang term. Since definitions are not necessarily accurate and automating the process of reducing them to a single sentiment is harder, we use the related tags. Next, we compute the corpus-wide statistic of a related tag and pick the one that occurs most frequently to be a transliteration for the slang term. To illustrate, we transliterate the slang-sentiment 'tight' to 'awesome' because of the following related tags in UD and occurrence strengths in the comments corpus - awesome 456, sweet 136, hot 429, sick 23, dope 182.... Since 'awesome' appears in our seed positive sentiment dictionary, the polarity of the slang 'tight' is recorded as positive. We create a dictionary of such transliterations ('wicked' transliterates to 'cool') and query UD for cases when the slang does not appear in the dictionary.

The process of spotting a sentiment proceeds as follows:

1. A shallow NL parse of a sentence to identify adjectives or verbs to suggest the presence of a sentiment.
2. Look for the spotted sentiment in the seed dictionaries or obtain the transliteration from the transliteration dictionary or UD. Identify and record the slang's polarity.
3. Increase the confidence in the spotted sentiment if there is also an artist/music related entity spotted by the first annotator.
4. If the confidence is greater than a tunable threshold, record the sentiment's polarity as a Boolean annotation.

Table 3 shows the accuracy of the annotator and illustrates the importance of using transliterations in such corpora.

Annotator Type	Precision	Recall
Positive Sentiment	0.81	0.9
Negative Sentiment	0.5	1.0
PS excluding transliterations	0.84	0.67

Table 3: Transliteration accuracy impact

Analysis of results indicated that the syntax and semantics of sentiment expression is hard to determine. Words incorrectly identified as sentiment bearing resulted in inaccurate transliterations which contributed to low precision, especially in the case of the Negative Sentiment annotator. Dependency parses of comments were expensive and minimally effective due to poor sentence constructions. Low recall often attributed to slangs not defined in UD. The slight increase in precision of the Postive Sentiment annotator when excluding the transliterations dictionary indicates the need for more selective transliterations in light of poorly structured sentences.

SPAM: 80% have 0 sentiments

> CHECK US OUT!!! ADD US!!!
> PLZ ADD ME!
> IF YOU LIKE THESE GUYS ADD US!!!

NONSPAM: 50% have at least 3 sentiments

> Your music is really bangin!
> You're a genius! Keep droppin bombs!
>
> u doin it up 4 real. i really love the album. keep doin wat u do best. u r so bad!
>
> hey just hittin you up showin love to one of chi-town's own. MADD LOVE.

Figure 3: Examples of sentiment in spam and non-spam comments.

4.2.3 Spam Annotator

Like many public spaces today this corpus suffers from a fair amount of spam – comments off topic of the message that often are a kind of advertising. While certain characteristics of spam made it harder to classify comments using traditional machine learning or pattern based techniques, these characteristics were quite useful in generating effective identification heuristics.

1. The majority of spam comments were related to the domain, had the same buzz words as many non-spam comments and were often less than 300 words long.
2. Like any auto-generated content, there were several patterns in the corpus indicative of spam. This annotator is aided by a finite seed of 45 such patterns found in the corpus using the Bayardo method.
3. Comments often had spam and appreciative content in the same sentence which meant that the annotator had to be aware of the previous annotation results.
4. Empirical observations of the corpus suggest that the presence of sentiment is pivotal in distinguishing spam content. Figure 3 illustrates the difference in distribution of sentiments in spam and non-spam content.

The first step of our algorithm simply spots possible spam phrases and their variations in text using the mined spam patterns; window-based and string similarity techniques. Classifying a comment as spam or non-spam is done using a set of rules over the results of all the three annotators. An example of such a rule would be that if a spam phrase, artist and music entities, and a positive sentiment were spotted; the comment was probably not spam. Table 4 shows the accuracy of the spam and non-spam annotators.

Analysis indicates that lowered precision or recall in the spam annotator was a direct consequence of deficiencies in the first two annotator results. For example, cases when

Annotator Type	Precision	Recall
Spam	0.76	0.8
Non-Spam	0.83	0.88

Table 4: Spam annotator performance

the comment did not have a spam pattern, and the first annotator spotted incorrect tracks, the spam annotator interpreted the comment to be related to music and classified it as non-spam. Other cases included more clever promotional comments that included the actual artist tracks, genuine sentiments and very limited spam content. (e.g. 'like umbrella ull love this song...'). As is evident, the amount of information available at hand in addition to grammatically poor sentences necessitates more sophisticated techniques for spam identification.

