

Evaluation of Music Information Retrieval: Towards a User-Centered Approach

Xiao Hu

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, IL, 61801, U.S.A.
xiaohu@illinois.edu

Jingjing Liu

School of Communication and Information
Rutgers, The State University of New Jersey
4 Huntington St.
New Brunswick, NJ 08901, U.S.A.
jingjing@eden.rutgers.edu

ABSTRACT

With the dramatic increase of online digital music, research on Music Information Retrieval (MIR) is flourishing more than ever. However, evaluation in MIR has been focused on system-centered approaches, where systems are evaluated against a pre-built ground truth dataset using system-focused measurements, and little attention has been spent on user experience. In this paper, we argue that MIR evaluation should take users, in addition to systems, into consideration. We suggest that some measures and models in the established area of Interactive IR in the text domain can be applied to the MIR domain. Novel evaluation measures that are unique to MIR are also proposed. The purpose of this paper is to encourage user-oriented, and thus more comprehensive, approaches to evaluating MIR systems.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *search process*, J.5 [Art and Humanities] – *Music*

General Terms

Measurement, Performance, Human Factors.

Keywords

evaluation, music information retrieval, user-centered evaluation, usefulness.

1. INTRODUCTION

As a crucial aspect of system development, evaluation of Information Retrieval (IR) systems has attracted continuous attention. Much of this effort can be seen from the annual TREC (Text REtrieval Conference) and the frequent appearance of workshops on IR evaluation at the annual conference of SIGIR. In the text IR area, some evaluation criteria and measurements beyond “relevance” have been proposed and used, and these alternative approaches have addressed many aspects of IR

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HCIR '10, August 22, 2010, New Brunswick, NJ, U.S.A..
Copyright 2010 ACM 1-58113-000-0/00/0004...\$5.00.

evaluation criteria, including user experiences in the search process, in addition to the well-accepted, system-focused measures such as *precision* and *recall*.

In recent years, with the popularity of digital music, research on Music Information Retrieval (MIR) is flourishing more than ever. The International Society of Music Information Retrieval (ISMIR) will have its 11th annual conference this year. In the last decade, a number of new algorithms and systems have been proposed and developed for a variety of MIR-related tasks: genre classification, artist clustering, music recommendation, playlist generation, etc. While these innovations greatly advance the state of the art of MIR, and some of the systems have been turned into real-world applications, the evaluation of MIR algorithms and systems is not as developed: the current evaluation paradigm is dominated by system-oriented approaches, while users, whom MIR systems serve, have rarely been considered in MIR evaluation frameworks.

During IR processes, users accomplish their tasks by interacting with IR systems. Hence, evaluations of IR systems need to take into account users' interactive processes of information searching and retrieval [2][3][18]. Just as other IR systems, MIR systems do not stand by themselves. The goal of MIR systems is to facilitate users' music information tasks, and thus the evaluation of MIR should inevitably take users into consideration. In this paper, we suggest that along with the maturity of system-centered MIR evaluation, it is needed to bring users into the picture. User-centered evaluation of MIR is grounded in the nature of music information seeking and as such has broad applicability in the evaluation of MIR.

2. RELATED WORK IN THE MIR DOMAIN

2.1 MIR Evaluation

Evaluation of MIR has been dominated by system-centered approaches. Since 2004, the annual event, Music Information Retrieval Evaluation eXchange (MIREX) [13] has become the main venue of system evaluation in the MIR community. MIREX is the counterpart of TREC in the music domain. Just like TREC, there is a variety of MIR *tasks* included in each year's MIREX, such as genre classification, mood classification, cover song identification, audio music similarity and retrieval, melody extraction, etc. For each task, systems developed by participatory research groups around the world are run against pre-built test collections and their performances are compared. The measures used in MIREX are all system-centered, including *accuracy* for classification tasks, *average precision* for retrieval tasks,

variations of *precision/recall* for (key, onset) detection tasks, etc. Unlike TREC, as of the year 2010, there has not been a task in MIREX that considers users' interactions with the systems.

MIR experiments outside MIREX are also primarily evaluated by system-centered approaches, that is, without involving users. Nevertheless, there are a few exceptions. Pauws, S. and Eggen [22] conducted a controlled user experiment to evaluate the quality of playlists generated by the algorithm that the authors proposed. They recruited twenty-two participants, each of whom used the proposed interactive system as well as a control system to generate playlists for two pre-defined situations ("soft music" and "lively music") over four experimental sessions. The researchers then compared the systems using participants' ratings on the resultant playlists. A post-experiment interview was conducted to elicit supplementary findings on the *perceived usefulness* of automatic music compilation. In a similar study, Pauws and van de Wijdeven [23] evaluated their "SatisFly" playlist generation system by conducting a user experiment with twenty-four participants. Each participant rated the playlists generated by the "SatisFly" system and a control system. The measures used in this study included *playlist quality* as calculated by users' ratings, *time spent on the task*, *number of button presses in accomplishing the task*, as well as *perceived usefulness* and *ease-of-use*. On the same line of research, a conclusive user evaluation was conducted in [28] to assess the "similar song" function of the E-Mu jukebox, a music recommendation system. Twenty-two participants used the test system as well as two control systems to perform a playlist generation task where 10 different songs must be chosen for a single imaginary music listening situation. In addition to the measures calculated in [23], this study also measured *the order of participants' preference* among the three systems by asking which system they liked most and least.

