PRELIMINARY ANALYSES OF INFORMATION FEATURES PROVIDED BY USERS FOR IDENTIFYING MUSIC

Jin Ha Lee **J. Stephen Downie** **M. Cameron Jones**
University of Illinois at Urbana-Champaign
Graduate School of Library and Information Science

ABSTRACT

This paper presents preliminary findings based on the analyses of user-provided information features found in 566 queries seeking help in the identification of particular music works or artists. Queries were drawn from the answers.google.com (Google Answers) website. The types and frequency of occurrences of different information features are compared with the results from previous studies of music queries. New feature types have also been developed to obtain a more comprehensive understanding of the kinds of information present in queries including such things as indications of uncertainty, associated use, and the "aboutness" of the underlying musical work. The presence of erroneous information in the queries is also discussed.

1. INTRODUCTION

The lack of empirical data regarding the various aspects of real-life music queries is an obstacle to truly understanding users' music search behaviors in real-life situations. This lack of understanding has negative implications for designing and evaluating music information retrieval (MIR) systems. This is especially so because the common assumptions of MIR researchers regarding the nature of music queries have been found to be remarkably different from the real-world situation [5]. In the past few years, a limited but increasing number of user studies have been conducted in an attempt to address this issue. In particular, previous studies of music information queries (e.g.,[1],[3],[8]) have identified a variety of different types of information features provided by users in queries presented to search intermediaries, and also provided some quantitative data (i.e., frequency counts of need types and feature types in the analyzed queries).

Notwithstanding these studies, there is still a lack of formal understanding regarding the nature and use of the information features present in queries. More specifically, a comprehensive understanding of the associations between the types of needs expressed by users and the information features they provide to help satisfy their needs does not yet exist. In the previous studies, the proportions of queries containing various features were reported at a very broad level, over all different types of queries expressing different needs. The usage statistics for each feature in each type of music query may be significantly different from the overall usage statistics as suggested in studies of other non-textual information searching behaviors [4], but we do not have enough empirical data about this at the present.

This paper presents an overview of the method and select results of a large-scale and ongoing series of analyses designed to provide the MIR community with a stronger empirical and theoretical foundation for the comprehension of real-world MIR queries. In this paper, we will limit our scope by reporting on the empirical data and interesting findings with regard to one specific type of MIR query we have labeled the "Identification" type (defined in section 2.1).

2. DATA COLLECTION AND ANALYSIS

2.1. Query Data Collection

Tague-Sutcliffe's [11] definitions of *query* and *search statement* are adopted in this paper. *Query* is defined as "the verbalized statement of a user's need" and the *search statement* is defined as "a single string, expressed in the language of the system, which triggers a search of the database". Music query is thus defined as the natural language statements in which users express their needs for music objects or information about those objects (i.e., metadata).

Queries are an essential component of IR processes and also an excellent source for collecting information to identify the kinds of information features that are relevant for various user tasks [3]. The particular website selected as the source of real-life query data is the Google Answers website [1], an online reference service provided by Google. Upon receiving the approval from the University of Illinois' Institutional Review Board (IRB), all the queries posted under the music category (2208) were collected in April 2005.

There are a variety of music information needs expressed in the queries we analyzed in the Google Answers data set. These include needs for identifying and/or locating recordings, obtaining lyrics, background information, or recommendations, etc. Of these, the most common type of queries was the ones expressing need for "identification". An "identification query" is one in which users seek help in identifying a *particular* piece of music (typically something that they have heard before), or the name(s) of *particular* artist(s) that they have encountered with regard to some particular

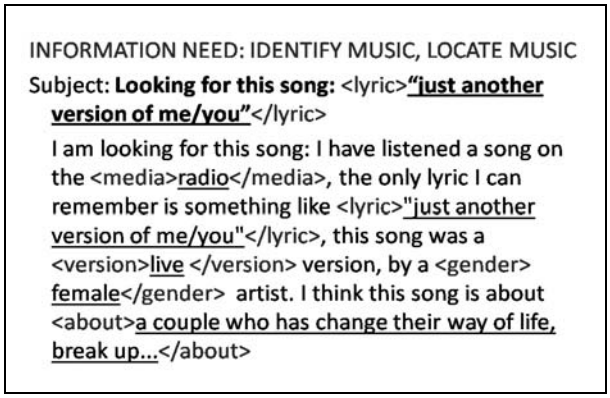
[1] http://answers.google.com/answers/

musical work (sometimes referred to as a "known-item search" in Library Science [9]). In this study we refer to this class of query as "Identification" queries.

