Q&A Websites: Rich Research Resources for Contextualizing Information Retrieval Behaviors

Jin Ha Lee
The Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, IL, 61820, USA
217.333.3280
jinlee1@uiuc.edu

Xiao Hu
The Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, IL, 61820, USA
217.333.3280
xiaohu@uiuc.edu

J. Stephen Downie
The Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, IL, 61820, USA
217.333.3280
jdownie@uiuc.edu

ABSTRACT
In this paper, we propose a framework to exploit user-generated context metadata from Questions and Answers (Q&A) websites where users provide natural language descriptions of information objects they aim to identify. Users' descriptions from Q&A websites are good sources of obtaining metadata regarding the real-life search context of real users in their information retrieval (IR) processes. We describe our suggested framework and how it may help address some known difficulties in exploiting user-generated metadata and metadata schemes for Information Retrieval in Context (IRiX).

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Retrieval models*

General Terms
Design, Experimentation

Keywords
Context metadata, User-oriented metadata scheme

1. INTRODUCTION

1.1 User Context in Information Retrieval
Belkin et al. [2] suggested a new approach to studying information retrieval (IR) based on a radically different premise, one which declared that "a fundamental element in an IR situation is the development of an information need out of an inadequate state of knowledge." In most approaches to complementing the IR process with user-related information, the information available about the user is often limited to that inferred from direct interactions between users and the system (e.g., transaction logs). What is missing in most cases is the capturing and exploitation of crucial contextualizing information such as user motivations and background knowledge [15]. As Belkin et al. [2] suggested, "for IR to be successful… information need must be represented in terms appropriate for just that task, with the remaining elements of the system (i.e., document representation, retrieval mechanism) being represented or constructed on the basis of that representation." In the following, we lay out an exploratory framework to represent the 'anomalous state of knowledge' (ASK) of IR system users by capturing and exploiting information about users' *real-life* IR processes, and integrating this information within a formal music representation system. While our framework development is in the domain of music information retrieval (MIR), the principles and techniques discussed here have broad application to all IR system interactions.

Recent user studies in MIR suggest the importance of exploiting context metadata in addition to content metadata for improving system success. Table 1 shows the types of content and context metadata in the MIR domain based on the framework suggested by Lee and Downie [12].

Table 1. Metadata Framework for Music Objects [7][12]

Content	**Musical**	Data derived directly from the music itself (e.g., melody, tempo, etc.)
	Textual	Text appearing along with the music (e.g., lyrics, etc.)
	Bibliographic	Traditional metadata that describes an object (e.g., title, author, etc.) and what it is *about* (e.g., plot, story of the lyrics, etc.)
Context	**Relational**	Data about the item's relationships (artificially created or socially constructed) with other music related items (e.g., genre, indications of similarity, etc.)
	Associative	Data indicating associated use with other works, media or events (e.g., use in TV, movies or commercials, use at special events, etc.)
	Perceptual	Description related to how the music object is perceived (e.g., mood, affect, rating, etc.)

Content metadata is defined as information that is intrinsic to an object, (i.e., can be derived from the music itself or the object embodying the music). Context metadata, however, describes the

extrinsic aspects (i.e., how the music is perceived and used among a group of users). The large-scale music seeking survey conducted by Lee and Downie [12] showed that people value music recommendations and reviews from others, and that they seek music by its associated use in other multimedia objects such as movies, commercials, etc. The necessity for providing access points that link music with external objects or events has already been noted in previous user studies [1][4][11]. This claim is also supported by [13] where Lee et al. discussed the challenges of users seeking music objects originating from outside of their culture and/or language – users in their study often exploited associative context metadata for their searches. To improve the user's experience of MIR, new kinds of metadata that better contextualize searches must be developed in addition to traditional bibliographic metadata [12]. Context metadata is dependent not just on the situation of an individual user, but also on the sociocultural situation of the group to which the user belongs. The heart of our framework is the exploitation of this sociocultural context of user groups.

