

# Multimodal Social Intelligence in a Real-Time Dashboard System

Daniel Gruhl · Meena Nagarajan · Jan Pieper · Christine Robson · Amit Sheth

Received: August 10th 2009 / Accepted: April 22nd 2010

**Abstract** Social Networks provide one of the most rapidly evolving data sets in existence today. Traditional Business Intelligence applications struggle to take advantage of such data sets in a timely manner.

The BBC SoundIndex, developed by the authors and others, enabled real-time analytics of music popularity using data from a variety of Social Networks. We present this system as a grounding example of how to overcome the challenges of working with this data from social networks. We discuss a variety of technologies to implement near real-time data analytics to transform Social Intelligence into Business Intelligence and evaluate their effectiveness in the music domain.

The SoundIndex project helped to highlight a number of key research areas, including named entity recognition and sentiment analysis in Informal English. It also drew attention to the importance of metadata aggregation in multimodal environments. We explored challenges such as drawing data from a wide set of sources spanning a myriad of modalities, developing adjudication techniques to harmonize inputs, and performing deep analytics on extremely challenging Informal English snippets.

Ultimately, we seek to provide guidance on developing applications in a variety of domains that allow an analyst to rapidly grasp the evolution in the social landscape, and show how to validate such a system for a real-world application.

---

Daniel Gruhl, Jan Pieper, and Christine Robson  
IBM Almaden Research Center  
650 Harry Road, San Jose, CA  
E-mail: {dgruhl, jhpieper, crobson}@us.ibm.com

Meena Nagarajan and Amit Sheth  
Ohio Center of Excellence on Knowledge-enabled Computing  
(Kno.e.sis), 3640 Colonel Glenn Highway, Dayton, OH  
E-mail: {meena, amit}@knoesis.org

**Keywords** BBC SoundIndex · Social Intelligence · Informal Text Analysis · Information Mashups · Semantic Annotation · Semantic Domain Models · Slang Sentiment Identification · Spam Filtering · Voting Theory

## 1 Introduction

There are over 100 million active users of MySpace<sup>1</sup>. Facebook has over 400 million users<sup>2</sup> and Twitter is growing at over 1300% a year<sup>3</sup>. Social networks produce more data every day than most companies see in a year. It is not surprising that with roughly the same number of users as the population of the United States, a wide variety of topics get discussed. With the ability to rapidly disseminate information, new topics can generate tremendous buzz in a matter of hours.

A substantial research and engineering effort is required to get a handle on this very large and rapidly evolving data set. This challenge is worth tackling as it enables us to tap into this wisdom of the crowds in near real-time. In this article, we explore some of the challenges and opportunities in building such “Social Intelligence” (SI) systems based on our experience with both a real world SI system, and subsequent research to explore some of the limitations it revealed.

We ground these observations in a near real-time SI system developed by the authors and others for the BBC “SoundIndex”. This system drew over 40 million datum a day from a variety of sources in order to construct a top 1000 music chart that continuously captures the buzz around popular music (see Figure 1).

---

<sup>1</sup> <http://www.myspace.com/pressroom?url=/fact+sheet/>

<sup>2</sup> <http://www.facebook.com/press/info.php?statistics>

<sup>3</sup> [http://blog.nielsen.com/nielsenwire/online\\_mobile/twitters-tweet-smell-of-success/](http://blog.nielsen.com/nielsenwire/online_mobile/twitters-tweet-smell-of-success/)

## 1.1 Business Intelligence

Business Intelligence (BI) systems have become a popular way to use large amounts of traditionally low value data that, when analyzed together, can provide an overview of the state of an organization. BI systems transform and analyze this data to derive aggregate statistics and identify trends and relationships. An important aspect of such systems is the focus on presenting the data in a way that the end user can manipulate and obtain actionable business insights by basing intuition on real-world observations.

As noted in Surowiecki (2004), under certain conditions, information gathered from crowds can exceed the accuracy and speed of that derived from experts. We seek to use this crowd-sourced information by taking advantage of the diversity and decentralized aspects of the crowd and analyzing it with the discipline of BI systems. Such a system differs from traditional polling methods in that while “passive” (specific responses cannot be elicited), it is considerably less expensive, able to cover a very broad audience, and has less risk of several kinds of data contamination (e.g., biasing the outcome with loaded questions). The lower cost of analyzing this data makes it easier for an analyst to “explore” the social landscape and develop their own intuition (Cody et al (2002); Koutsoukis et al (1999)).

## 1.2 From Social Intelligence into Business Intelligence

The task of taking Social Intelligence and transforming it into Business Intelligence is not easy. The data sets are multimodal, socially-affected content and tend to be very large. For example, the current number of “tweets” on Twitter per second<sup>4</sup> is rapidly closing in on the peak number of transactions per second for major credit cards<sup>5</sup>.

All of these present challenges in processing the user-generated content, mining the prevailing opinions, aggregating between very different modalities and presenting the findings in an actionable way. For the establishing problem, the BBC SoundIndex, this was made even more challenging because the expected audience for the analysis tool was teenagers who may not have extensive training in business modeling tools.

We have found that the approaches taken for the BBC SoundIndex apply well to developing other Social Intelligence systems. We have also seen that an SI system can be leveraged for many different applications, such as sense-making, anthropological studies, and studying the evolution of communities, topics and

even color preferences (Locke, 2004). In all of these cases, the ability to quickly and easily examine the selections, comments and preferences of millions of users provides a powerful environment for experiments.

## 1.3 Contributions

This article summarizes our experiences with building a deployed Social Intelligence application — the BBC SoundIndex — to provide near real-time information about the popularity of various artists, tracks and genres of music. We pay special attention to describe what it takes to build such an application that wishes to exploit the wisdom of the crowds.

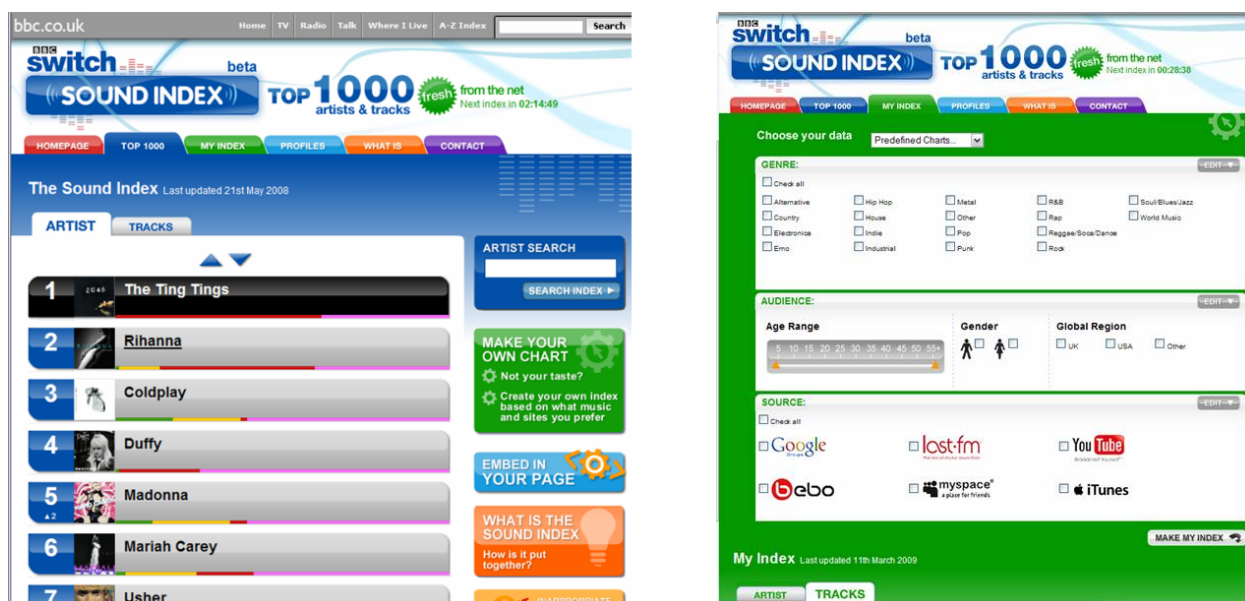
We found challenges and interesting research questions in many different areas during this project. The system had to be able to process Informal English text, and structured and semi-structured data from a variety of modalities, all of which reflect some aspect of crowd-sourced intelligence.

We discuss high throughput analysis of user-generated content from social networks, metadata extraction from such sources, tagging of entities, sentiment identification and spam elimination. Lastly, and in some ways most importantly, the resultant information had to be presented in a way that facilitates the extraction of actionable business insight. Specifically, we cover the following contributions:

- Data acquisition from sites that are prominent in a social networking setting. We present a system to handle the acquisition and processing of data, even when the data arrives sporadically.
- Semantic annotation of artist and track mentions. We leverage semantic domain models, specifically the MusicBrainz RDF, and the statistical and natural language properties of entity mentions to spot and disambiguate entity mentions in informal text.
- Elimination of spam and off-topic discussions from artist pages. We present a system that accounts for spam including a special type of non-traditional spam content — discussions surrounding an artist but unrelated to their music.
- Aggregation of sentiment expressions surrounding an artist and their work. We detail a method that identifies traditional and unconventional sentiment expressions (slang), detects opinion polarities and aggregates positive vs. negative expressions for an artist.
- Normalization of high level extracted concepts into explorable data structures such as hypercubes. This transformation allows traditional data exploration tools to be employed on social data.

<sup>4</sup> <http://en.wikipedia.org/wiki/Twitter>

<sup>5</sup> 2007 Federal Reserve Payments Study, [www.frbservices.org/files/communications/pdf/research/2007\\_payments\\_study.pdf](http://www.frbservices.org/files/communications/pdf/research/2007_payments_study.pdf)



**Fig. 1** The BBC SoundIndex — an example of a Social Intelligence application which ingested tens of millions of datum a day to create near real-time charts of artist and track popularity.

- Application of voting theory to create “information mashups” by providing non-linear information aggregation algorithms. This is very important when the data driving an SI system comes from sources with very different modalities (e.g., comments, sales, counts of video views, track listens, etc.).
- Exploration of two high value presentation approaches to present Social Intelligence information for rapid understanding. Ultimately an SI system is of value only when it can help an analyst make a better, more timely decision.

## 2 BBC SoundIndex

Our work began with a real-world application. As a project for the BBC, we developed a replacement for the traditional music charts which lists the top songs and artists each week. The goal was to use data from social networking sites to generate a near real-time application listing top artists and songs, including the analytical capabilities to explore different genres and population demographics. The application is called BBC SoundIndex, and we ran a live beta on the BBC website from January through September of 2008. Figure 1 shows the top artist chart on the left and the analytical controls on the right.