Among our data set of 2208 queries, 575 queries were for identifying music and 110 were for identifying artist(s). After filtering out the queries posted by the same inquirer multiple times, 566 unique identification queries were identified and analyzed.

2.2. Data Encoding

Content analysis is the principal method used for analyzing our query data. Content analysis is defined as a "systematic, replicable technique for investigating the manifest and the latent content of any artifact of communication to understand themes and orientations" [7]. This method enables researchers to systematically collect and organize unstructured information into a standardized format that allows one to make inferences about the characteristics of recorded material [7]. Query data were manually encoded. After identifying the needs expressed in each query, all the instances of various information features present in the query text were marked up with tags that identify the type of feature. Such tagging allows for the easy extraction and examination of all instances of the features. Figure 1 shows an example of a marked-up query record.


INFORMATION NEED: IDENTIFY MUSIC, LOCATE MUSIC

Subject: **Looking for this song:** <lyric>**"just another version of me/you"**</lyric>

I am looking for this song: I have listened a song on the <media>radio</media>, the only lyric I can remember is something like <lyric>"just another version of me/you"</lyric>, this song was a <version>live </version> version, by a <gender> female</gender> artist. I think this song is about <about>a couple who has change their way of life, break up...</about>

Figure 1. Example of a marked-up query record

For this encoding process, a standardized set of categories of features was needed as the basis for developing the tag set. The first step was to converge all the categories of needs and features from the previous MIR user studies (i.e.,[1],[3],[8]), and establish an initial set of features (categories) as the pre-coding scheme. The main reason for starting with the categories from previous studies rather than developing them from scratch was to maintain some comparability of the features with the ones used in the previous studies. These features were regarded as tentative and subject to revision based on the further analysis of queries. By an iterative coding process, the features are, ***and continue to be,*** refined to a sufficient level so that they are mutually exclusive, unambiguous, and comprehensive when taken together for expressing the information provided in the music information queries.

3. DISCUSSION

Table 1 compares the information features identified in this study with the ones from the previous study of Google Answers music queries by Bainbridge et al. [1]. The primary difference between this study and [1] is that this study focused only on the information features used in Identification type queries, where Bainbridge et al. collected features from music queries of all types.

Table 1. Comparison of the information features identified in Bainbridge et al. [1] and this study

Features used in queries	% of queries containing the feature	
	[1] N=502	This study N=566
LYRIC	28.9	60.6
DATE	31.9	59.2
GENRE	32.7	35.5
ARTIST NAME	*55.0	19.3
ORCHESTRATION	13.5	**16.8
ABOUT [LYRIC STORY]***	2.6	15.4
WHERE HEARD/PLACE HEARD	24.1	14.7
AFFECT/MOOD	2.4	14.0
WORK TITLE	35.6	13.6
AUDIO/VIDEO EXAMPLE	4.4	10.8
SIMILAR (work/artist)	4.6	9.2
TEMPO	2.4	7.6
NATIONALITY (of music/artist)	12.5	4.2
LANGUAGE	2.0	3.7
COLLECTION TITLE	12.2	2.7
LABEL	5.4	0.1
LINK (to bibliographic info)	2.9	-

* *PEFORMER (47.8%) + COMPOSER (7.2%)*
** *MUSICAL INSTRUMENT (12.8%) + VOCAL (4.0%)*
*** *LYRIC STORY* was used in Bainbridge et al. [1]

From this table, we can observe a notable increase in the use of *LYRIC*, *DATE*, *A/V EXAMPLE* and *AFFECT/MOOD* information in Identification queries. A similar decrease in *ARTIST NAME*, *WORK TITLE*, *NATIONALITY*, and *COLLECTION TITLE* is also present. Note that *ARTIST NAME* and *WORK TITLE* are NOT zero even though these are Identification queries. Many Identification queries contain just one of the two features, or contain guesses (many of which are wrong) about what the title or artist might be (see section 3.4).