1.2 Dynamic Nature of Information Objects

Although many IR researchers perceive the necessity for exploiting context metadata to better assist users [4], the development of an operational MIR system that actually exploits context metadata has been hindered by pragmatic considerations based upon the evolving identity of information objects. The identity of an information object develops in a dynamic way – it is constantly modified through its life [7]. The development of an identity starts with a set of attributes that are determined at the creation of an object (e.g., the main entities in bibliographic records). As the information object moves through its life, additional sets of attributes are obtained. For instance, new metadata can be generated based on how the information object is perceived and used by people in a particular culture. For music objects, the associative use of music in multimedia objects including movies or TV commercials, or similarity-based groupings of performers can be examples of such metadata [12]. However, formal metadata creation rarely happens beyond the initial, traditional bibliographic stage for it is very labor-intensive to add new metadata to information objects. Context metadata is rarely formally created because the information upon which it is based comes into being *after* the object has been used and has interacted within a dizzying array of contexts.

1.3 Exploiting User-Generated Metadata and Metadata Schemes

Systems have been proposed to exploit user-generated metadata and metadata schemes to overcome the aforementioned difficulties with creating context metadata. The ideas of creating decentralized ontologies [3] and of employing subject experts/enthusiasts and users for generating metadata [5] have become more pervasive. Collaborative annotation tools such as del.icio.us (a social bookmarks manager) and Flickr (an online photo sharing system) are examples of such tools [3]. Despite the initial enthusiastic reactions toward those systems, the limitations of such metadata schemes soon became evident. Metadata content and schemes in these systems suffered from a lack of vocabulary control and standardized structure. Disparate, irrelevant, incomplete and inaccurate user vocabularies resulted in metadata content with a high noise ratio, making it problematic for IR [3]. Part of the problem lies in neglecting the distinction between merely describing an object and describing the object *in order to find it*. Idiosyncratic and uncommon descriptions may be interesting, but are much less useful in assisting other users in their information seeking. Another critical issue is the difficulty in offering incentives for people to continue providing manual descriptions of information objects.

2. OUR CONTEXT-BASED FRAMEWORK

Our framework attempts to minimize the problems of user-generated metadata and metadata schemes in multiple ways. The underlying idea of the framework is that information seeking is a socially situated action. There may be unlimited ways to search for an information object, but cultural and social contexts affect the ways in which users search for information, leading users to prefer certain ways over others repeatedly. If an information object is identified based on one user's description, this description may contain useful clues for other users who are seeking the same object.

Locating meaningful sources for the extraction of context metadata information is the most important task of our project. We deem Q&A websites to be one of the best sources for the following reasons. First, the questions from Q&A websites contain descriptions of information objects in the terms used by real users in their real-life information searches. Second, we assume that the user's description in the question will generally be a sincere and best description of the sought object (i.e., people would not *intentionally* provide incorrect information about the sought object).

We have collected over 2000 authentic questions on music information objects from Naver.com (a popular Korean Web search portal hosting one of the largest Q&A systems), and Google Answers (an online reference system in a Q&A format). As of May 13, 2005, there were 319,592 questions posted and answered under the music category on Naver.com, and 2,235 questions under the music category of Google Answers. Rather than collecting any casual descriptions given about a particular object, we are focusing only on the descriptions given for the search purpose. We believe that these descriptions reflect the most realistic context of information searching and contain information that will better assist the next user's search task.

We began to classify the collected questions based on the categories suggested in [14]: directional, holdings, ready reference, exact reproduction, description, readers advisory, bibliographic instruction, research, citation list, analysis, and critique. We have been analyzing user descriptions of music objects in the questions by taking a grounded theory approach in order to identify the most influential and commonly used features for receiving successful answers in each category. Based on our findings, we are designing a formalized metadata scheme to collect these features in a systematic and efficient way.

Figures 1 and 2 provide a general view of our framework. There are two main stages –preparation and implementation. The preparation stage consists of multiple steps explained as follows:

1. **Data Collection**: Questions posed by real users on various Q&A Websites are collected.

2. **Classification of Questions**: All questions are classified into categories based on their functions (e.g., identifying an object, locating an object, etc.).

3. **Identification of Features in User Descriptions**: The features that are commonly used and have higher impact in obtaining a successful answer are identified by an analysis.