### 2.1 Vision

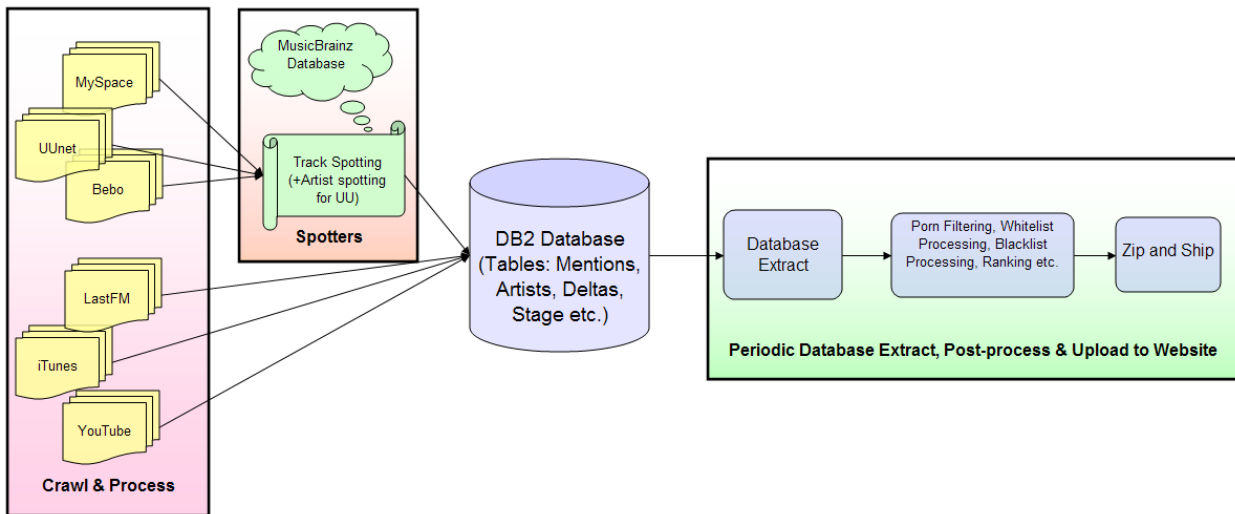
The BBC had noticed an increasing drift in the artists and songs reported in popular music charts and what their subject matter experts (i.e., DJs) thought. This is

not surprising as the methodology for generating these charts dates back over half a century. In the 1950’s and 1960’s, record sales and radio plays were good predictors of music popularity. Since both recording music onto phonograph 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 counting is the goal, in reality these numbers are derived from polling a relatively small number of record stores and radio stations. Challenges in polling are well known (Pliny the Younger wrote about them in 105 A.D. (Balinski and Laraki, 2007b)). This raises the obvious concern — are the sample record stores really representative of the way most people get their music? Is the radio still the most popular medium for music? In 2007 less than half of all teenagers purchased a CD<sup>6</sup> and with the advent of portable MP3 players, fewer people are listening to the radio.

With the rise of new ways in which communities are exposed to music, the BBC saw a need to rethink how popularity is measured. Could the wealth of information in online communities be leveraged by monitoring online 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,

<sup>6</sup> NPD Group: Consumers Acquired More Music in 2007, But Spent Less



**Fig. 2** SoundIndex architecture. Data sources are ingested and if necessary transformed into structured data using the MusicBrainz RDF and data miners. The resulting structured data is stored in a database and periodically extracted to update the front end.

traded and sold in digital environments? Furthermore, could the notion of “one chart for everyone” be replaced with a notion that different groups might wish to generate their own charts reflecting the popularity of people like themselves (as the Guardian<sup>7</sup> put it – “What do forty something female electronica fans in the US rate?”). Providing the data in a way that users can explore the data of interest to them was a critical goal of our project.

## 2.2 Basic Design

The basic design of the SoundIndex, which we will describe in more detail in the remainder of the paper, is shown in Figure 2. We looked at several multimodal sources of social data, including “discussion sites” such as MySpace, Bebo and the comments on YouTube, sources of “plays” such as YouTube videos and LastFM tracks, and purely structured data, such as the number of sales from iTunes. We then clean this data, eliminate spam and perform analytics to spot songs, artists and sentiment expressions. The cleaned and annotated data is combined, adjudicating for different sources, and uploaded to a front end for use in analysis.

## 2.3 Challenges in Continuous Operation

Real world events often impact systems attempting to use social data. Every time the news reports that MySpace has been a subject of a DDOS attack, one can be

<sup>7</sup> <http://www.guardian.co.uk/music/2008/apr/18/popandrock.netmusic>

assured that those trying to pull their data are suffering as well. When sites report wildly inconsistent numbers for views or listens (e.g., total number of “all time” listens today is less than the total number yesterday), they also causes problems for the kind of aggregation we wish to do.

Combine these unforeseeable events associated with the data scale, multiple daily updates to the front end, and running a distributed system with components on different continents, and it becomes clear why this kind of system is challenging to build and operate.

## 2.4 Testing and Validation

Is the social network you are looking at the right one for the questions you are asking? Is the data timely and relevant? Is there enough signal? All of these questions need to be asked about any data source, and even more so of social networks given the varied user-interests they cater to. When developing an SI application, such as the SoundIndex, it is critical to validate the final output. We used a combination of point polling with groups in the target audience along with ongoing verification with subject matter experts to help identify problems with the system and raise confidence that the resulting data (our top 1000 list) is credible.

## 2.5 Lessons Learned

Deploying this system brought to light many areas which warranted further research and development. They include the gathering of data from low reliability sources,

the need for enhanced natural language processing capabilities for application to the kinds of Informal English often found in these social networks, well developed adjudication technologies to support adjudication of multimodal sources, and dashboards to support the visualization and understanding of these very high-dimensional and complex data sets. We will explore each of these in turn in the remainder of the paper.

### 3 System and Control

Our Social Intelligence system can best be envisioned as a five step pipeline (see Figure 3). Like many BI systems there is no a static corpus that is analyzed once and a result drawn. Our system is rather a dynamically evolving flow of information where the output is more likely to be monitored in updating lists and dashboards than corresponding to a fixed report. This introduces substantial engineering challenges that must be overcome – it is a large step from processing data once to a system which runs continuously.

**Data Acquisition:** Fetching the data from the source sites using web crawlers and site-specific APIs, transforming structured data, web pages and user comments into common formats, and ingesting the data into the database. This data may be either unstructured (in the form of comments) or already structured (for example record sales).

**Pre-Processing:** As a first stage of processing unstructured comment data, individual posts are broken out, and repeated posts and some types of obvious spam are removed. In general this stage is fairly heuristic and as such is one of the stages that changes the most as the system is applied to new domains (e.g., electronic health records).

**Language:** Ingested comments are passed through a UIMA (Ferrucci and Lally, 2004) chain of annotators to identify important entities<sup>8</sup>. In the case of music this is the artist and track mentions and if a comment has any associated positive or negative sentiments. At this time we also check that the comment is on topic and unlikely to be spam.

<sup>8</sup> The UIMA (Unstructured Information Management Architecture) is a system to allow multiple annotators to provide annotations on the same underlying subject of analysis placing their results in a common data structure. These annotations can be as simple as a notation that a span of characters represents a certain part of speech or entity, but can also be as complex as providing links into larger ontologies, provide some notion of the valance of a sentiment being expressed, or even provide cross references to other annotations provided by other annotators.

**Facets:** We employ aspects of traditional business intelligence solutions in our SI system, one of which is the concept of “dimensions” or “facets” - a way of thinking about features and how they may change. In music, dimensions may have to do with features of the user making a comment (e.g., gender, location), the post itself (e.g., data source, time of post), and derived content of the post (e.g., the artist/album being discussed, associated sentiment expressions) etc. This data can then be envisioned as a hypercube where the dimensions of the cube are the features that were extracted. Once transformed this way the cube can be projected to ignore dimensions we do not care about in a given analysis (e.g., perhaps gender) and grouped by those we do (e.g., perhaps age or number of album sales).

**Adjudication:** While manipulating the hypercube can reveal interesting features of the underlying data, for this application (and many other Social Intelligence applications) we want a Top-*N* list. Consequently, we need to project this hypercube to a single value which can be used to rank the list of artists, tracks, albums, etc. Multiple projections are fed into a voting scheme for final rank aggregation.

#### 3.1 Control

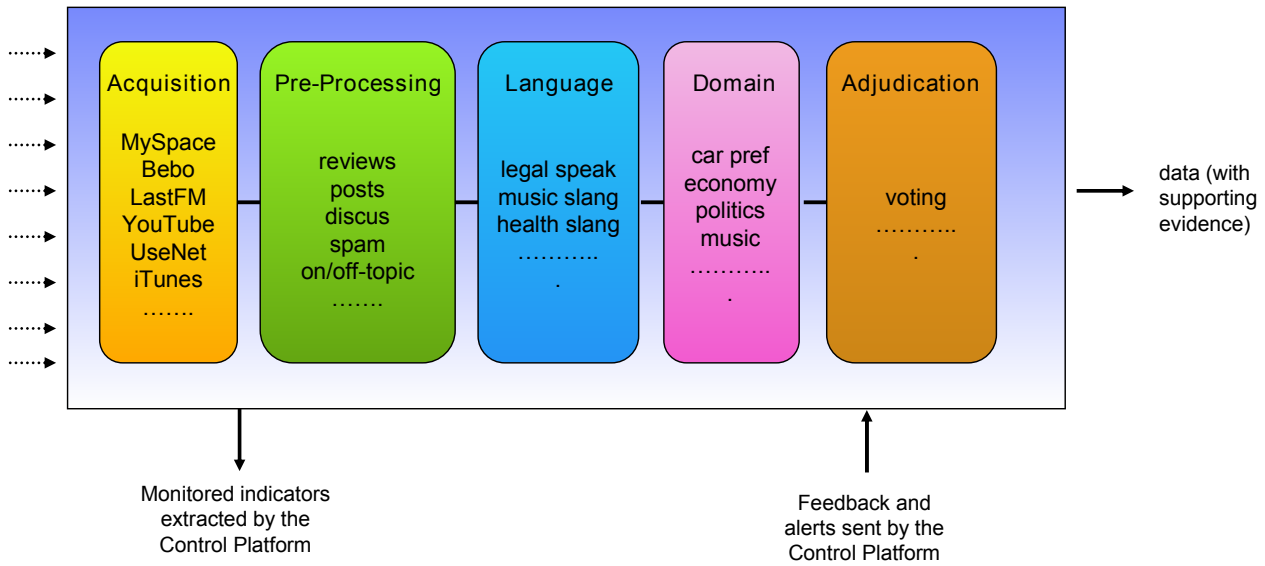
As alluded earlier, keeping an information workflow of many complex parts running 24 hours a day seven days a week is a difficult task. This is further complicated by the challenges of fetching and processing tens of millions of datum a day given the variety of sources, the quantity of data involved, and the fact that none of these source sites are under our control.

The control system runs over the entire pipeline to ensure that data is flowing in the expected types and quantities and to detect problems. This control layer takes what corrective action it can, and alerts an operator when a problem cannot be fixed without human intervention (Bhagwan et al (2009); Alba et al (2008b)).

#### 3.2 Human Intervention

It is worth noting that systems of this type cannot compare to the “gold standard” of human adjudicated results. Any system where human subject matter experts are available to curate results will outperform a system that seeks to operate largely autonomously.

However, in a very large data volume system, most posts will need to be processed fully automatically. Identifying when and how a subject matter expert can most effectively intervene is a difficult challenge. In many



**Fig. 3** System Component Architecture of the Data Acquisition and Processing System. In addition to the linear flow a control layer monitors the data at each stage and performs “sanity checks” that allow errors (either in the source, or in the system itself) to be caught as quickly as possible. It is worth noting that this is a general purpose architecture that the authors have employed in many domains. For each domain different “plug ins” are appropriate - for example, legal speak is a less useful plugin for music, but very helpful when evaluating SEC filings.

cases, for a system of this type, human expert time is best employed on developing and scoring training sets that deal with evolving or changing conditions (e.g., particular challenges around emerging artists, particular changes in how language is being used) and less on fixing misclassifications on day-to-day posts.

#### 4 Data Sources

For the Top 1000 list we developed for the BBC SoundIndex we relied exclusively on online communities. We used sites where users comment on artists and songs, listen to music, view videos, enjoy remakes, covers and parodies, purchase downloads, and so forth. Most of these sites are considered Social Networking Systems (SNSs), and all of them include some aspects of social networking. Each site is aimed at enabling music fans to become more engaged in the communities surrounding their favorite artists. Looking at multiple sources helps to remove bias that may exist within one source and increases the authority of the combined output.

##### 4.1 Source Characteristics

We found it helpful in our review of these sources to consider characteristics of each source along a few axes. These descriptions can be used to describe any data source for an SI system, and help inform how best to use the data source.

**Intentional vs. Unintentional:** Intentional sources require the user to take active steps, such as selecting a

song or video. Unintentional sources eliminate or override the selection process, often using an automated system, such as one which automatically plays music based on previously stated preferences.

**Creative vs. Passive:** Some sources require creative actions by the user, such as authoring a comment on a bulletin board. Passive sources simply observe user behavior, e.g., which songs or videos the user selects. While creative sources usually provide a lower volume of content, we usually wish to “weight” them higher because the effort involved implies deeper interest. Table 1 identifies all sources we explored within the modalities of intention and passiveness.

