# An Analysis of Lyrics Questions on Yahoo! Answers: Implications for Lyric / Music Retrieval Systems

*Sally Jo Cunningham, Simon Laing*

Computer Science Department
University of Waikato
Hamilton 3240 New Zealand

*{sallyjo, simonl} @cs.waikato.ac.nz*

**Abstract**  This paper analyzes 237 questions posted to Yahoo! Answers, a popular community-driven question and answer service. The questions are all natural language and are self-categorized by their poster as being related to music lyrics, and as such they provide a rich context for understanding lyrics-related information behavior outside the constraints imposed by specific lyrics retrieval systems. We categorize the details provided in the queries by the types of music information need and the types of music details provided, and consider the implications of these findings for the design of music/lyric systems and for music retrieval research.

**Keywords**  User studies, multimedia document retrieval, music digital libraries

## 1  Introduction

Creating a useful and usable music retrieval system is a notoriously difficult task. A music document may consist of a symbolic representation of a work (eg, a score or MIDI encoding), an audio file (eg, MP3), an image (eg, a CD cover), textual metadata (a work's title, artist, composer, etc.), lyrics, a video of a performance—or a combination of any or all of the above [4].  Significant problems have yet to be resolved with document / query representation schemes, retrieval algorithms, and interface support in this challenging research area.

This paper focuses on identifying problems in developing systems for supporting lyrics-based information needs. At first glance it would appear that creating a lyrics-based music digital library would be one of the more straightforward development efforts in music retrieval, given that text-based retrieval is a better understood endeavor than image, video, and audio retrieval. This paper is a preliminary investigation into whether or not existing music retrieval research can address (or is addressing) support for lyrics retrieval systems.

Our approach is based on developing an understanding of what people want to find, and how they describe what they want, when they are trying to satisfy a lyrics information need. To that end, we analyze a set of lyrics related questions posted on Yahoo! Answers, an open Web-based question and answer forum.  Once this understanding emerges of what lyrics seeking behavior 'in the wild' (that is, outside the constraints of a retrieval system, and as expressed in natural language) then we can identify remaining problems in supporting lyrics retrieval.

## 2  Previous work

At present music retrieval research is only lightly informed by an understanding of user needs. For a variety of reasons—including intellectual property law, limited access to a significant and standard music testbed, and lack of access to usage records for emerging commercial music systems—it has been difficult for researchers in music retrieval to develop or exploit data concerning the music information behavior of target users. This situation is particularly problematic in that the common assumptions of 'typical' music behavior made by retrieval researchers and music system developers have been found to differ markedly from actual music behavior in the real world [4].

Query log analysis of music related interactions on Web search engines (eg, [12]) yield extremely coarse-grained information on music behavior; sessions are generally short, queries are generally brief, and the log provides no insight into the searchers' motivations, intended use of retrieved music documents, or satisfaction with the search results. Few usage studies exist of music digital libraries or specific music collections (eg, [5], [8]). These types of investigations are necessarily limited to providing insights into the usability of features implemented in the system studied; log data cannot suggest additional functionality or document types appropriate for the users. For both search engines and digital libraries, the user's information need is obscured by the requirement of complying with the query formats of a specific system.

What is required, then, is a source of authentic music information behavior and needs. Earlier examinations of music behavior are based on information requests harvested from music-related

newsgroups [3], question-answer services [7], and archives of mailing lists [2]. These resources are seeing use to the extent of providing immense quantities of raw data on a scale similar to web logs; however, manual analysis methods limit in practice the size of a harvested dataset to at most a few hundred requests. This type of investigation complements log analysis with a finer-grained understanding of music behavior.

Most technical music retrieval research focuses on integrating lyrics with audio: for example, aligning lyrics to audio signals (eg, [9]); or using lyrics as a basis for thematic or genre clustering and classification of related audio files (eg, [10]). Lyric retrieval has proved to be a special case of text retrieval, inspiring additional research into problems such as identifying and matching multiple (non-identical) lyrics for a single song [6] and supporting search over lyrics that are syllabicated as performance instructions [13].