3.1. Prevalence of Date/Time Information

Information features related to the time dimension (i.e., date information) were prevalent throughout the analyzed queries. The specificity of this information ranges from vague (e.g., "old", "recent") through a specific decade, date range, to a particular year, and sometimes even to a specific date and time. 199 out of 566 queries (35.2%) mentioned the date when the

sought music was released and 156 queries (27.6%) mentioned the date when the user heard the sought music. The association between date and genre information was also observed. In 8 queries, date was also used to represent a particular music style (e.g., has a "70's feel to it"). All together, 335 out of 566 (59%) of all analyzed queries contained some kind of information on the time dimension.

So why is the date information so frequently used in identifying a particular musical work? One argument can be made regarding the *experience* of music. Experiencing music does seem to capture that particular moment in the users' life and later bring back the memories and feelings "at the time" to the users [2]. Most of these users who want to identify certain music have experienced the sought music in some point of time and space, and the time information seems to work as a major retrieving clue for recalling that memory.

3.2. Other Notable Features

3.2.1. SIMILAR Feature

The *SIMILAR* feature was used to mark up references to known artists or works used to describe attributes of the sought music. Music similarity has been typically associated with playlist generation or music recommendation, but our analysis shows that it is also used for seeking a particular music object. In the analyzed queries, 46 queries (8.1%) contained references to similar artist(s) whereas only 7 queries (1.2%) referenced similar musical work(s).

3.2.2. AFFECT/MOOD Feature

The *AFFECT/MOOD* feature was used in 79 queries for identifying music (14.0%), a considerably larger proportion of queries than 2.4% reported over all types of queries in [1]. Some examples of the moods that our study has found include: "lovely" (3), "mellow" (3), "fun(ny)" (3), "funky" (3), "dreamy" (2), "hypnotic" (2), "happy" (2), "sad" (2), "quirky" (2). This finding suggests that mood information is an important "hook" in the minds of the searchers when they cannot remember the names of the artist or title of work.

3.2.3. ABOUT Feature *and LYRIC STORY* Feature

In the initial coding process for this study, the *ABOUT* feature was used to mark up general descriptions where the user tried to explain what the "subject" of the music was "about". Most of these descriptions turned out to be "lyric stories" which was one of the features included in [1]. However, some users were describing what the related music video was "about" or what the album cover looked like. Others were describing an overall "theme" of the music rather than the "lyric story", for example, "Christmas", "Halloween", "love", "war", "jealousy", and so on.

3.3. New Information Features

Table 2 presents the information features that were not listed in the previous study by Bainbridge et al. [1]. Among these features, the most common one was the *MEDIA* feature (44.0%) used for part of text where the user describes the media from which he/she heard the sought music. Various expressions of certainty or uncertainty as to the accuracy of the user-provided information were used in conjunction with other features (30.7%). Other common features include: the *LYRIC DESCRIPTION* feature (30.0%) referring to part of text where the user provides additional explanation of the lyric fragments he/she is providing such as the type (e.g., refrain, chorus), part (e.g., at the beginning, end of the song), and frequency of occurrence (e.g., repeated *X* times), and *ASSOCIATED USE* (30.0%) for the part describing the use of music in other work(s) including movies, commercials, video games, and so on.

Table 2. New information features identified in queries for identifying music

Features used in queries	% of queries containing the feature
MEDIA (e.g., radio, TV)	44.0
EXPRESSION OF (UN)CERTAINTY	30.7
LYRIC DESCRIPTION	30.0
ASSOCIATED USE (e.g., movie, ad)	30.0
GENDER OF ARTIST	20.5
MUSICAL STYLE DESCRIPTION	19.8
RELATED WORK TITLE	15.9
PRIOR SEARCH INFORMATION	13.4
SCENE (where music was used)	13.3
RELATED EVENT	4.2
REGION	2.6
VERSION INFORMATION	1.9
MELODY DESCRIPTION	0.7

3.4. Providing Incorrect Information

In music queries, consistently misheard lyrics, misinterpretations of genre, misunderstood artist names, and inexact released dates are rather common. However, it is interesting that, notwithstanding these errors, users, at least in some cases, are still able to obtain correct identifications of their sought-after music [10]. In our Google Answers data set, we can point to a number of examples where the user was still able to find the correct music although the user's information was incorrect like the ones that follow:

Query (#501789):