**Structured vs. Unstructured Data:** Within each data source there will typically be elements of both structured and unstructured data that an SI system can make use. User demographic information, sales ranks, song listens and video views are relatively clean sources of structured data. In contrast, user comments are unstructured data that require entity spotting, spam detection, sentiment mining etc. to extract the expressed opinions. For examples of the crawled structured and unstructured data in our system, see Table 2.

	Intentional	Unintentional
Creative	Comments: MySpace, YouTube, Bebo, Twitter, Facebook, Blogs	–
Passive	Active Views: YouTube; Sales Data	Passive Listens: LastFM

**Table 1** Sources by Modality

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 2** Examples of structured and unstructured data.

## 4.2 Corpus Details

Music popularity and opinions on music are the subject of heated discussions in online communities. For our initial investigations, our choice of online data corpora was 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 are also spreading word about trends and acting as primary decision-makers for music sales (Mediamark, 2004). We exploit this majority’s appreciation of music in online music communities to complement traditional metrics for ranking popular music.

While this provides us with a list of highly targeted sites of quality data, data acquisition has been an evolving process. The BBC SoundIndex was launched with data from MySpace, LastFM, Bebo, YouTube, Usenet, and Sales Rank from Amazon.com and Apple iTunes. Since then we have explored additional data sources such as Facebook, Twitter, and blogs.

### 4.2.1 MySpace

MySpace is a popular social networking site that has a section dedicated to music artists and fans. Major, independent, and unsigned music artists have all 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. Our system crawls and analyzes comments from these artist pages to determine the popularity of specific songs and the artist overall. In addition, it also gathers user demographic information such as the age, gender and location of the comment poster to derive demographic trends.

### 4.2.2 LastFM

LastFM is a website that aggregates music listening statistics from a diverse group of users who listen to

music from their computers. It uses this data to create customized play lists for its users. A similar service is available from Pandora<sup>9</sup>, but for the SoundIndex project we focused on data from LastFM. The most valuable information is a list of songs, albums, and artists that were most listened to in the last one week and last six months. This provides insight into what people like based on their listening behavior, which can differ from what people discuss. LastFM also provides demographic information similar to that of MySpace, and can be considered a structured data source, because the data is collected from LastFM playlists of each listener- a relatively noise-free dataset.

### 4.2.3 Bebo

Bebo is a social networking site similar to MySpace, and is especially popular in Europe. It provides official artist pages and a top 100 list of the most popular artists and songs. Our system uses these top 100 listings to discover new artists and crawls all discovered artist pages for user comments. Bebo also provides user demographic data, but only age and gender are most useful as many members put clever answers into the location attribute, such as “where do you live? — In my head”.

### 4.2.4 YouTube

YouTube is a video sharing website that hosts many music videos. Our crawler had to be adapted to query the site for specific artists and then retrieve view counts and member comments for individual videos. We found that YouTube was very popular at the time and provided roughly an order of magnitude more data than any other data source. The unevenness between sources poses a serious question about how to create a joined ranked list, which we explore in more detail in Section 9.

### 4.2.5 Usenet

The Usenet is one of the earliest Internet discussion systems still in existence today. It is a precursor to HTML based discussion forums and online bulletin boards. The Usenet is a rich data source with a worldwide user base and discussions on a broad range of topics. One particular advantage of Usenet data is that it is plain text with well established protocols and tools for retrieval of new content. On the other hand, forums are often quite broad and cover many artists, music styles and even other topics mixed in. This confounding increases the noise for the subsequent tasks of artist and song spotting. It is also worth noting that the demographics

<sup>9</sup> www.pandora.com

for this source appear to be somewhat older than for many of our other sources.

#### 4.2.6 Sales Rank

Another important data source is sales data or sales rank as provided by Amazon.com and Apple iTunes. The original Billboard charts are based on music sales and airplay data. One problem with sales data is that it cannot capture the buzz about an upcoming release, but it still provides verification of the buzz once the song or album is commercially available. An advantage of sales data is that it is fairly structured and clean.

### 4.3 Additional Data Sources

Our initial work with the BBC focused on the six data sources described above. One reason to limit the number of sources was the need to enter into licensing and service agreements with each site. Another reason is that the Internet constantly creates new and exciting ways to interact and communicate, some of which were not available or popular when we first started this project. We have since explored the following additional sources, as they contain highly relevant data.

#### 4.3.1 Facebook

Facebook is a popular social networking site that provides public pages for artists and fans. While user demographic data is often not public due to privacy controls on Facebook, the site has some unique features. Facebook's interaction model is similar to that of blogs — the official artist posts a message and followers can comment on the specific topic. Facebook also offers a light-weight feedback mechanism where users can “like” something with a single mouse-click. However, one drawback of this interaction model is that the frequency of new official announcements may influence the overall activity on an artist's page.

#### 4.3.2 Twitter

The recent popularity of microblogging sites such as Twitter has caught our attention as well. We adapted our system to retrieve data about artists using their query API and found that it provides a stream of comments similar to MySpace. One notable difference is the timeliness of twitter data. Users will often “tweet” as an event is happening. For example, there was a large spike in comments during the release of Madonna's new song “celebration” on July 31, 2009. Such immediate feedback makes microblogging sites an interesting and

potentially valuable data source. However, one serious drawback is that it is currently difficult to detect on-topic comments amongst the thousands of off-topic tweets every day. This problem is compounded by the short nature of posts (140 character maximum), which provides little context for the subsequent data processing.

#### 4.3.3 Blogs

With the explosive growth of the “blogosphere” as a means for self-expression and information dissemination, there is great potential for mining this data set for Social Intelligence. Several online services index and enable searching over blog postings, including Technorati<sup>10</sup> and Blogpulse<sup>11</sup>. As of June 2008, Technorati has an index of 112.8 million blogs, which are searchable via an API.

### 4.4 Source Exploration

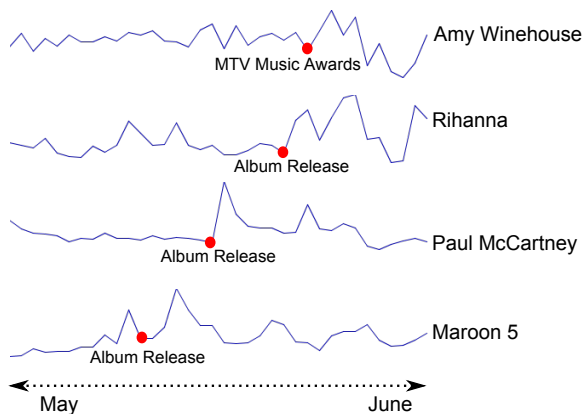
We conducted a proof of concept study in 2007 to show that these data sources do provide near real-time data. The goal of this exploration was to show that it is possible to assess popularity trends, correlate chatter with external events (like artists winning awards) and identify the beginning and persistence of trends to enable marketing focus on early-adopter segments without the lag from sales data. Over a period of 26 weeks (January through June 2007) 788,384 unique comments were collected from the top 100 artists on MySpace. Figure 4 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 buying decisions and fan loyalties.

We conducted some initial 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.

<sup>10</sup> <http://technorati.com/>

<sup>11</sup> <http://www.blogpulse.com/>



**Fig. 4** Spikes in comment volumes and rises in popularity occur after newsworthy events in 2007.

- 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 complemented by semantic domain models.

## 5 Data Acquisition

The data acquisition component gathers data from a diverse set of sources. It must do so in a way that is scalable to millions of datum and extensible to changes in the data sources and to additional data sources.

### 5.1 Prioritization of the Data Acquisition Process

Given constrained network bandwidth, we need to prioritize how to examine the different sites and assign a frequency with which to revisit each artist page. As an example, we seeded our set of artists by looking at the “top artists” lists from social networking sites such as MySpace, and from published top charts such as Billboard’s “Top Singles” charts.

Using this seed list, we can then identify the artist pages of interest and begin retrieving information on fan preferences (i.e., comments) from these pages. The access policies of these sites usually require a wait of a few seconds between fetches to reduce the load on their servers. This limits the amount of data that can be retrieved from a single site, but does not impact overall crawl rate when fetching data from multiple sites.

The list of artists can be quite extensive. Our initial set consisted of nearly 50,000 artists. With the required

wait between requests, our system could only check a few thousand artists per hour and exhaustive rescans of all comments for a popular artist could take days.

We were concerned that rapidly emerging artists could be missed for extended periods of time in the fast changing environment of a social network music community. One insight is that not all artists receive user feedback with the same frequency. We therefore implemented a prioritization scheme for the crawlers.

Others have explored perfect crawl scheduling schemes, but often these require developing a schedule based on high quality estimates on rate of source change. Since we do not have such information we have used two data gathering schedules that arbitrarily split the available crawl bandwidth, balancing the need to cover popular artists while also discovering emerging trends:

1. **Exhaustive crawl:** A process that scans all the artists at the rate of about one thousand per hour. Each scan collects all the comments generated since the last scan and generates new estimates for the comment rate and its uncertainty.
2. **Priority crawl:** An additional process that scans a subset of the pages (one thousand artist pages) in 6 hour cycles. These are the artists who have the highest variance/uncertainty in their comment incidence rate<sup>12</sup>.

Taken together these two crawls do a good job of keeping the information we know about fresh while also discovering new sources as they emerge.

### 5.2 Ensuring a Continuous Feed

There were a variety of interesting challenges that emerged around the crawling tasks in various domains. Some of them were operational ones that we expected; such as the need to limit how often a source is contacted. Others revolved around source availability, for example maintenance cycles, machine failures, overloading, etc. However, some of the more interesting challenges emerged in cases where assumptions made by the data sources turned out to be mistakes. For example, one source assumed that “there will never be more than 1000 comments on a video”.

Finally, there were challenges that really cannot be predicted, for example YouTube going offline because

<sup>12</sup> 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. We measure time 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 be back-solved to give requisite crawl rate for any error bound. For more details, including prioritization scheme, see (Grace et al, 2008)

the world wide DNS entries were corrupted. In all these cases, lessons learned were encoded in the control framework to catch when similar events happen, work around them when possible, and alert the operator otherwise.

## 6 Pre-processing

Pre-processing is the task of transforming the crawled data into a normalized format. For structured data, such as song listening information from LastFM, or vital statistic information from a medical system, the normalization function simply maps the data fields into a internal schema. Unstructured data, such as the user comments on MySpace or surgical notes from a EMR system, requires source-specific mapping functions that extract individual comments and metadata and potentially supplement this information with other structured data such as poster demographic data.

This transformation of data is by far the most brittle step of the entire system. Web page layout and service APIs often change without notice, requiring manual adjustments to the site-specific mapping functions. Thus this is one of the areas the control framework must monitor most closely, allowing rapid notification to the operator when changes in the amount of generated information may indicate a broken parser.

Once the data is normalized, it is stored in a relational database (DB2). Relational models have the advantage of being easily extensible (e.g., through the addition of new columns) for aspects of future data sources. Each data element (e.g., comment) is uniquely identified by a combination of features around it - for music, these might be user, data source, artist and timestamp (best estimate or exact).

Many data flow system combine the fetch and format steps into a single module, but we have found with SI systems this is not a good idea. We differentiate the task of fetching the content as is from the source provider (crawling), and the task of breaking that content into usable chunks (pre-processing) as the failure modes of the two are quite different and need to be handled in very different ways. Given the difference in WAN and LAN speeds and the restrictions on how often a site can be hit, it is much faster to re-pre-process the data than it is to re-crawl it. Thus staging all the crawled data in "source form" makes more sense for SI systems than it may for many others.

## 7 Semantic Annotation of User Comments

In order to support the kinds of roll-ups, aggregations and trending that we would like to perform in an SI sys-

tem, the wealth of unstructured user-generated content requires transformation into a structured form by identifying particular entities such as mentions of artists, albums and tracks within the posts. In this Section, we describe our approach for extracting three types of metadata from user comments: artist and track mentions, sentiment orientations and spam annotations.