Given the amount of effort involved the obvious question arises – why filter spam? For our end goal of using comment counts to lead to positions on the list, filtering spam is key. This is corroborated by the volume of spam and non-spam content observed over a period of 26 weeks for 8 artists; see Table 5. The chart indicates that some artists had at least half as many spam as non-spam comments on their page. This level of noise would significantly impact the ordering of artists if it were not accounted for.

Gorillaz	54%	Placebo	39%
Coldplay	42%	Amy Winehouse	38%
Lily Allen	40%	Lady Sovereign	37%
Keane	40%	Joss Stone	36%

Table 5: Percentage of total comments that are spam for several popular artists.

4.3 Generation of the hypercube

We use a data hypercube (also known as an OLAP cube[5]) stored in a DB2 database to explore the relative importance of various dimensions to the popularity of musical topics. The dimensions of the cube are generated in two ways: from the structured data in the posting (e.g., age, gender, location of the user commenting, timestamp of the comment, artist), and from the measurements generated by the above annotator methods. This annotates each comment with a series of tags from unstructured and structured data. The resulting tuple is then placed into a star schema in which the primary measure is a relevance with regards to musical topics. This is equivalent of defining a function.

$$M : (Age, Gender, Location, Time, Artist, ...) \rightarrow M \quad (1)$$

In our case, we have stored the aggregation of occurrences of non-spam comment at the intersecting dimension values of the hypercube. Storing the data this way makes it easy to examine rankings over various time intervals, weight various dimensions differently, etc. Once (and if) a total ordering approach is fixed this intermediate data staging step might be eliminated.

4.4 Projecting to a list

Ultimately we are seeking to generate a "one dimensional" ordered list from the various contributing dimensions of the cube. In general we project to a one dimensional ranking which is then used to sort the artists, tracks, etc. We can aggregate and analyze the hypercube using a variety of multi-dimensional data operations on it to derive what are essentially custom popular lists for particular musical topics. In addition to the traditional billboard "Top Artist" lists, we can slice and project (marginalize dimensions) the cube for lists such as "What is hot in New York City for 19 year old males?" and "Who are the most popular artists from San Francisco?" They translate to following mathematical operations:

$$L_1(X) : \sum_{T,...} M(A = 19, G = M, L = \text{“}NewYorkCity\text{''}, T, X, ...) \quad (2)$$

$$L_2(X) : \sum_{T,A,G,...} M(A, G, L = \text{“}SanFrancisco\text{''}, X, ...) \quad (3)$$

where,

X = Name of the artist
T = Timestamp
A = Age of the commenter
G = Gender
L = Location

Note that for the remainder of this paper we aggregate tracks and albums to artists as we wanted as many comments as possible for our experiments. Clearly track based projections are equally possible and the rule for our ongoing work.

This tremendous flexibility is one advantage of the cube approach.

F

5. EXPERIMENTS

To test the effectiveness of our popularity ranking system we conducted a series of experiments. We prepared a new top-N list of popular music to contrast with the most recent Billboard list. To validate the accuracy of our lists, we then conducted a study.

5.1 Generating our Top-N list

We started with the top-50 artists in Billboard's singles chart during the week of September 22nd through 28th, 2007. If an artist had multiple singles in the chart and appeared multiple times, we only kept the highest ranked single to ensure a unique list of artists. MySpace pages of the 45 unique artists were crawled, and all comments posted in the corresponding week were collected.

We loaded the comments into DB2 as described in Section 4.1. The crawled comments were passed through the three annotators to remove spam and identify sentiments. The tables below show statistics on the crawling and annotation processes.

Number of unique artists	45
Total number of comments collected	50489
Total number of unique posters	33414

Table 6: Crawl Data

38%	of total comments were spam
61%	of total comments had positive sentiments
4%	of total comments had negative sentiments
35%	of total comments had no identifiable sentiments

Table 7: Annotation Statistics

As described in Section 4.3, the structured metadata (artist name, timestamp, etc.) and annotation results (spam/non-spam, sentiment, etc.) were loaded in the hypercube.