## 3    Data gathering and analysis

Yahoo! Answers is an internet based reference site that allows users to both submit and answer questions. Unlike some earlier 'ask an expert systems' (eg, Google Answers), there is no charge to post a question and no financial reward to answer questions. Instead, the system is driven by a 'points' and 'levels' arrangement that rewards posters of correct answers with status within the Yahoo! Answers community.

When posting a question to Yahoo! Answers, the user is required to specify one or more categories for it. We focus in this paper exclusively on Entertainment & Music > Music > Lyrics posts. Yahoo! Answers sees heavy use; as of September 2009, the Lyrics subcategory alone contained over 226,000 questions that had been 'resolved' (that is, had received at least one acceptable response).

We harvested 250 questions posed on a single day in September 2009, from the newly posted ('open') section of the Lyrics category. Twelve were discarded as duplicates and one discarded as off topic, leaving 237 questions for analysis. The average question length was approximately 58 words; the longest question contained 291 words (a request for an explanation of a song's meaning, including the full lyrics), and the shortest a mere 7 ('*What are some of.....? your favourite lyrics?*'). By contrast, audio queries to conventional search engines are far more brief (eg, [12] report an average of 3.1 terms in a 2006 study of the metasearch engine Dogpile).

Grounded theory ([11]) was used to develop categories to elicit characterizations of the desired outcome for the queries (Section 4) and the information features provided by the poster (Section 5). Initial categories were established by bringing together features from previous studies of natural language music–related questions (eg, [1], [3], [7]). These categories were regarded as tentative and were revised based on examination of the Yahoo! Answers Lyrics queries. An iterative coding process was employed, continuing until the two researchers agreed on both the coding categories and the codes assigned to each question.

## 4    Characterizing the desired outcome

At this point, we examine the types of music information that the posters have specified that they would like to receive as a response to their question—that is, the types of music document or details that they are seeking (Table 1).

| Category | No. of queries | % (of 237) |
|---|---|---|
| Lyrics | 51 | 21.6% |
| Metadata | 95 | 40.3% |
| Identification | 36 | 15.3% |
| Copy | 6 | 2.5% |
| Example of type | 16 | 6.8% |
| Explanation | 16 | 6.8% |
| Feedback | 18 | 7.6% |
| Creative Practice | 7 | 3.0% |
| Other | 7 | 3.0% |

Table 1. Desired responses to questions

- *Lyrics*: requests for the complete lyrics to a song, or for specific lines (sometimes in a specific performance of a song)
- *Metadata*: requests for the title of a song and/or its artist / composer ('who it's by').
- *Identification:* questions asking some variation on 'what is this song?' without further specification of the desired result.
- *Copy*: requests to obtain a copy of an audio or video version of a song (by downloading or streaming).
- *Example of type*: requests for a song that fits into a specified category or genre (eg, a 'love song').
- *Explanation*: requests for 'the meaning' of a song and/or portions of the lyrics
- *Feedback*: the question solicits feedback on original song lyrics.
- *Creative Practice*: requests for technical or creative process information to be used in creating new songs.
- *Other*: questions that fall outside the above categories.

A close examination of the questions and their posted answers indicates Metadata and Identification can be collapsed into a single category; the desired result in both is a single song matching the given

criteria, with title and/or artist provided as a response. The Copy category is obviously closely related; a link to a song's audio will allow the poster to verify whether that song is indeed the requested music, and further the site hosting the audio (eg, YouTube) commonly includes music metadata such as title and author. A further breakdown of the 95 Metadata requests indicates that the title is the primary identifier for a song: 91 questions request a title, 25 ask for both title and artist, and 4 request the artist only.

The next largest category is that of requests for full or partial Lyrics for a specific song—the only surprise being that this is not the largest category, given that the poster has explicitly tagged the question as Lyrics focused. Most Lyrics requests appear to assume that there is only one set of lyrics for a song—they ask for '*the words*' or '*the lyrics'*.