4. **Formalization of Metadata Scheme**: Based on the result of feature analysis, we generate a formalized metadata scheme for each category of questions. In Figure 2, the graphic representations of the schemes are depicted arbitrarily, as we cannot predict the actual schemes before completing the analysis of the questions.

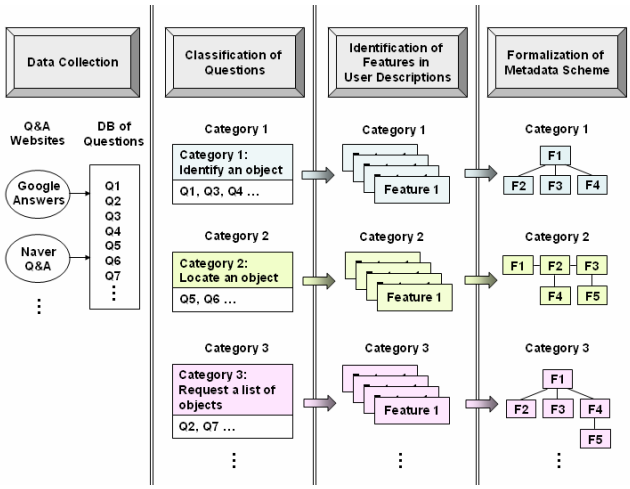

Figure 1. Four Steps of the Framework Preparation Stage

The second stage is illustrated in Figure 2. Here, we only present the "identify object" category of questions (i.e., a kind of known-item search), but the general idea of our framework can be deduced in its application to other question categories.

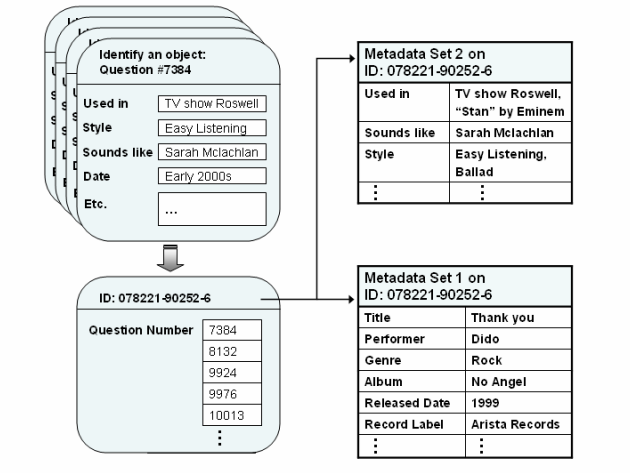

Figure 2. The Implementation Stage

The preliminary analysis of the questions in the "identify object" category revealed the following types of metadata as dominant types: words in lyrics, date, name of performer, genre/style description, associated use, instrument and mood. Based on the formalized metadata scheme, we are proposing to develop a structured interface to help users better formulate their questions. When the question is answered, the answerer creates a pointer to the sought object in the already established catalog of music objects. This way, all the user descriptions related to the same object can be collocated. Only the questions that received relevant answers, as verified by the inquirer's rating, are used. We assume that the descriptions in questions that did not receive any answers from other people, or received poorly rated answers are less likely to be useful. We recognize that there do exist cases where this is not true. However, we assume that these cases will be less common and by using only the cases where the relevancy of the answer was verified by the searcher, we hope to generate a relatively solid set of metadata. Through this process, a secondary set of context metadata is generated in addition to the primary set of metadata records that already resides in the database.

In our framework, rich user-generated descriptions provided in the IR processes address the sparseness of metadata. We use authentic questions given by real users as the data source, and therefore the metadata extracted is exactly the kind of information that people use for their IR tasks. Moreover, unlike the one-time creation of "formal" metadata, users' metadata constantly evolves and reflects a temporal dimension of context [10]. The situation of our framework between two opposing approaches, manual annotation and automatic extraction of metadata, is also notable. By establishing a *user-oriented structure* for metadata collection, we hope to minimize the problems of other approaches; mainly the time-consuming and costly nature of manual metadata creation and the unsatisfying and untested results for large sample sizes in automatic extraction [6].