### 7.1 Named Entity Recognition (NER)

Identifying named entity mentions in text becomes increasingly complicated when there is insufficient context surrounding the discourse (Nadeau and Sekine, 2007). The language used on social networking sites is in the Informal English domain — a blend of abbreviations, slang and context-dependent terms, delivered with an indifferent approach to grammar and spelling. This task is also more challenging when the words referring to named entities are also words used commonly in everyday language, such as "Yesterday," which could refer to the previous day, a Beatles song (one of 897 songs with that title), or a movie (there are three such movies).

Annotating named entities can take advantage of semantic domain models that provide information on how entities in a domain might relate to one another. This information can complement traditional statistical natural language processing (NLP) techniques to assist in entity spotting. In our work, the problem of entity identification is approached as that of entity spotting. First, we identify a known list of named entities in the text and subsequently disambiguating its reference.

Domain dictionaries have been widely used in NER, including the use of Wikipedia (Bunescu and Pasca, 2006), Wiktionary (Muller and Gurevych, 2008), DBLP (J Hassell and Arpinar, 2006), and MusicBrainz (Alba et al, 2008a). Recently, Bunescu and Pasca exploited the linked and textual features of Wikipedia to perform named entity disambiguation (Bunescu and Pasca, 2006), an important step in accurately extracting named entities. These provide inspiration for our work, demonstrating that it is possible to do efficient and accurate NER using domain knowledge supplemented with NLP techniques. Our work differs in how we systematically constrain a domain knowledge base in annotating a set of known named entities in Informal English text.

#### 7.1.1 Domain Model Restrictions to Improve Precision

We performed semantic annotation of track and album name mentions with respect to the MusicBrainz RDF<sup>13</sup>. MusicBrainz is a knowledge base of instances, metadata

<sup>13</sup> <http://wiki.musicbrainz.org/RDF>

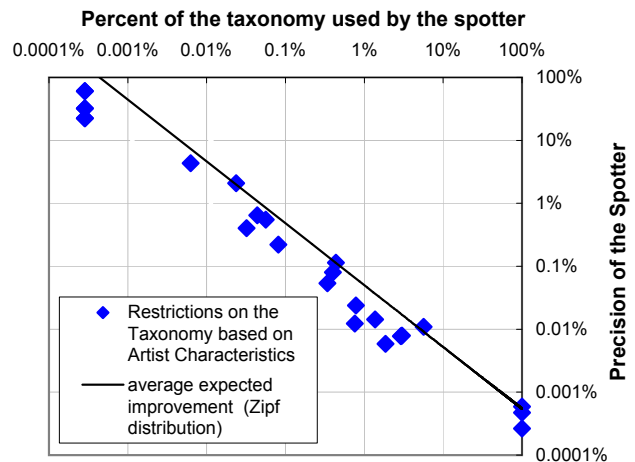
and relationships in the music domain. We used a snapshot of MusicBrainz that contained 281,890 artists who had published at least one track and 4,503,559 distinct artist/track pairs.

When working with ontologies it is often possible to restrict the set of entity candidates and thereby increase spotting precision. For example, instead of spotting millions of possible track names, it may be sufficient to only look for tracks of a specific artist when processing comments from that artist’s page (Gruhl et al, 2009). The music domain is particularly well suited to these types of restrictions, as there is such a large overlap between song titles and Informal English utterances.

We verified this approach experimentally by spotting song titles on a data set of comments from MySpace pages for Madonna, Rihanna and Lily Allen. These artists were selected to provide variety while each being popular enough to draw comment. Starting with the complete list of song-title entities from MusicBrainz (the case where no information is known as to which artists may be contained in the corpus), the MusicBrainz RDF is then successively reduced with tighter restrictions (the most restricted entity set contains just the songs of one artist ( $\approx 0.0001\%$  of the MusicBrainz RDF)). At every restriction, a naive exact string spotter is used to identify song title entities appearing in comments on the pages of these three artists.

The restrictions for our experiment were derived manually from user comments such as, “I’ve been a fan for 25 years now,” “send me updates about your new album,” etc. We consider three classes of constraints – career, age and album based restrictions that meaningfully reduce the MusicBrainz RDF using metadata about the artists and tracks in the domain model. The goal is to verify if such real-world constraints can be used to systematically reduce the size of the entity spot set and therefore improve precision of the spotter.

**Findings:** The results of domain model reduction are shown in Figure 5. The 22 plotted experiments reflect restricting the RDF based on various specific characteristics of the artists gleaned from the comments on their MySpace pages. For example, for Lily Allen the comment “Happy 25th B-DAY!” restricts the RDF to only 0.081% of artists in MusicBrainz who have this birth year. Since the test corpus represents three specific artists, the name of the artist is the most effective narrowing constraint. In every case, restricting the domain model improved spot precision. An important finding from this study is that restricting a domain model’s scope effectively improves spotting precision, regardless of the type of restriction used, as long as the restriction only removes off-target artists.



**Fig. 5** Shows the average expected improvement in precision for a naive spotter on song title entities as the domain model (MusicBrainz) is restricted. Each data point reflects an experiment where the model was restricted based on characteristics of the artist, as gleaned from comments on their MySpace pages. Note that the graph is in log-log scale.

An important area of interest stemming from this work is to develop automatic constraint selection based on the user comments themselves, for example, a “birth-date note” detector, a gender of artist identifier, a recent album release detector, etc. Once a robust set of content based constraint detectors are developed we can begin to experiment on “free domain” spotting - that is spotting in domains where less focused discussions are expected, e.g. Twitter messages.

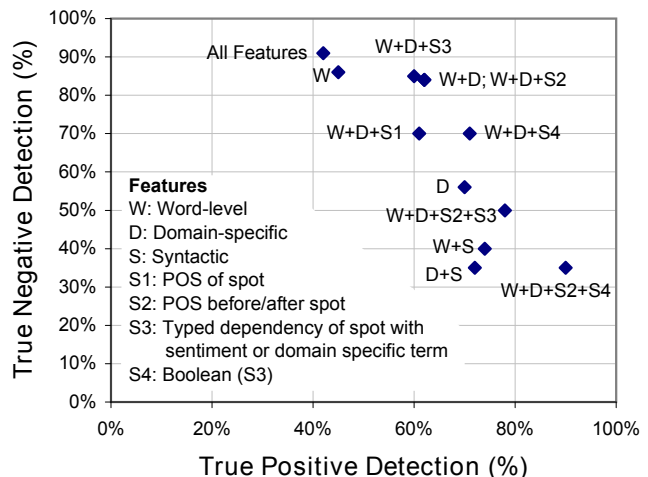
### 7.1.2 NLP to Further Improve Precision

Although the naive spotter has the advantage of spotting all possible mentions (modulo spelling errors), it generates a large number of false positives. For example, the comment “Got your new album **Smile**. Simply loved it!” contains a valid mention of the track Smile; while the comment ‘Keep your **smile** on. You’ll do great!’ does not. To improve the precision as much as possible while reducing recall as little as possible, we chain the results returned by the naive spotter with an NLP engine that uses the context around a spot to decide whether a spot is indeed valid. We approach the task of deciding whether a spot is valid or not as a binary classification problem. We train a Support Vector Machine (SVM) classifier using syntactic, word-level and domain-specific features pertaining to a spot found in a comment. While the approach of using features surrounding positive and negative examples to train and test a classifier is not new (Joachims, 1998), it is important to experimentally validate the usefulness of different feature sets for entity recognition in this domain.

**Data, Features and Experiments:** Our training and test data sets for the three artists were obtained from hand-tagged data by four of the authors. Positive and negative test sets contained all spots that at least three authors had confirmed as valid or invalid spots and spots where two authors had agreement on the validity of the spot and the other two were not sure. This yielded a training set of 550 valid spots and 550 invalid spots and a test set of 120 valid spots and two test sets of 229 invalid spots each. The larger test set of invalid spots was further divided into two equal sets to test the generality of the features.

We experimented with three types of features to make a decision about the validity of a spot (also see Figure 6): Syntactic natural language features (S)<sup>14</sup>; word-level (W) and domain-specific features (D) surrounding a spot. The syntactic features include the part-of-speech (POS) tags of the spot and the surrounding words (S1, S2). Word-level features include the capitalization or presence of quotes surrounding a word. Domain-specific features were used following the observation that spots that co-occurred with sentiment expressions and domain-specific words such as ‘music’, ‘album’, ‘song’, ‘concert’, etc., were more likely to be valid spots. In order to identify sentiment expressions in comments, we curated a sentiment dictionary of common positive and negative expressions (see Section 7.2) from UrbanDictionary (UD)<sup>15</sup>. The syntactic grammatical relationship or typed dependencies between the spot and the domain and sentiment terms were also employed as a feature (S3). For example, the typed dependency between the sentiment expression ‘loved’ and the spot ‘Smile’ in the comment ‘Got your new album **Smile**. Simply *loved* it!’, is *nsubj(loved-8, Smile-5)* implying that **Smile** is the object of *loved*. In addition to the syntactic dependency features, we also tested the simple presence or absence of these features (S4).

The efficacy of the three types of features and various combinations thereof in predicting the labels assigned by the human annotators was then evaluated. All experiments were carried out using the SVM classifier from Chang and Lin (2001) using 5-fold-cross validations. Figure 6 reports those combinations for which the accuracy in labeling either the valid or invalid data sets was at least 50% (random labeling baseline). Accuracy in labeling valid and invalid spots refer to the percentage of true and false positives that were labeled correctly by the classifier.



**Fig. 6** Classifier accuracy in % of true positives and true negatives for different feature combinations (higher = better).

**Findings:** The experiments revealed both expected and surprising findings about the usefulness of feature combinations for this data. For identifying valid spots, the following combinations were most useful (see Figure 6):

- Word-level, domain-specific and contextual syntactic tags (POS tags of tokens before and after the spot) and the boolean typed dependency features (indicating the presence of a syntactic relationship between the spot and a sentiment or domain-specific term): 90% labeling accuracy
- Word-level, domain-specific and contextual syntactic tags and the POS tags for the typed dependency features (the syntactic relationship between the spot and a sentiment or domain-specific term): 78% labeling accuracy

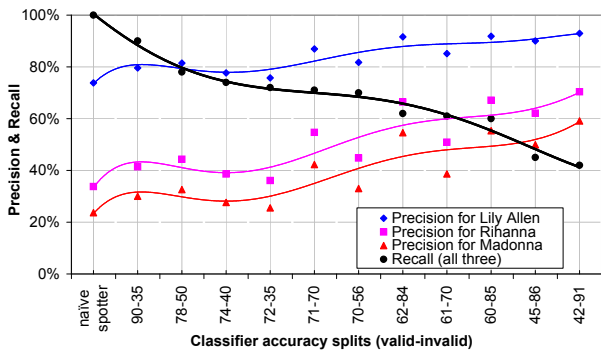
This suggests that *local word descriptors along with contextual features* are good predictors of valid spots in this domain. For identifying invalid spots, the following combinations were most useful:

- All features: 91% labeling accuracy
- Word-level: 86% labeling accuracy
- Word-level, domain-specific and the POS tags for the typed dependencies: 85% labeling accuracy
- Word-level, domain-specific: 84% labeling accuracy
- Word-level, domain-specific and contextual syntactic tags: 84% labeling accuracy

It is interesting to note that the POS tags of the spot itself were not good predictors for either the valid or invalid spots. However, the POS typed dependencies were fairly effective in catching invalid spots. In other words, the actual syntactic relationship between a potential spot and a sentiment or domain-related word helps to decide if a spot is a music related entity or

<sup>14</sup> Obtained using the Stanford NL Parser <http://nlp.stanford.edu/software/lex-parser.shtml>

<sup>15</sup> [www.urbandictionary.com](http://www.urbandictionary.com)



**Fig. 7** NLP PR curves for three artists and feature combinations. This chart helps select the tuning point of such a classifier for a particular SI application, based on the trade-off between not missing anything (high recall) and limiting the amount of data an analyst needs to examine (high precision). A 42-91 split corresponds to features (see Figure 6) that resulted in 42% and 91% accuracies in identifying true positives and negatives.

a common word in English. This suggests that not all syntactic features are useless, contrary to the general belief that syntactic features tend to be too noisy to be beneficial in informal text.