The data represented by each cell of the cube is the number of comments for a given artist. The dimensionality of the cube is dependent on what variables we are examining in our experiments. Timestamp, age and gender of the poster, geography, and other factors can all be dimensions in hypercube, in addition to the measures derived from the annotators (spam, non-spam, number of positive sentiments, etc.).

For the purposes of creating a top-N list, all dimensions except for artist name are collapsed. The cube is then sliced along the spam axis (to project only non-spam comments) and the comment counts are projected onto artist name axis. Since the percentage of negative comments was very small (4%), the top-N list was prepared by sorting artists on the number of non-spam comments they had received independent of the sentiment scoring.

In Table 9 we show the top 10 most popular Billboard artists and the list generated by our analysis of MySpace for the week of the survey.

5.2 System Details

All experiments were run on Xen virtual machines hosted on an IBM 3650 with quad-core processor running at 2.66GHz. The VMs run Redhat Enterprise Linux WS release 4 update 5. Each is allocated 1-2 GB of physical RAM. Data storage is done on a 4 drive SATA Raid5 array on which each VM has an image file served through QEMU-DM. Data Management was done via DB2 v9.1 EE for Linux.

5.3 The word on the street

Having fetched more that 50,000 comments, gone to great lengths to remove the spam and parse the informal English found within, tallied and scored and ultimately derived an alternative Top-N list for popular music, the obvious question raised is – does it work? Do people actually post on-line about music they prefer? Could a list generated from casual comments on a social networking site be a more accurate representation than that offered up by the record industry itself? Fully answering this question would (and will) require numerous studies beyond the scope of this paper, but we were able to perform a casual preference poll of 74 people in the target demographic.

At the conclusion of the data sampling week, we conducted a survey among students of an after-school program (Group 1), Carnegie Mellon (Group 2), and Wright State (Group 3). Of the three different groups, Group 1 comprised of respondents between ages 8 and 15; while Group 2 and 3 comprised primarily of college students in the 17-22 age group. Table 8 shows statistics pertaining to the three survey groups.

The survey was conducted as follows: the randomly chosen 74 respondents were asked to study the two lists shown in Table 9. One was generated by Billboard and the other through the crawl of MySpace. They were then asked the following question: 'Which list more accurately reflects the artists that were more popular last week'? Their response along with their age, gender and the reason for preferring a list was recorded.

Groups and Age Range	No. of male respondents	No. of female respondents
Group 1 (8-15)	8	9
Group 2 (17-22)	21	26
Group 3 (17-22)	7	3

Table 8: Survey Group Statistics

The sources used to prepare the lists were not shown to the respondents, so they would not be influenced by the popularity of MySpace or Billboard. In addition, we periodically switched the lists while conducting the study to avoid any bias based on which list was presented first.

Billboard.com	MySpace Analysis
Soulja Boy	T.I.
Kanye West	Soulja Boy
Timbaland	Fall Out Boy
Fergie	Rihanna
J. Holiday	Keyshia Cole
50 Cent	Avril Lavigne
Keyshia Cole	Timbaland
Nickelback	Pink
Pink	50 Cent
Colbie Caillat	Alicia Keys

Table 9: Billboard's Top Artists vs our generated list

6. RESULTS

The raw results of our study immediately suggest the validity of our hypothesis, as can be seen in Table 10. The MySpace data generated list is prefered over 2 to 1 to the Billboard list by our 74 test subjects, and the preference is consistantly in favor of our list across all three survey groups.

	Group 1	Group 2	Group 3
MySpace-Generated List	15	30	6
Billboard List	2	17	4

Table 10: Experiment Results: number of people who prefered each list

A more fine grained statistical analysis of the data only improves upon the initial suggestion of the data. 68.9% of all subjects prefered our list. Calculating the standard error for these 74 responses, we have:

$$\frac{s}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})} = 0.054$$

computed using $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})}$, the estimated

standard deviation, and $\bar{x}$, the mean of the data points $\{x_1, x_2, x_i, ..x_n\}$.

This estimated standard deviation of the sample mean provides the result that 68.9± 5.4% of subjects prefer our list to the Billboard list. Looking specifically at Group 1, the youngest survey group whose ages range from 8-15, we can see that our list is even more successful. Even with a smaller sample group (resulting in a higher standard error), 88.2± 8.1% of subjects prefer our list to Billboard. This striking result shows a 6 to 1 preference for our list from younger listeners.