A major motivating factor for developing this kind of framework is to overcome the limitations of current natural language processing (NLP) technologies by collecting more structured descriptions. One might argue that forcing users to use a structured interface for inputting their queries rather than allowing them to describe their needs in natural language is moving backwards from user convenience. However, we argue that providing a structure to better formulate their queries will eventually benefit users by increasing the likelihood of receiving relevant answers as studies have shown that the NLP technologies work better with structured information. We believe our framework will contribute to more efficient and successful applications of NLP in processing user descriptions [8]. Also note that the structure we suggest will be based on how users *already* describe music objects in their real-life music information seeking, not an arbitrary structure, or one determined by the constraints of the system or data.

3. CONCLUSIONS AND FUTURE WORK

The most fundamental aspect of our framework is the dual representation of information objects. The system contains two sets of metadata: First is the already established authoritative set of metadata in the database containing the relatively stable and definite attributes (e.g., bibliographic metadata such as creator and title). Second is the user-oriented set of context metadata that describes dynamic attributes that are highly context-dependent, such as the use of an information object, the similarity of a certain object to other objects in a given culture, and even commonly

made mistakes (i.e., incorrect information) of various user groups. Such polyrepresentations, from different cognitive representations of both users' information situations and the sought-after objects, can improve the performance of IR systems [9].

In our future work, we will extend our study to other categories of questions beyond the "identify object" class. In terms of evaluating the eventual success of our framework, we propose a comparative examination of a set of search results of real user-generated questions with the searches being supported/not supported with framework-derived context metadata.

4. ACKNOWLEDGMENTS

We thank the Andrew W. Mellon Foundation for their moral and financial support. This project is also supported by the National Science Foundation (NSF) under Grant Nos. NSF IIS-0340597 and NSF IIS-0327371.

5. REFERENCES

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

[2] Belkin, N. J., Oddy, R. N., and Brooks, H. M. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation, 38*, 2, 1982, 61-71.

[3] Cayzer, S. Semantic blogging and decentralized knowledge management. *Communications of the ACM, 47*, 12, 2004, 47-52.

[4] Downie, J. S., and Cunningham, S. J. Toward a theory of music information retrieval queries: system design implications. In *Proceedings of the ISMIR 2002*, Paris, 2002.

[5] Greenberg, J. Metadata and the World Wide Web. In *The encyclopedia of library and information science (Vol. 72).* New York: Marcel Dekker, 2002, 244-261.

[6] Greenberg, J. Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging, 6*, 4, 2004, 59-82.

[7] Gilliland-Swetland, A. Setting the Stage. In: Murtha Baca ed: *Introduction to Metadata: Pathways to Digital Information*. Getty Information Institute, 2000.

[8] Hu, X., Downie, J. S., West, K., and Ehmann, A. Mining music reviews: promising preliminary results. In (accepted) *Proceedings of the ISMIR 2005*, London, 2005.

[9] Ingwersen, P. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the SIGIR 1994,* 101-110.

[10] Ingwersen, P. and Järvelin, K. Information retrieval in contexts. *SIGIR 2004 IRiX Workshop,* Sheffield, UK, 29th July 2004.

[11] Kim, J. and Belkin, N. J. Categories of Music Description and Search Terms and Phrases Used by Non-Music Experts. In *Proceedings of the ISMIR 2002*, Paris, 2002.

[12] Lee, J. H. and Downie, J. S. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In *Proceedings of the ISMIR 2004*, Barcelona, 2004.

[13] Lee, J. H., Downie, J. S., and Cunningham, S. J. Challenges in cross-cultural/multilingual music information seeking. In (accepted) *Proceedings of the ISMIR 2005*, London, 2005.

[14] Pomerantz, J. A linguistic analysis of question taxonomies. *Journal of the American Society for Information Science and Technology, 56*, 7, 2005, 715-728.

[15] Ruthven, I. "and this set of words represents the user's context…" *SIGIR 2004 IRiX Workshop*, Sheffield, UK, 29th July 2004.