### 7.1.3 Overall Named Entity Spotter Accuracy

We chained two annotators together, first annotating comments using the naive spotter, followed by eliminating false positives using NLP analytics. This approach allows the more time-intensive NLP analytics to run on less than the full set of input data, as well as providing a certain amount of control over the precision and recall of the final result.

We measured the overall improvement in spotting accuracy for this mining chain. Ordered by decreasing recall, Figure 7 shows an increase in precision for the different feature combinations after boosting the naive spotter with the NLP analytics. For example, the precision of the naive spotter for Madonna’s spots was 23% and increased to almost 60% using the feature combination (all features) corresponding to the 42% and a 91% accuracy in identifying true positives (valid spots) true negatives (invalid spots) as shown in Figure 6.

The final result of this annotator is a structured representation of spotted artist and track/album names which is loaded into an OLAP cube, supporting the kinds of roll-ups, aggregations and trending that we would like to perform in an SI system.

## 7.2 Understanding Sentiment

Another metadatum used for gauging popularity is the sentiment associated with a comment post. In a clas-

sic setting with a paragraph or multiple paragraphs of text, the task would typically be to extract the sentiment and the polarity as directed toward an artist or their work mentioned in the comment. Given the nature of our data, where a comment is one or two sentences long and there is typically only one artist and/or their work mentioned, we make a simplifying assumption that the spotted sentiment expressions are always associated with the spotted entities in the comment. We do not explicitly verify if there is an attachment or syntactic dependency. Empirical evaluation also suggests that for sources such as MySpace where each artist has a page, few other artists are mentioned on pages other than their own. If no entity was spotted, for example in comments such as “I loved seeing you yesterday”, the sentiment is assumed to attach to the artist whose page the comment was found on.

Our approach for quantifying crowd preferences is related to work in opinion mining (OM) from public boards such as blogs, reviews and forums (Esuli, 2006). Our mining of sentiments about artists or their music differs from past work in OM because of the nature of our corpus and our goal 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 Hatzivassiloglou and McKeown (1997), Turney and Littman (2003) Esuli and Sebastiani (2005) and Kamps et al (2004).

One of the unique challenges that we faced, compared to previous efforts in this area, was the varied number of ways that users, typically in the teen demographic, tend to express sentiment. Slang sentiment expressions such as “wicked” to mean “good” or “tight” to mean “awesome” are commonplace.

Slang expressions of opinions are harder to detect because their usage has changed over time – for example, the usage of the word ‘sick’ has changed from bearing a negative to a positive connotation.. While we borrow from past work in using linguistic cues that identify tokens of words that might indicate sentiment expressions (traditional or slang) and Turney’s (Turney and Littman, 2003) work in identifying polarities, we also rely on an external domain resource to assist in this process given the informal nature of user-generated content. Our system first mines a dictionary of traditional and slang sentiments from UrbanDictionary.com (UD) and uses this to assist in the identification of sentiment expressions and their polarities.

### 7.2.1 Building a Sentiment Dictionary

Our sentiment dictionary maps frequently used sentiment expressions to their orientations, i.e., positive or

negative. The dictionary is built off a popular slang dictionary, UrbanDictionary.com (UD), that provides a set of related tags and user-defined and voted definitions for a term. Since glossary definitions are not necessarily accurate and automating the process of reducing them to a single sentiment is problematic, our system uses the related tags. For example, the slang expression ‘wicked’ has the following tags associated with it - ‘cool, awesome, sweet, sick, amazing, great’ etc. It is worthy to note that related tags are only indicators of possible transliterations. The tag ‘sick’ for example, appears as a related tag of both words ‘good’ and ‘bad’.

Our algorithm for mining a dictionary of sentiment expressions and their orientations taps into crowd agreements around the most current usage of slangs and proceeds as follows. Starting with a seed of five positive and five negative sentiment expressions (good, awesome, bad, terrible, etc.), UD is queried to obtain the top five related tags for each seed word. For each obtained related tag, we calculate its Semantic Orientation score with respect to the known positive and negatively oriented seed words. If the orientation of the related tag is toward the positive seed words, we pick the top positive seed word appearing in the list of words associated with the related tag as its transliteration. This process of obtaining new top five tags and determining their transliterations and orientations continues until no new tags are found.

We borrow Turney’s work for calculating the Semantic Orientation scores of words (Turney and Littman, 2003), with one modification. Instead of using the entire Web for co-occurrence statistics, we limit our calculations to UrbanDictionary where slang usage is unbiased by co-occurrences outside the slang context.

### 7.2.2 Sentiment Annotator: Experiments and Findings

Gauging the sentiment associated with a user comment, i.e., identifying word tokens that might express a sentiment and finding its polarity proceeds as follows:

1. Perform a shallow NL parse of a sentence to identify adjectives or verbs that suggest the presence of a sentiment (Hatzivassiloglou and McKeown, 1997).
2. Look for the spotted sentiment in the mined dictionary. If the word is not found, then compute the word’s possible transliteration using support from the corpus. To illustrate, transliterate the slang-sentiment “tight” to “awesome” because of the following co-occurrence strengths of “tight” with expressions in the mined dictionary sentiment words in the corpus: “tight-awesome” has a co-occurrence count of 456, “tight-sweet” 136, “tight-hot” 429, etc. Since

“tight” co-occurs the most with “awesome”, the polarity of the slang “tight” is recorded as positive. The suggested transliteration is picked only if statistically significant. The former method of transliteration via the mined dictionary associates a larger confidence with the spot compared to the latter that relies on weak corpus indicators of meaning.

3. Increase the confidence in the spotted sentiment if there is also an artist/music related entity spotted by the named entity spotter.
4. If the confidence is greater than a tunable threshold, record the sentiment and its polarity as an annotation value.

This annotator was evaluated in a proof of concept study using the MySpace corpus described in Section 4.4, containing 600,000 comments gathered over a period of 26 weeks. Precision and recall figures were calculated over a random sample of 300 comments for nine artists. The comments were hand labeled for the presence and orientation of sentiment expressions. Tunable cut-off thresholds for annotators were determined based on experiments. 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

Results indicate that the syntax and semantics of sentiment expression in informal text is difficult to determine. Words that were incorrectly identified as sentiment bearing, because of incorrect parses due to sentence structure, resulted in inaccurate transliterations that in turn lowered precision (especially in the case of the Negative Sentiment annotator). We also experimented with the dependency relationships between entities and sentiments expressions but found them to be both expensive and minimally effective, most likely due to poor sentence constructions.

Low recall was consistently traced back to one of the following reasons: A failure to catch sentiment indicators in a sentence (inaccurate NL parses) or a word that was not included in our mined dictionary, either because it did not appear in UD or because it had insufficient support in the corpus for the transliteration (e.g., ‘funnn’). Evaluating the annotator without the mined dictionary, i.e., only with the corpus support, significantly reduced the recall (owing to sparse co-occurrences in the corpus). However, it also slightly

improved the precision, indicating the need for more selective transliterations. This speaks to our method for mining a dictionary of transliterations that does not take into account the context of the sentiment expression in the comment – an important near-term investigation for this work. Empirical analysis also revealed that the use of negation words such as ‘not’, ‘never’ was rare in this data and therefore ignored at this point. However, as this may not be true for other data sources, this is an important area of future work.

### 7.3 Dealing with Spam

Like many online data sets today, this corpus suffers from a fair amount of spam — off-topic comments that are often a kind of advertising. A preliminary analysis shows that for some artists more than half of the comment postings are spam (see Table 4). This level of noise could significantly impact the data analysis and ordering of artists if it is not accounted for.

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

**Table 4** Percentage of total comments that are spam for several popular artists.

A particular type of spam that is irrelevant to our popularity gauging exercise are comments unrelated to the artist’s work; such as those making references to an artist’s personal life. Eliminating such comments and only counting those relevant to an artist or their work is an important goal in generating ranked artist lists. Since many of the features in such irrelevant content overlap with the ones of interest to us, classifying spam comments only using spam-phrases or features was not very effective. However, a manual examination of a random set of comments yielded useful cues to use in eliminating spam.

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. Some examples include “Buy our cd”, “Come check us out..”, etc.
3. Comments often had spam and appreciative content in the same sentence which implied that a spam annotator would benefit from being aware of the previous annotation results.

SPAM: 80% have 0 sentiments

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

NON-SPAM: 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.

**Fig. 8** Examples of sentiment in spam and non-spam comments.

4. Empirical observations also suggested that the presence of sentiment is pivotal in distinguishing spam content. Figure 8 illustrates the difference in distribution of sentiments in spam and non-spam content.

These observations also speak to the order in which our annotators are applied. We first perform named entity spotting, followed by the sentiment annotator and finally the spam filter.

#### 7.3.1 Spotting Spam: Approach and Findings

Our approach to identifying spam (including those irrelevant to an artist’s work) differs from past work, because the small size of individual comments and the use of slang posed new challenges. Typical content-based techniques work by testing content on patterns or regular expressions (Mason, 2002) that are indicative of spam, or by training Bayesian models over spam and non-spam content (Blosser and Josephsen, 2004). Recent investigations on removing spam in blogs use similar statistical techniques with good results (Thomson, 2007). 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.

Our spam annotator builds off a manually assembled seed of 45 phrases found in the corpus that are indicative of spam content. This seed was assembled by empirical analysis of frequent 4-grams in comments that are indicative of spam content. First, the algorithm spots these possible spam phrases and their variations in text using a window of words over the user comment

and computing the string similarity between the spam phrase and the window of words in the user comment. These spots along with a set of rules over the results of the previous entity and sentiment annotators are used to decide if a comment is spam. As an example, if a spam phrase, artist/track name and a positive sentiment were spotted, the comment is probably not spam. By looking for previously spotted meaningful entities we ensure that we only discard spam comments that make no mention of an artist or their work.

Table 5 shows the accuracy of the spam and non-spam annotators for the hand-tagged comments. Our analysis indicates that lowered precision or recall in the spam annotator was a direct consequence of deficiencies in the preceding named entity and sentiment annotators. For example, in cases where the comment did not have a spam pattern from our manually assembled list, 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 in a comment (one or two sentences) in addition to poor grammar necessitates more sophisticated techniques for spam identification. This is an important focus of our ongoing research.

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

**Table 5** Spam annotator performance

## 8 Generation of the Hypercube for Facets

Many modern Business Intelligence systems are designed to support exploration of the underlying dimensions of “facets” and the correlations and relationships between them. This can be modeled as a high dimensionality data hypercube (also known as an OLAP cube (Codd et al, 1993)). The system stores this in a relational database (e.g., DB2) to allow the analyst to explore the relative importance of various dimensions for the task at hand (for the BBC the popularity of musical topics). The dimensions of the cube are generated in two ways: from the structured data (e.g., age, gender, user location, timestamp, artist), and from the measurements generated by a chain of annotator methods that annotate artist, entity and sentiment mentions in each comment<sup>16</sup>. The resulting tuple is then placed into

<sup>16</sup> Due to the scale of data being examined these annotations are generated automatically by the UIMA analysis chain. While

a star schema in which the primary measure is a relevance with regards to musical topics. This is equivalent of defining a function.

$$M : (Age, Gender, Location, Time, Artist, \dots) \rightarrow M$$

In other words, we define a multidimensional grid where the dimensions are quantities of interest. What those are depends on the domain; for music, they are values such as age, gender, etc. We “fill in” each point in the grid with the number of occurrences of non-spam comments made by a poster of that age, gender and location posting at that time about that artist, song, album etc. Storing the data this way makes it easy to examine rankings over various time intervals, to weight various dimensions differently, etc.; for example, exploring how teenage male posters tastes have differed between the US and UK from week to week.

Interesting and relevant projections are one of the most important findings an SI analyst can discover. Once identified they can be coded as fixed outputs of the system for use in applications such as dashboards, alert systems and workflows that make use of this social information. For the SoundIndex we use primarily one such projection to ordered lists for each data source.