We can further calculate a confidence level for our data using the common statistical method, the t-distribution. This method is generally accepted to be usable for sample sizes of more than 30 observations without the need to establish that the data is normally distributed. [4]

We employ the standard t-test to determine a critical value, denoted $T(\frac{\alpha}{2}, n-1)$ for n samples and a confidence interval of $k = 1 - 2\alpha$. The confidence interval for the confidence level k is than given by the critical value times the standard error:

$$T(\frac{\alpha}{2}, n-1)\frac{s}{\sqrt{n}}$$

Solving this for the confidence level which shows a preference of MySpace list versus the Billboard list (i.e., a more then 50% preference for the MySpace list) gives:

$$\bar{x} - T(\frac{\alpha}{2}, n-1)\frac{s}{\sqrt{n}} \geq 0.5$$

$$0.69 - T(\frac{\alpha}{2}, 73)0.055 \geq 0.5$$

which solves using t-test tables to $\alpha = 0.001$.

This gives a 99.9% confidence interval that a randomly polled group of similar individuals will show an overall preference for the MySpace data generated list over the Billboard list. Thus, we can say with a high degree of confidence that on-line sources are a *better* indicator than the traditional record charts for people in our sample group.

We can also tentatively conclude that our list is prefered equally by men and women. Groups 2 and 3 had equal preference for the myspace list (approximately 64% and 60%, respectively), however group 2 was mostly female (55%), whereas group 3 was mostly male (70%).

Another interesting observation: after concluding the survey, we asked some of the subjects which they thought was the most popular set (as opposed to the one they preferred). That is, the correlation between perceived popularity and preference. 83% of subjects believed that their prefered list was also the most popular list, similarly distributed across those who preferred the MySpace generated list and those that preferred Billboard.

We conclude that new opportunities for self expression on the web provide a *more* accurate place to gather data on what people are really interested in than traditional methods. The even stronger results from the younger audience suggests that this trend is, if anything, accelerating.

7. FUTURE WORK

7.1 Further Experimentation

In validating our lists, we focused on college students, but surveying additional demographics such as high school students (who comprise an age range that is representative of the majority of MySpace users) and new college graduates might yield additional insights as to the impact of our lists. This audience might be useful in determining exactly how close our lists are to the industry Top-N, and determining trends in how good a predictor our list is to the industry Top-N list a week or two later.

7.2 Future Research Directions

One approach we considered but postponed due to time constraints was deeper analysis on the social network graph to determine which individuals are most likely to influence their online communities. While we have just shown that statistical sampling is no longer necessary, we would like to explore whether targeted "advertising" of new opinions (music tracks, new artists) affect the likelihood that the community(s) will adopt the opinions of these individuals.

There are many other topics where we could employ our methodology to gauge popularity and sentiment. Sports teams, movies, and video games are just a few – but in order to accurately assess popularity and sentiment, an active corpus with many user-generated comments must be available. Online forums are a starting point, but they are dependent on the online "footprint" that these topics have in the online forums, blogs, and larger social networking sites. For example, trying to track popular topics for the San Francisco Symphony would mean we would have to crawl many smaller data sources where the postings may contain many topics unrelated directly to the symphony itself. Using message boards with lesser information about participants (such as Usenet) would not give us the ability to easily determine age, gender and geographic preference correlations.

7.3 Broken English

Broken English is not limited to social networking sites. Fragmented grammar appears in call center transcripts, chat logs from instant messaging clients, email messages, text messages, voice-to-text transcripts with poor precision, etc. Our analysis framework can be used to create metadata, rewrite acronyms, etc. in all of these domains. As electronic communication methods becomes less formal, our annotators become increasingly valuable. We plan to continue research with other media such as these to extend our work into new domains.

7.4 Enterprise Applications

We chose the music domain and built content annotators to create our own Top-N list due to the impact that music has on popular culture. This approach is by no way limited to teenage opinion surveys as the same text mining techniques can also be applied to call center transcripts, instant messaging chat logs, and emails.