I only caught some of the lyrics.**<other>** I only heard it once**</other>**. **<uncertainty>**I am not sure that the wording is absolutely correct **</uncertainty>**. IT was a **<gender>**female **</gender>** singer, very **<style>**plain **</style>**, **<affect>**beautiful**</affect>**,

<tempo>slow**</tempo>**. **<lyric>**"there will be no black flag on my door"**</lyric>** **<lyric>**"I am in love"**</lyric>** **<lyric>**"I will go down with vengence"**</lyric>**

Answer:

...I'm pretty certain that the song you heard was Dido's "White Flag". A snippet of the lyrics are:

"I will go down with this ship And I won't put my hands up and surrender There will be no white flag above my door I'm in love and always will be"

...Search Strategy (on Google):

"be no black flag" dido "white flag" lyrics...

Query (#342762):

Looking for a song. They [sic] lyrics go **<lyric>**"My Mamma done told me, When I was in knee-socks"**</lyric>** It's a **<genre>**jazzy**</genre>** number.

Answer:

…Well, you were close. The lyrics refer to "knee pants" not "knee socks" and the genre is blues rather than jazz….

For human intermediaries, the incorrect information present seems to be easily overcome. However, within an automated system, such errors can be catastrophic. This is one reason why we are noting expressions of *UNCERTAINTY* via our <uncertainty> tag to see if consistent clues can be derived to signal to an automated system that they should deal with the contained information in some special way (perhaps using some type of re-weighting approach) to improve search robustness.

4. CONCLUSION AND FUTURE RESEARCH

Our preliminary results show that the Identification type queries display different feature distributions than those found in previous, more general studies. For future research, we will continue to explore the differences in search behaviors for the queries expressing other needs. We will also compare these features to the ones used in queries for other non-textual cultural objects as certain similarities have come to our attention, especially with regard to image searching. A number of interpretive attributes including abstract concept (e.g., *ABOUT*), external relation (e.g., *SIMILARITY*) and reactive attribute (e.g., *UNCERTAINTY*) observed here were also found in Jörgensen's study [6] of image descriptions.

5. ACKNOWLEDGMENTS

This work supported by The Mellon Foundation, and the National Science Foundation (NSF IIS-0327371).

6. REFERENCES

[1] Bainbridge, D., Cunningham, S. J., & Downie, J. S. "How people describe their music information needs: A grounded theory analysis of music queries", *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR),* Baltimore, USA, 2003.

[2] Cunningham, S. J., Bainbridge, D., & Falconer, A. "More of an art than a science: Supporting the creation of playlists and mixes", *Proceedings of the 7th International Conference on Music Information Retrieval,* Barcelona, Spain, 2004.

[3] Downie, J. S., & Cunningham, S. J. "Toward a theory of music information retrieval queries: System design implications", *Proceedings of the 3rd International Conference on Music Information Retrieval,* Paris, France, 2002.

[4] Fidel, R. "The image retrieval task: Implications for the design and evaluation of image databases", *The New Review of Hypermedia and Multimedia*, *3,* 181-199, 1997.

[5] Futrelle, J. & Downie, J. S. "Interdisciplinary communities and research issues in music information retrieval", *Proceedings of the 3rd International Conference on Music Information Retrieval,* Paris, France, 2002.

[6] Jörgensen, C. "Attributes of images in describing tasks", *Information Processing and Management, 34*(2/3), 161-174, 1998.

[7] Krippendorff, K. *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage, 2004.

[8] Lee, J. H., Downie, J. S., & Cunningham, S. J. "Challenges in cross-cultural/multilingual music information seeking", *Proceedings of the 6th International Conference on Music Information Retrieval,* London, UK, 2005.

[9] Lee, J. H., Renear, A., & Smith, L. C. "Known-item searching: Variations on a concept", *Proceedings of the 69th ASIS&T Annual Meeting*. Austin, USA, 2006.

[10] Lee, J. H., & Renear, A. "How incorrect information delivers correct search results: A pragmatic analysis of queries", *Proceedings of the 70th ASIS&T Annual Meeting.* Milwaukee, USA, 2007.

[11] Tague-Sutcliffe, J. "The pragmatics of information retrieval experimentation, revisited", *Information Processing and Management, 28*(4), 467-490, 1992.