SoundIndex seeks to generate a “one dimensional” ordered list from the various contributing dimensions of the cube. In general the simple case of projecting to a one dimensional ranking can be performed as a linear projection which is then used to sort the artists, tracks, etc. It 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, the cube can be sliced and projected (marginalize dimensions) to form lists such as “What is hot in Dayton among 19 year old males?” and “Who are the most popular artists from San Jose?” These questions can be translated into the following mathematical operations:

$$L_1(X) : \sum_{T, \dots} M(A = 19, G = M, L = \text{“Dayton”}, T, X, \dots)$$

$$L_2(X) : \sum_{T, A, G, \dots} M(A, G, L = \text{“SanJose”}, X, \dots)$$

where,  $X$  =Name of the artist,  $T$  =Timestamp,  $A$  =Age of the commenter,  $G$  =Gender,  $L$  =Location.

These projections produce a “Top- $N$ ” list for each source that can later be non-linearly combined via voting (see Section 9). For the more complex BI style anal-

the results are spot checked by a human periodically to assure language drift has not occurred, in general, they run unsupervised.

ysis, the entire cube (or more specifically, updates to it) can be passed on to the front end system for user exploration.

## 9 Adjudicating Multiple Data Sources

We have now established a series of techniques for working with data from individual social networking sites. This Section presents our efforts to combine data from multiple sources in a manner that does justice to each individual site. Our initial insight was that voting theory provides a wide range of methods to consolidate different opinions and we explored their application to the SI space (Alba et al, 2008a).

In creating the BBC SoundIndex, we used data sources with different modalities. That is, each data source had different characteristics, as described in Section 4.

There is no clean partitioning of music fans — it is likely that many fans use different sites in parallel. In addition, content varies across sites, with some sites having a virtual monopoly on specific content (such as Jeffree Star on MySpace). Thus no single site can be deemed as the most authoritative. Furthermore, a large user base is not necessarily an indication of influence. It is necessary to take this into account when generating any BI-style application where it is key to capture different types of endorsements from as many sources as possible. Working with multiple data sets has the feel of a “Mash-Up,” i.e., things that were not designed to work together combined for a new purpose.

We must consider how relevant each source is as an indicator of community interest. As an example, listening to a song on a radio stream does not indicate the same level of interest as an in-depth discussion of a song or artist in an online forum. Furthermore, we must consider the problem of popularity vs. population. Every data source has different populations, and the users posting on a particular site are typically a small percentage of the total users. Thus it is not possible to simply “sum counts” across different data sources.

In order to combine these data sources in a manner that effectively indicates community interest, we must adjudicate. Our approach for top- $N$  lists as in the BBC SoundIndex project is to apply election techniques to artist popularity. A long democratic history of debating and modifying voting methods has provided a wide range of methods for adjudication.

Using voting theory to adjudicate top- $N$  lists has two stages: selecting possible voting systems, and judging which system is the best. Every voting system has different ways to make an election fair. For example, in the United States Presidential Election, a delegate system is employed, where votes are allocated to states in

proportion to the population. One of the goals of this system is preventing any single state or group of states to disproportionately overwhelm the election with a higher voting percentage of citizens.

In voting theory, the way in which “fairness” is measured is by choosing a social welfare function (SWF), typically from a group of established SWF’s, and then comparing the effectiveness of different voting systems. SWFs take the form of defined mathematical criteria for success, enabling us to quantitatively measure the success of a voting system for the criteria we have set. For example, the Spearman Footrule SWF measures overall distribution of influence from each voting sector in an election. This SWF would be a good way of measuring how well the US Presidential Election system fulfills the goal of distributing influence among all the states. Therefore, choosing a SWF depends on what need to be prioritized in an election system.

### 9.1 Best Adjudication Schemes

We considered 10 voting systems for adjudication of top- $N$  lists: total votes (as in a popular election), weighted votes (e.g., by population), semi-proportional methods (e.g., equal votes per source), a delegate system (e.g., non-linear by population), simple rank (summing rank across sources), Nauru inverted rank (summing  $\frac{1}{rank}$ ), run-off, and round robin. (Alba et al, 2008a)

We made use of two classic SWFs, each of which had a slightly different voting theory election that works best. We employ the Spearman Footrule distance (Diaconis and Graham, 1977), a SWF which emphasizes the preservation of position in the top- $N$  rankings. The other SWF was based on the Precision Optimal Aggregation method introduced by Adali et al (2006), which measures how many artists from each source’s top- $N$  list made it into the overall top- $N$  list.

The Runoff Method performed best when evaluated against the Spearman Footrule SWF, reflecting a good overall distribution of influence from each source. A Runoff Election is conducted by selecting the winning candidates from each source in a fixed order. Order is set randomly at the start of the election. Once the same candidate has been selected by at least 50% of sources it is added to the elected list and further mentions of it are ignored. This repeats on unselected candidates to fill the remainder of the list.

The Nauru Method (Inverted Rank) performed best when evaluated against the Precision Optimal Aggregation SWF. The Nauru method is a positional voting system, that is, an election based on rank rather than proportional vote count. Within each source, the rank

(1 = best) can be determined by simply sorting artists by number of votes. These ranked lists are then combined from different sources. Nauru is a method that rewards higher ranks more and lessens the impact of a single bad ranking. Thus, the Precision Optimal Aggregation SWF is satisfied by the resulting list having at least one candidate from each source near the top, whereas the Spearman Footrule SWF measures Nauru lower, as order is not being preserved.

Since this work on harmonizing multiple sources to generate Top- $N$  lists (Alba et al, 2008a), we have noted that a slight variant of the Nauru method can be created to either increase or decrease the “distance” between ranks by performing an operation on the denominator of the overall calculated rank. We have found this modification to be the most successful method.

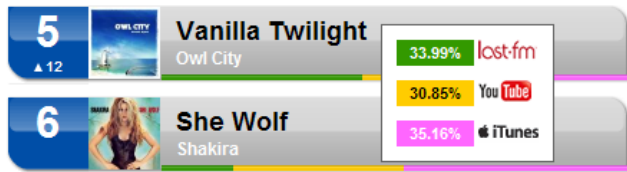
## 9.2 Successful Mashups

These voting systems all outperform majoritarian or semi-proportional methods, e.g., those used in the US state and federal election processes. They also dramatically outperform the common practice of simply adding “counts.” We validated this using a large data set gathered around the 50th Grammy Awards ceremony on Feb 10th 2008. For the two SWF’s described above, the best adjudication methods produce results that are up to 45% better than the standard summation.

We also observe that this process is a matter of fine-tuning. If we ignore absolute ranking and simply generate a bucket of top 40 artists, the top 10 artists from almost every ranking will appear in that list. The challenge is not in understanding which artists are generally popular, but in gauging relative popularity. This is symptomatic of the data sources — viewing a video on YouTube or making a comment on MySpace are such different forms of endorsement that cannot easily be compared. Thus the ordering of top artists is a particularly difficult problem if the artist’s support base varies from site to site.

## 10 Dashboards

All of the effort required to assemble this SI data is only worthwhile if the data is presented in a way that the user can make sense of it. The goal of a Business Intelligence system is to enable the understanding of the data by analysts. For an SI system, the goal is to enable users to rapidly grasp the evolution in the social landscape. The challenge is to allow non-specialist users to explore this complex data without having to understand all the subtleties of the system.



**Fig. 9** Individual entries in the SoundIndex indicate how the different sources contribute to their rank.

Our solution for the SoundIndex is the well-known concept of music charts, providing a set of predefined Top- $N$  lists, that even novice users are familiar with. Secondly, we enable drill-down into the data and the ability to explore the data by creating customized charts. The resulting user interface is similar to the *dashboards*<sup>17</sup> often provided in Business Intelligence applications. Two types of visualizations have proven particularly useful for rapid understanding and drill down in the music domain: ordered lists of aggregate data and timeline visualizations. These dashboard-type visualizations are both simple and familiar, something we have found to be true for financial analysis applications as well.

### 10.1 Top 10 Lists as Social Intelligence Dashboards

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”<sup>18</sup>. 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 (McIntyre, 1990).

Top- $N$  lists are a good way to visualize popularity data, particularly when analytics provide insights into where this popularity is coming from. In one glance an analyst can easily see what is most popular, and then drill down to understand further details. The SoundIndex Chart provides a visual clue about the underlying sources that contributed to the rank of an artist as shown in Figure 9. The colored bar underneath the song title and artist name indicates, which sources support

<sup>17</sup> Dashboards in Business Intelligence are analogous to a dashboard in a car. A collection of displays which often hold gauges of information that might be important or might be peripheral depending on the task at hand.

<sup>18</sup> en.wikipedia.org/wiki/Seven-Wonders-of-the-World

the artist. The figure also shows a dialog with detailed information that pops up when hovering over an entry.

The Internet empowers people to engage online and the data used in our SI application mainly stems from sites that invite user participation. We therefore wanted to empower the users of the SoundIndex to explore different dimensions of our data set and allow them to create their own charts. This helps users to develop an intuition about the data, and allows them to focus on the specific aspects that matter most to them. The analytical capabilities exposed in the BBC SoundIndex enabled users to explore different genres, population demographics, and data sources as shown in Figure 1.

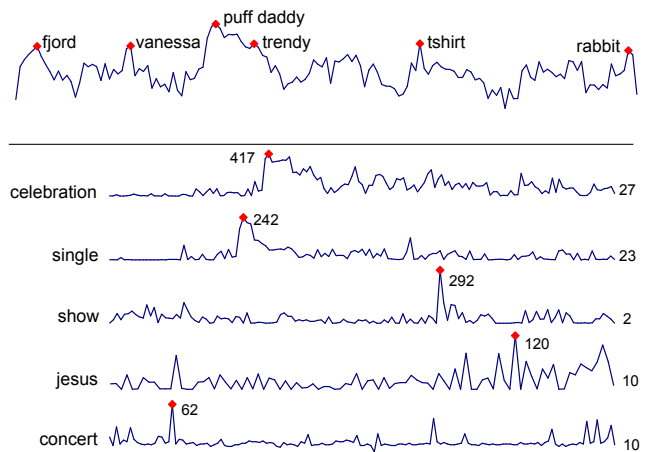
## 10.2 Timelines

The SoundIndex Charts are frequently updated. We indicate “movement” in the charts in the standard way, showing the previous rank below the current one (e.g. in Figure 9 “Vanilla Twilight” moved from position 12 to position 5). However, as changes occur much more frequently (the SoundIndex releases a new chart every six hours), timeline visualizations can capture the rank history of an artist over a longer period of time, e.g. using sparklines as shown in Figures 4 and 10.

Sparklines provide a clean effective visualization of time series data (Tuft, 2006). For analysts and other users of Business Intelligence applications, these simple visualizations provide a high volume of information in an easily understandable and summarizable format.

As part of our ongoing work, we explored the use of sparklines to visualize Twitter data annotated with our system. Twitter is a particularly interesting data source, because tweets are often posted as an event unfolds. Figure 10a shows tweets mentioning Madonna from July 28<sup>th</sup> through August 8<sup>th</sup> 2009. One interesting aspect of applying sparklines to this data set was to decide which labels to display in the graph. We found that annotating the highest local maximums with differentiating terms that summarize this time period worked well. The terms are identified by filtering for stop words and then calculating TFIDF for each term in one hour intervals.

Figure 10b shows the volume of tweets for the top five most common terms used in association with Madonna over the same time period. During this time, Madonna released her latest single, titled “Celebration”, and Twitter users were discussing her upcoming shows and concerts, and the release of several photographs of Madonna with her boyfriend, Jesus. It is worth noting that if projecting onto the “topic span” dimension, some of these discussions will drop out as they are not music related.



**Fig. 10** (a) Sparkline showing volume of tweets mentioning Madonna. The highest local maximums are annotated with differentiating terms for that time period (largest TFIDF term). (b) Drill down on data showing in the same timeframe the volume of tweets per hour for the five most common terms used in association with Madonna.

However, it is easy to see that they may be of interest for different SI applications.

These visualizations are complementary. While Figure 10a provides an overview of the volume of tweets about Madonna, Figure 10b is more effective in identifying big events and trends. The differentiating terms in Figure 10a provide an additional view into the data, revealing fringe discussions of interest. For example, this data set includes a discussion of Madonna’s plans to take her children on a fjord tour, speculation about collaboration with other artists, such as Vanessa and Puff Daddy, and a discussion of her fashion sense including t-shirts and hair bows resembling rabbit ears.