For a minority of the Lyrics requests, the lyrics desired are to a specific performance or version of a song and so may not necessarily be the authoritative lyrics (eg, one question presents an audio link and asks, *'Can anyone decipher the lyrics up to the 25th second? plz? i am a nice guy?'*). There may not even exist an authoritative version for some songs, or portions of a song: for example, a 'freestyle' improvisation (*'the song is called "close my eyes" by Matisyahu. I can not find the lyrics for the freestyle he does in the middles of the song'*). Some queries explicitly request non-authoritative versions of the lyrics: for example, *'what's the lyrics of the song paradise by the kpop group "the melody"? the english translation, hangul and romanization please!'* The goal of existing work on identifying multiple  sets of lyrics for a single song [6] is to identify the authoritative version and eliminate 'mistakes' in other lyrics; these queries suggest that alternative lyrics should not necessarily be rejected, and that the identification of different versions may be more difficult than previously anticipated (for example, in matching translations to the original).

*Example of type* questions are not answered by a specific song (a 'known item search'), but instead seek to elicit one or more songs that match a type description. Picking out an answer from a set of potential matches is problematic; the standard default for a music retrieval system is to present textual metadata (eg, song title and artist), which is unlikely to convey the point of similarity between a song and the type description (eg, *'A Happy optomistic, catchy song'*). Providing appropriate support for browsing remains an open problem in music retrieval; coming to a deeper understanding of the song facets that are used to judge a match is required to drive interface development (eg, tempo? lyrics? affect?).

*Explanation* questions (*'What Is This Song About?'*) require a deep understanding of the semantics of the lyrics, and are unlikely to be addressable by automated retrieval systems.

Similarly, requests for *Feedback* and critique of original lyrics written by the poster and assistance in the *Creative Practice* of creating audio are well beyond the capacities of existing digital libraries. However, these questions highlight that a great deal of music behavior is embedded in a social context—we listen to music at social gatherings, talk about the latest hits in casual conversation, and play songs on the radio or a CD as we drive. It seems appropriate that a music retrieval system should support music experts, aficionados, and keen novices in discussion and in community-based reference services—that the vision of a music digital library could include people as well documents and software.

## 5   Characterizing the information features provided

The features or characteristics used to describe the 204 Lyrics, Bibliographic Details, Copy, Identify, Explanation, and Example queries are as follows:

| Category | No. of queries | % (of 204) |
|---|---|---|
| Lyric fragments | 113 | 47.9% |
| Storyline | 24 | 10.2% |
| Video references | 18 | 7.6% |
| Metadata | 95 | 40.3% |
| Genre/Style | 42 | 17.8% |
| Orchestration | 30 | 12.7% |
| Similarity | 11 | 4.7% |
| Where heard | 50 | 21.2% |
| Undesired result | 7 | 3.4% |
| Other | 2 | 0.8% |

Table 1.  How the information needs are described

- *Lyric fragments*: the remembered portions of a desired song.
- *Storyline* or message: a description of 'what happens' in a song, or a message conveyed by the song (eg, 'I love her and miss her').
- *Video references*: details about a video including the desired song (most frequently a music video for the song itself), provided either as a link to a video file or as a text description of the action occurring in the video.
- *Metadata*: bibliographic details, further broken down into Title, Artist, Collection Title, Date, Remix, and Tempo.
- *Genre or style*: can be a standard genre such as R&B, or a genre constructed by the poster (eg, 'contemporary, modern').
- *Orchestration:* an indication of the instruments and vocal parts in a recording.
- *Similarity:* another song or an artist that is similar to the desired song(s).

- *Undesired result*: another song, artist, or performance that is not the desired result.
- *Where heard:* the circumstances in which the poster heard a song performance or broadcast.

Finding a song based on the lyrics can be surprisingly difficult. Frustratingly, a person may remember the *Storyline* or gist of the song but not recall any of the lyrics themselves (*'the the song talks about hating someone so much they wish they would some how die'*). The lyrics for a song can be difficult to understand as sung, making it difficult to construct a text search based on the known partial lyrics (*'I have NO CLUE what ANY of the lyrics are except two words because I saw someone mouth them while the song was playing behind me… All I know is in the chorus it's "something something git'cha git'cha"'*). A related difficulty is the mondegreen—a misheard lyric that may seem plausible but is incorrect (*'someone in the backmground singing fly high or sky high or something like that'*). It can be difficult to decide how to enter lyrics as search terms; should "*git'cha git'cha*" be entered written? As *Get Ya Get Ya*? *Get You Get You*? *Gitcha Gitcha*? Moreover, some lyrics are not dictionary words (*'Cannot remember any of the lyrics for the life of me besides the chorus lyics which simple go: ooo ooo OOoo ooo ooo, ooo ooo OOoo ooo ooo (repeat)'*). These problems push conventional IR matching techniques such as latent semantic analysis to their limits and beyond.