One commercial example of how this can be applied to an enterprise is by performing similar analysis on electronic communication (emails and corporate IMs). Many applications exist that monitor email and instant messaging behavior between employee accounts, but these tend to monitor message flows and not the actual content. By adding in the content analysis, we can use similar annotators to monitor employee sentiment, behavior trends, topics of interest, and compliance with legislative regulations.

Conventional wisdom around market intelligence suggests statistical surveys of sample populations as the preferred method to determine prevailing opinions. While statistically valid, these surveys require active participation and hopes that the sample population will not bias the survey. We believe that a new methodology (like ours) for market intelligence that gathers opinions from a large population (ideally, an entire population) would more accurately determine prevailing opinions. Previously, this was considered infeasible due to the difficulty of reaching 100% of the population. With the penetration of online social networking sites (e.g., MySpace, Orkut) and acceptance of blogging by GenX/iGen populations, including topic-specific blogs such as Slashdot and Blogger, data mining the online opinions of large portions of this population can be quickly implemented.

8. CONCLUSION

Online communities are a virtual gold mine of GenX/iGen music opinions. Regardless of a musician's genre, label, or age, one is hard pressed to find a band without a MySpace profile, and most popular bands have a fairly active fan community. Even more traditional bands such as the Beatles and the Rolling Stones have active presences. Providing fans with the ability to deliver personal messages and feel as though they have spoken directly to the band has proved to be very appealing, and, as we have shown, very valuable for us to gauge popularity and buzz within these communities.

9. REFERENCES

[1] T. W. Adorno. A social critique of radio music. *Kenyon Review*, pages 208–217, 1945.

[2] M. Balinski and R. Laraki. A theory of measuring, electing, and ranking. *PNAS*, 104(21):8720–8725, May 2007.

[3] J. Blosser and D. Josephsen. Awarded best paper! - scalable centralized bayesian spam mitigation with bogofilter. In *LISA '04: Proceedings of the 18th USENIX conference on System administration*, pages 1–20, Berkeley, CA, USA, 2004. USENIX Association.

[4] A. C. Cameron. Statistical inference for univariate data.

[5] E. Codd, S. Codd, and C. Salley. *Providing OLAP (on-line Analytical Processing) to User-analysts: An IT Mandate*. Codd & Date, Inc., 1993.

[6] A. Esuli. Survey of techniques for opinion mining. `http://medialab.di.unipi.it/web/Language+Intelligence/OpinionMining06-0%6.pdf`.

[7] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624, New York, NY, USA, 2005. ACM Press.

[8] D. Ferrucci and A. Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, 2004.

[9] D. Freitag. Information extraction from html: application of a general machine learning approach. In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 517–523, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[10] D. Freitag and N. Kushmerick. Boosted wrapper induction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 577–583. AAAI Press / The MIT Press, 2000.

[11] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[12] M. R. Inc. Teen market profile. `http://www.magazine.org/content/files/teenprofile04.pdf`.

[13] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using wordnet to measure semantic orientation of adjectives, 2004.

[14] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. In *NIPS*, pages 3–10, 2002.

[15] H. D. Lasswell. Listening to popular music. *The Communication of Ideas*, 1948.

[16] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, 2006.

[17] J. Mason. Filtering spam with spamassassin. In *Proceedings of HEANet Annual Conference*, 2002.

[18] D. Mayzlin and J. A. Chevalier. The effect of word of mouth on sales: Online book reviews. *Yale School of Management Working Papers*, 2003.

[19] M. McIntyre. Hubbard hot-author status called illusion. `http://www.scientology-lies.com/press/san-diego-union/1990-04-15/hubbar%d-hot-author-status-illusion.html`.

[20] D. Riesman. Listening to popular music. *American Quarterly*, 2(4):359–371, 1950.

[21] J. Roberto J. Bayardo. Efficiently mining long patterns from databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, New York, NY, USA, 1998. ACM Press.

[22] S. Soderland. Learning to extract text-based information from the world wide web. *Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, 1997.

[23] A. Thomason. Blog spam: A review. In *Fourth Conference on Email and Anti-Spam CEAS 2007*, 2007.

[24] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, 2003.

[25] Wikipedia. Billboard charts. `http://en.wikipedia.org/wiki/Billboard_charts#The_Billboard_Hot_100`.

[26] Wikipedia. Wonders of the world. `http://en.wikipedia.org/wiki/Seven_Wonders_of_the_World`.