The explanations behind these keywords reveal the most important lesson from developing SI visualizations: enable drill down. For an analyst to grasp these changes in the social landscape, it is essential that they be enabled with the actual data behind the summarization. For data mined from social networks, this provides a simple rule: always let your analyst drill down to the original comments.

## 10.3 Other Visualizations

Beyond Top- $N$  lists and sparklines, there exist a myriad of options for working with SI data. Since the various dimensions that might be selected over are loaded into an OLAP cube, there are many options for importing this data into commercially available analysis tools. We have explored this data using the IBM Cognos Business Intelligence software, and found that this is an excellent option for advanced analysts.

The overall goal of an SI system, however, should be to summarize and simplify the data into an easily understandable fashion, while enabling drill down to explore the source and means of the summaries.

## 11 Validation

Matching the right crowd to the right question is key in developing an effective SI application. In general, the right approach is to have (hopefully several) subject matter experts look at the results and “sanity check” them. This is often the easiest way to catch problems in the system early on. Once the subject matter experts believe the system to function well, it is important to check the results against a “ground truth”.

With SoundIndex, the predicted *popularity* of artists and songs is being examined. What we need to ascertain is how well the system creates a list that is preferred by the target audience, compared to current alternatives. The evaluation presented in this section is effectively a case study that we have used as a model for verifying Social Intelligence applications in a variety of domains.

### 11.1 Focus Groups for “Ground Truth”

Before beginning development of the full SoundIndex system, we wished to evaluate the effectiveness of our general approach. That is, we wanted to know if it was even possible to use data from SNSs to create a better Top-10 list of popular artists. To that end, we conducted a series of experiments using MySpace data and prepared a top- $N$  list of popular music to contrast with the most recent Billboard list. Then we conducted a user study to validate the accuracy of our lists.

#### 11.1.1 Generating our Top- $N$ List

We started with the Billboard’s top-50 singles chart during the week of September 22nd through 28th, 2007. If an artist had multiple singles in the chart and thus appeared multiple times, we only kept the highest ranked single to ensure a unique list of artists, leaving us with 45 subjects. MySpace pages of these 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 5. The crawled comments were passed through the three annotators to remove spam and identify artist, track mentions and sentiments. The tables below show statistics on the crawling and annotation processes.

As described in Section 8, 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

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

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

**Table 6** Crawl Data (above) and Annotation Statistics (below)

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 are all 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 the 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 7 we show the top 10 most popular Billboard artists and the list generated by *our* analysis of MySpace for the week of the survey. While some top artists appear on both lists (e.g., Soulja Boy, Timbaland, 50 Cent, and Pink), there are important differences. In some cases, our MySpace Analysis list clearly identified rising artists before they reached the Top-10 list on billboard (e.g., Fall Out Boy and Alicia Keys both climbed to #1 on Billboard.com shortly after we produced these lists). Overall, we can observe that the Billboard.com list contains more artists with a long history and large body of work (e.g., Kanye West, Fergie, Nickleback), whereas our MySpace Analysis List is more likely to identify “up and coming” artists. This is consistent with our expectations, particularly in light of the aforementioned industry reports which indicate that teenagers are the biggest music influencers (Mediamark, 2004).

#### 11.1.2 The Word on the Street

Using the above lists, we performed a casual preference poll of 74 people in the target demographic. We conducted a survey among students of an after-school program (Group 1), Wright State (Group 2), and Carnegie Mellon (Group 3). Of the three different groups, Group 1 was comprised of respondents between ages 8 and 15; while Group 2 and 3 were primarily comprised of college students in the 17–22 age group.

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 7** Billboard’s Top Artists vs. our generated list

The survey was conducted as follows: the 74 respondents were asked to study the two lists shown in Table 7. 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.

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.

### 11.1.3 Results

The raw results of our study immediately suggest the validity of the system, as can be seen in Table 8. The MySpace data generated list is preferred over 2 to 1 to the Billboard list by our 74 test subjects, and the preference is consistently 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 8** Results: number of people who preferred each list

More exactly,  $68.9 \pm 5.4\%$  of subjects prefer the SI derived 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),  $88.2 \pm 8.1\%$  of subjects prefer the SI list to Billboard. This striking result shows a 6 to 1 preference for our list from younger listeners. We conclude that, to some extent, the billboard.com list represents the preferences of an older demographic than our MySpace analysis list—specifically a demographic with more buying power than 8-15 year olds.

A t-test for confidence suggests that respondents will prefer the SI list to the alternative with  $\alpha = 0.001$ . Thus a 99.9% confidence interval that a randomly polled group of similar individuals will show an overall preference for the SI generated list over the Billboard list. We can say with a high degree of confidence that the SI system is *better* at producing charts than the traditional method — at least for people in our sample group.

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.

### 11.2 Expert Validation

Our user study validated our general methodology — that is, information gleaned from SNSs can be used to better gauge music popularity. Having proved that this is possible, we went on to refine our technique, including adding additional data sources, and addressing the problems of adjudication.

As we began to work on a full system, we turned to expert validation rather than user studies to judge if our approach was successful. For the SoundIndex project, we were very fortunate to have access to some of the world’s experts on what music is popular — BBC Radio One DJs. The DJs worked with us through several iterations of the system, as we included more data sources, and modified the means of adjudication.

For each change we made, the DJs validated the new Top-N lists outputted by the system. In almost all cases they were able to spot results that looked odd and, upon examination, bugs were found. A few of these were simple workflow problems (i.e. accidentally dropping a source), but by working with the DJs we also began to understand the problem of adjudicating sources, which led to our voting approach.

This iterative design and validation provided a system that both we and our experts were very satisfied with. When we reached the point of launch, the DJ’s agreed that the system was not only an improvement on traditional bulletin boards, but also a novel way to explore music popularity.

### 11.3 After the Launch

Having validated our system with subject matter experts, SoundIndex faced its most critical evaluation: public launch. The positive response in the media provided the most satisfying validation of our end-to-end system. The Guardian UK described the SoundIndex as the “first definitive music chart for the internet age.”<sup>19</sup>

<sup>19</sup> www.guardian.co.uk/music/2008/apr/18/popandrock.netmusic

Several industry representatives not only enjoyed the system, but expressed a desire to have the raw data exposed so that additional SI-type systems could be built on top of it. ReadWriteWeb would “love to see the BBC offer some chart widgets, so that bands could display their rankings on their social networking profiles, and people could display their custom charts.”<sup>20</sup> and TechCrunch wished that the BBC would “release some of this data, perhaps on a platform, so that UK startups like Songkick can incorporate it into their service.”<sup>21</sup>. The Washington Posts noted that the “SoundIndex also lets users sort by popular tracks, search by artist, or create customized charts based on music preferences or filters by age range, sex or location.”<sup>22</sup>

The blogosphere acknowledges that “With the Sound Index the BBC has stolen a march on others (this could have sat well within Google/Yahoo! etc) and if the Sound Index is promoted/developed properly it could be a major draw to the BBC online music pages.”<sup>23</sup> and the SoundIndex “is an interesting twist on the traditional charts, potentially making a person’s actions as important as their songs as they attempt to create an internet buzz.”<sup>24</sup> This is exactly the kind of validation we could hope to see for an SI system.

Music experts from the BBC saw the SoundIndex as a paradigm shift in the industry. “It is seldom that one gets the opportunity to create a paradigm shift for an industry, especially one as established as ours. By bringing the People’s Chart to life through SoundIndex, the IBM team has done exactly that for the music charts industry” - Geoffrey Goodwin (Head, BBC Switch). “SoundIndex is the early leader in solving the very hard problem of charting music in this new digital age” - Stephen Davies (Head of Digital Enterprises, Audio and Music - BBC Worldwide)

#### 11.4 Ongoing Validation

The biggest challenge in validating the SoundIndex system is that the data sources are constantly changing. Since the time of launch, Facebook has eclipsed MySpace as the most popular SNS on the web. Sources we did not include in the launch, such as Twitter, have taken off dramatically. In such a constantly changing landscape, there is no one tuning of an SI system that provides the gold standard.

<sup>20</sup> [readwriteweb.com/archives/bbc\\_launches\\_sound\\_index.php](http://readwriteweb.com/archives/bbc_launches_sound_index.php)

<sup>21</sup> [eu.techcrunch.com/2008/05/20/bbcs-sound-index-is-good-but-we-wont-get-the-data/](http://eu.techcrunch.com/2008/05/20/bbcs-sound-index-is-good-but-we-wont-get-the-data/)

<sup>22</sup> [www.washingtonpost.com/wp-dyn/content/article/2008/05/20/AR2008052000519.html](http://www.washingtonpost.com/wp-dyn/content/article/2008/05/20/AR2008052000519.html)

<sup>23</sup> [www.nickburcher.com/2008/04/bbc-sound-index-great-new-way-of.html](http://www.nickburcher.com/2008/04/bbc-sound-index-great-new-way-of.html)

<sup>24</sup> [blog.webometrics.org.uk/2008/04/bbcs-sound-index-latest-and-greatest.html](http://blog.webometrics.org.uk/2008/04/bbcs-sound-index-latest-and-greatest.html)

The best approach is the one taken by the BBC-continuous validation with focus groups as the system progresses. Such work with focus groups is also important as a compliment to expert validation. For instance, while the expert validation approach was invaluable to the development of SoundIndex, there were a few hitches. In some cases the small group of experts, in talking to each other, became convinced that a particular artist *should* be popular.

Upon examination of the source data we found very little evidence to support that they were. This kind of issue is a classic case of the concern Surowiecki (2004) raised with the independence of such groups. It is easy for a small group of experts to self-reinforce an opinion. This is particularly possible when the landscape changes so rapidly, as is the case with music popularity. This is where continuous evaluation and iterative system tweaking make the most difference. Even in domains such as healthcare, where the data sources are (hopefully) less changeable, it is still impossible for a system to give good results without constant tuning. The “right” approach will differ depending on what the system seeks to show, but the general methodology holds — initially validate with experts and then compare the results against the current best approach, or when possible, ground truth data. Only then can confidence be developed that the SI system is producing results an analyst can rely on.

## 12 Related Work

Moving into the Social Intelligence domain we have borrowed heavily from the areas that have informed natural language text processing and Business Intelligence. In this Section, we provide references to some of the background literature that we use. While there are a number of full “buzz tracking” systems in existence, many focus on providing reference to specific mentions (e.g., Google and Yahoo news both offer alerts on news, web sites and blogs). The alternative are services, which plot popularity of particular terms or phrases, either in terms of what is mentioned or searched upon (e.g., Facebook’s Lexicon, Google Trends, Nielson’s Blogpulse). While these (and others) cover some aspects of Social Intelligence, we are not aware of systems at the moment that combine broad data acquisition, deep natural language analytics, and traditional Business Intelligence visualization and manipulation workbenches.

### 12.1 Metadata Extraction

Providing a unified view of data extracted from a number of domains requires resolving heterogeneities between their elements (Sheth, 1999). Traditionally this

is attempted on low level features (e.g., combining ‘address’ elements between Google Maps and Craigslist for a housing mashup). However, unstructured data sources lack the well defined structure of such features. Thus metadata extraction needs to be performed before integration. Extracting data from Social Networks draws from two domains of literature — *Information Extraction* (IE) and *Natural Language Mining* (NLM).

IE exploits the structure of the content in the gathered data. Metadata is extracted via manually coded or automatically induced wrappers that are specific to the particular data source (Kushmerick, 1997; Freitag and Kushmerick, 2000). In our work, the user’s demographic information, post timestamps, etc., are examples of this kind of structure.

This contrasts with NLM, which analyzes the unstructured part of the content to extract information nuggets (Freitag, 1998; Soderland, 1997). An SI system can employ these approaches to process user posts from sites such as MySpace and Bebo. The SoundIndex uses a series of UIMA (Ferrucci and Lally, 2004) annotators that are driven by the basic entity spotting. For the music domain, such metadata extracted from unstructured posts includes specific artist and track mentions.