A word or phrase in the lyrics may be too common to be helpful in constructing a search, but the manner in which it is sung can be distinctive enough to be useful (*"Free-ee-e-e-ee"*). Combining facilities for text and 'sung' audio in a query would neatly solve this problem (eg, [9]).

Posters are sometimes able to point to songs *Similar* to the desired result, or conversely to indicate songs that are known to not be an answer to the question (*'its definitely not Land of 1000 dances'*). Facilities for indicating closeness/distance of results to an exemplar would be useful for these queries and also to represent a song's degree of membership in a *Genre*.

*Metadata* provided is frequently tentatively presented as likely to contain errors (*'im not sure of the name of it i believe it's called "spirit"'*; *'i think it is by nirvana or rhcp or something like that'*)—understandably, since if the person had the correct metadata then they could answer their question themselves. The challenge for a retrieval system is to gracefully identify similar values to those suggested, for query refinement or ranking of results (eg, terms related to spirit, groups whose music is similar to that of Nirvana or the Red Hot Chili Peppers).

Where the poster heard the song might be useful in answering the question (*'What was the song played at the end of GH on 9/22/09?'*)—or it might not (*'What's the name of a song I heard at Red Lobster?'*). Again, this type of detail suggests the value of a community-based answering service to work with heavily context dependent questions.

## 6   Conclusions

This paper analyzes a set of Lyrics-related questions to tease out the types of details presented to describe the information need (Section 5) and the expected responses (Section 4); the findings can inform further music retrieval research and development by suggesting new directions in search facilities, browsing structures and interfaces, and document representation.

## References

[1]   D. Bainbridge, S.J. Cunningham, and J.S. Downie. How people describe their music information needs: a grounded theory analysis of music queries. In *4th International Conference on Music Information Retrieval (ISMIR)*, Baltimore, Maryland, 2003.

[2]   Cunningham, S.J., Bainbridge, D., Falconer, A. More of an art than a science: playlist and mix construction. In *International Conference on Music Information Retrieval (ISMIR '06)*, Vancouver.

[3]   J.S. Downie and S.J. Cunningham. Towards a theory of music information retrieval queries: system design implications. In *3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[4]   J.S. Downie. Music Information Retrieval. *Annual review of information science and technology*, Volume 37, pages 295-340, 2003.

[5]   M. Itoh. Subject search for music: quantitative analysis of access point selection. In *1st Annual International Symposium on Music Information Retrieval*, Amherst MA, 2000.

[6]   P. Knees, M. Schedl, and W. Gerhard. Multiple lyrics alignment: automatic retrieval of song lyrics. In *6th International Conference on Music Information Retrieval (ISMIR '05)*, pages 564-569, London, UK, 2005.

[7]   J.H. Lee, J.S. Downie, and S.J. Cunningham. Challenges in cross-cultural / multilingual music information seeking. In *6th International Conference on Music Information Retrieval (ISMIR '05)*, London, UK, 2005.

[8]   J.R. MCPherson and D. Bainbridge. Usage of the MELDEX digital music library. In *2nd Annual International Symposium on Music Information Retrieval*, Bloomington IN, pages 19-20, 2001.

[9]   M. Muller, F. Kurth, D. Damm, C. Fremery, and M. Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *European Conference on Digital Libraries 2007*, pages 112-123, Berlin, 2007.

[10]  R. Neumayer and A. Rauber. Integration of text and audio features for genre classification in music retrieval. In *29th European Conference on Information Retrieval*, pages 724-727, Rome, Italy, 2007.

[11]  A. Strauss and J. Corbin. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage, 1990.

[12]  D. Tjondronegoro, A. Spink, and B.J. Jansen. Multimedia web searching on a meta-search engine. In *12th Australasian Document Computing Symposium*, pages 80 – 83, Melbourne, Australia, 2007.

[13]  B. Wingenroth, M. Patton, T. DiLauro. In *ACM/IEEE Joint Conference on Digital Libraries '02*, pages 308-309, Portland, Oregon, 2002.