## 12.2 Rank Ordering and Popularity

Top- $N$  lists are a familiar tool for expressing importance or popularity. As such, it is an important framework to present SI information in, and understanding its influence has been the subject of study in politics, arts and economics (Adorno, 1945; Mayzlin and Chevalier, 2003). While these studies have been motivated by the communication industry’s monitoring of popularity trends and consumer behavior, results have also been used to understand how a culture’s operation shapes “hit lists” (Riesman, 1950).

Social science teaches that fundamental to popularity is the presence of a vocal, popular minority compelled to share their opinions with a larger audience. These influencers are not a new phenomenon – Lasswell (1948)’s comments on communication models summarizes the role that a community and their opinions can have on trends and popularity lists. Communications within a society were shown to often be an inextricable part of outcomes such as voting preferences.

We leverage this in our work by noting that popularity often develops in and spreads through social communication channels. Thus SI is based on the hypothesis that one can determine what is popular or important by measuring 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 will be popular tomorrow.

## 12.3 Rank Aggregation as Voting

Creating ranked lists from other ranked lists is a classic problem in voting theory. This has been used in the past to improve search applications on the Web and to combat “spam” (Dwork et al, 2001). Combining results of multiple selection criteria has a long history in information retrieval, where various metrics are employed to evaluate rankings and combine results (Baeza-Yates and Ribeiro-Neto, 1999). For information retrieval, success is often measured using a weighted harmonic mean of precision and recall called an F-measure (Makhoul et al, 1999). We can consider this a kind of voting where the social welfare function is precision and recall.

For the SoundIndex, which is seeking a measure of popularity, a very different social welfare function is needed. Each site’s opinion should matter and be counted towards the total ranking. Rank aggregation is employed in database applications to combine similar results (Fagin et al, 2003). The social welfare function again differs, thus these rank aggregation methods cannot be directly employed for music popularity.

Adali et al (2006) study rank aggregation, measuring the quality of different aggregation methods with respect to the effect that an individual source can have on the outcome. This study inspired one of the social welfare functions we employ to study voting systems.

Many voting systems have been proposed since the field’s inception in the 18<sup>th</sup> century (Saari, 1994), especially since Arrow (1951, 2nd ed., 1970) showed that perfect voting schemes do not exist. Objective comparative effectiveness of voting systems was established by Bergson (1938), through the development of the Social Welfare Function. Recent work by Balinski and Laraki (2007a) defines a more general Social Decision Function. These define a mathematical criteria for the success of a voting system using a list of characteristics which are desired (e.g., if the majority of voters prefer a single candidate to all others, that candidate is elected).

Within voting theory, the problem of finding Top- $N$  lists can be expressed as a multiple-winner voting system. To do so, each source is asked, “What artists should fill the top- $N$  slots of our ranked list of popular artists?”. Simple majority and plurality are the most popular majoritarian voting systems (Riker, 1982). While 50% of the votes are required to win a simple majority, the plurality system simply chooses the winner that attains the maximum number of votes without having to surpass the threshold. Plurality voting can be extended to the multiple-winner scenario by counting and sorting votes to select top- $N$  winners while also normalizing for source population sizes (an often difficult task with

social network sites who notoriously like to claim the largest numbers of users).

Whilst majoritarian methods use information from binary comparisons between choices, positional voting methods use the source's preference orderings as well. Thus each "voter" ranks the candidates in order of preference. In a multi-source SI system each source provides a ranking that is used to generate a combined preference ordering. We have found that variations of the popular Borda Count (de Borda, 1981) work well. They are known to generate a complete, transitive social ordering supported by a broad consensus, rather than the choice that is favored by a majority.

### 13 Conclusions

Social Intelligence applications are an exciting new area of research and development that has become possible through the rise of user participation in digital media. In this paper we present the BBC SoundIndex and show that creating such an end to end system requires skills and research from a wide range of disciplines. We are certain that many of our approaches can be improved upon. We hope that the reader is left with a good overview of what is required to build an SI system and hope to spark interest in tackling some of the exciting challenges that we encountered on the way.

The data that drives a Social Intelligence application has some important differences from traditional enterprise data, and those differences can make it difficult to leverage for Business Intelligence applications. Spam can make even focused data streams very noisy — certainly more so than traditional sales data. Ascertaining what the data relates to can be difficult as well, as rather than being facts collected for a specific purpose, social data is often opinions expressed in Informal English. While in business settings people often avoid comparing "apples to oranges", almost every source of social data was constructed differently and for different purposes, thus every aggregation in SI is an "apples to oranges" problem.

However, once these challenges are met, the rewards are striking. We have seen bands leap to the top of the charts within hours of a defining event, often before the subject matter experts were even aware the event had happened. Charts assembled via SI can challenge the conventional wisdom that focused polling is the best way to garner preferences. We have seen the excitement as end users, who traditionally were provided a fixed report, suddenly were able to play with the data and develop an understanding of the social landscape that matters to them.

There is an opportunity with SI applications to bring the benefits of very large data set analysis out of the en-

terprise and to the end user. In doing so we have found that all users are analysts in the areas that interest them. All they need are the tools.

### Acknowledgements

We thank our teammates on the SoundIndex project, without whom this work would not have been possible: Alfredo Alba, Varun Bhagwan, Julia Grace, Kevin Haas, Nachiketa Sahoo, and Tyrone Grandison. We particularly thank Alfredo, Varun, Kevin, and Tyrone for all their systems expertise, Julia Grace for her visualization skills, and Nachi for assistance with spam detection and the focus group study. We thank Geoff Goodwin, Head of BBC Switch, for his vision, support and encouragement, and Beth Garrod, BBC Producer, for her domain expertise and guidance. We thank Marti Hearst, Ron Fagin, Sandra Yuen and Amy Vandiver for their advice and guidance. We thank Bill Cody, Steve Dill, Jeff Pierce, and Laura Haas for their support and encouragement of this project.

### References

- Adali S, Hill B, Magdon-Ismael M (2006) The impact of ranker quality on rank aggregation algorithms: Information vs. robustness. In: ICDEW '06, p 37
- Adorno TW (1945) A social critique of radio music. *Kenyon Review* pp 208–217
- Alba A, Bhagwan V, Grace J, Gruhl D, Haas K, Nagarajan M, Pieper J, Robson C, Sahoo N (2008a) Applications of voting theory to information mashups. In: ICSC, pp 10–17
- Alba A, Bhagwan V, Grandison T (2008b) Accessing the deep web: when good ideas go bad. In: OOPSLA Companion '08, ACM, pp 815–818
- Arrow KJ (1951, 2nd ed., 1970) *Social Choice and Individual Values*. Yale University Press
- Baeza-Yates RA, Ribeiro-Neto B (1999) *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA
- Balinski M, Laraki R (2007a) A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences* 104(21):8720
- Balinski M, Laraki R (2007b) A theory of measuring, electing, and ranking. *PNAS* 104(21):8720–8725, DOI 10.1073/pnas.0702634104
- Bergson A (1938) A Reformulation of Certain Aspects of Welfare Economics. *The Quarterly Journal of Economics* 52(2):310–334
- Bhagwan V, Grandison T, Alba A, Gruhl D, Pieper J (2009) *Mongoose: Monitoring global online opinions via semantic extraction*. In: SQAM workshop at IEEE 2009 International Conf. on Cloud Computing

- Blosser J, Josephsen D (2004) Scalable centralized bayesian spam mitigation with bogofilter. In: USENIX conference on System Administration
- de Borda JC (1981) Memoire sur les elections au Scrutin. Histoire de l'Acad. R. des Sci.
- Bunescu RC, Pasca M (2006) Using encyclopedic knowledge for named entity disambiguation. In: EACL, The Association for Computer Linguistics
- Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Codd E, Codd S, Salley C (1993) Providing OLAP (online Analytical Processing) to User-analysts: An IT Mandate. Codd & Date, Inc.
- Cody WF, Kreulen JT, Krishna V, Spangler WS (2002) The integration of business intelligence and knowledge management. IBM Systems Journal 41(4)
- Diaconis P, Graham R (1977) Spearman's Footrule as a Measure of Disarray. Journal of the Royal Statistical Society Series B (Methodological) 39(2):262-268
- Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the Web. Proceedings of the 10th Int Conf on World Wide Web pp 613-622
- Esuli A (2006) Survey of techniques for opinion mining. Language and Intelligence Reading Group
- Esuli A, Sebastiani F (2005) Determining the semantic orientation of terms through gloss classification. In: CIKM '05, ACM Press, pp 617-624
- Fagin R, Kumar R, Sivakumar D (2003) Efficient similarity search and classification via rank aggregation. Proceedings of ACM SIGMOD
- Ferrucci D, Lally A (2004) Uima: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng
- Freitag D (1998) Information extraction from html: application of a general machine learning approach. In: Conference on AI/Innovative Applications of AI
- Freitag D, Kushmerick N (2000) Boosted wrapper induction. In: Proceedings of the 17th National Conf. on AI/Innovative Applications of AI
- Grace J, Gruhl D, Haas K, Nagarajan M, Robson C, Sahoo N (2008) Artist ranking through analysis of online community comments. IBM Tech Report
- Gruhl D, Nagarajan M, Pieper J, Robson C, Sheth A (2009) Context and domain knowledge enhanced entity spotting in informal text. In: ISWC
- Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Association for Computational Linguistics
- J Hassell BAM, Arpinar I (2006) Ontology-driven automatic entity disambiguation in unstructured text. In: ISWC '06
- Joachims T (1998) Text categorization with support vector machines. In: Lecture Notes in Computer Science: Machine Learning, Springer Verlag
- Kamps J, Marx M, Mokken R, de Rijke M (2004) Using wordnet to measure semantic orientation of adjectives. URL [citeseer.ist.psu.edu/kamps04using.html](http://citeseer.ist.psu.edu/kamps04using.html)
- Koutsoukis NS, Mitra G, Lucas C (1999) Adapting online analytical processing for decision modelling: the interaction of information and decision technologies. Decision Support Systems 26(1):1 - 30
- Kushmerick N (1997) Wrapper Induction for Information Extraction. PhD thesis, U. Washington
- Lasswell HD (1948) Listening to popular music. The Communication of Ideas
- Locke LA (2004) Super searches. Time Magazine
- Makhoul J, Kubala F, Schwartz R, Weischedel R (1999) Performance measures for information extraction. Proceedings of DARPA Broadcast News Workshop
- Mason J (2002) Filtering spam with spamassassin. In: Proceedings of HEANet Annual Conference
- Mayzlin D, Chevalier JA (2003) The effect of word of mouth on sales: Online book reviews. Yale School of Management Working Papers
- McIntyre M (1990) Hubbard hot-author status called illusion. <http://www.scientology-lies.com/press/san-diego-union/1990-04-15/hubbard-hot-author-status-illusion.html>
- Mediamark (2004) Teen market profile. [www.magazine.org/content/files/teenprofile04.pdf](http://www.magazine.org/content/files/teenprofile04.pdf)
- Muller C, Gurevych I (2008) Using wikipedia and wiki-tionary in domain-specific information retrieval. In: Working Notes for the CLEF 2008 Workshop
- Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Linguisticae Investigationes
- Riesman D (1950) Listening to popular music. American Quarterly 2(4):359-371
- Riker WH (1982) Liberalism Against Populism. Waveland Press, Inc., Prospect Heights
- Saari DG (1994) Geometry of Voting, vol 3 of Studies in Economic Theory. Springer-Verlag
- Sheth AP (1999) Changing focus on interoperability in information systems: From system, syntax, structure to semantics. Interoperating Geographic Info Sys
- Soderland S (1997) Learning to extract text-based information from the world wide web. KDD '97
- Surowiecki J (2004) The wisdom of crowds. Doubleday
- Thomason A (2007) Blog spam: A review. In: Fourth Conference on Email and Anti-Spam CEAS 2007
- Tufte E (2006) Beautiful Evidence. Graphics Press
- Turney PD, Littman ML (2003) Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans Inf Syst 21(4):315-346