These three studies conducted user experiments in evaluation and addressed aspects that by no means can be covered in system-centered approaches. The user-dependent measures are essential for understanding users' experience of MIR systems and for achieving the ultimate goal of MIR systems. Interestingly, these three studies were all on the task of playlist generation, with similar design and scale, were all conducted in the same research lab, Philips Research Laboratories, and were conducted at least five years ago. This indicates that user-oriented evaluation has not been well adopted and has limited influence in MIR.

2.2 User Studies in MIR

Despite of the sparseness of user-centered evaluation, MIR researchers have long paid attention to users. User studies in MIR have primarily focused on identifying users' music information needs and the features users often employed to describe their needs. For example, McPherson and Bainbridge [21] analyzed server logs of the MELDEX digital music library and discovered its usage patterns. Itoh [17] surveyed 21,177 online catalog search logs in an academic music library and identified access points of music scores in online environments. Researchers also analyzed forum postings and requests on Q/A sites (e.g., Google Answers) to identify user needs and information features used in musical queries [1][11][14][19]. Besides these approaches, ethnographic methods (e.g., interviews) and surveys were often used in exploring users' music information seeking behaviors [12][20]. In a study using multiple user study methods, Vignoli [27]

conducted interviews, user experiments with existing products, as well as online surveys to investigate how music listeners organize and access their digital music collection.

3. USER-CENTERED EVALUATION IN TEXT IR

In the area of text IR, for the past over three decades, *relevance* [24] has been a major criterion of evaluation and has been overwhelmingly used in TREC practice. However, in recent years, researchers have been arguing that relevance is not sufficient in evaluating IR systems because evaluation studies are routinely pursuing information seeking tasks outside of the traditional, so-called Cranfield paradigm and are taking a broader view of tasks, users, and contexts [15][18]. Relevance-based measures such as *precision* and *recall* are not good for evaluating Interactive Information Retrieval (IIR) systems, because while users may modify or develop search tasks during search processes, the two measures cannot quantify the "informativeness" of interactions [7]. In addition, neither *precision* nor *recall* is a highly significant factor of *user satisfaction* towards a given retrieval system. Depending on their information seeking tasks, users may not be concerned about retrieving all the documents relevant to their search tasks; for many users, they are happy if they can get a good answer in a short amount of time [16]. Further, the purpose of an IR system is to help users accomplish a task, and therefore IR system evaluation should consider both task success as an outcome and the value of support that IR systems provide over the entire information seeking episode as a process [9]. Relevance-based measurements that only focus on topical matches between the documents and the query terms fail to address these requirements.

There have been alternatives to *relevance*, such as *efficiency*, *satisfaction* [26], and *utility* [5][9], to name just a few. Kelly [18] pointed out that very complex IIR activities involving both Behavioral Science and Computer Science require pluralistic approaches and methods, and that "a single, prescribed model would be deleterious" (p. 202). The evaluation methods and measures to be used depend, to a large extent, upon the goal of the evaluation. For example, the evaluation may be used for a system that is able to retrieve relevant documents, or for a system that can help its users accomplish specific tasks.

From a phenomenological perspective and based on the nature of information seeking, Belkin and colleagues [4][9] have recently suggested *usefulness* as a criterion for evaluation of IIR systems. *Usefulness*, as they argues, can be used to evaluate system support from the aspects of both task outcome and task process in the accomplishment of a task. In terms of the measurements under the usefulness framework, *usefulness* itself can be a measurement of how much the search results contribute to task accomplishment. In addition, other measures can include but are not limited to: *task accomplishment* (how well the user finishes the task; how many steps the user goes through; how long it takes) and *support of the system to the information seeking goal* (to the general task and to each possible sub-task; acceptance or rejection to system suggested search strategies and/or query reformulation; usefulness and the use of retrieved documents; recall and precision of single search; etc.). It should be noted that while *usefulness* is an alternative to *relevance*, the authors did not mean to disregard relevance and its measures; instead, relevance measures are part of the usefulness framework. While some of these measures have

been used in previous studies, the usefulness framework is not a simple repetition of previous efforts. It suggests that determination of which measures to use in evaluation depends crucially upon the specification of a leading task or goal whose accomplishment itself can be measured. This multiple-measure approach in general echoes Kelly's notion [18] that there is not a single best method in IIR evaluation.

4. PROPOSED MEASURES IN USER-CENTERED MIR EVALUATION

As previously mentioned, while MIR evaluation has been dominated by system-centered approaches which typically measure how well the systems (algorithms) classify music and how relevant their retrieved music was, rare effort has been spent on measuring how well the systems support users' completion of music information seeking tasks or what users' search experiences are like. We believe that the concept of *usefulness* introduced in section 3 could be well applied to MIR system evaluation. We suggest that the measurements of the usefulness framework are applicable to the MIR domain. Further, due to the unique entertaining nature of music that is different from textual information in regular text retrieval, we also suggest additional aspects that are not included in the original usefulness framework proposal, such as *entertainability* and *social life support*. In addition, other IIR measurements such as *learnability* are also important in evaluating MIR systems. Our proposed measures are as follows.

Measure on music information task accomplishment. This measure as Cole et al. [9] proposed in their IIR evaluation framework focuses on search outcome. Since the goal of MIR activities is to support users to fulfill their music information needs represented by their MIR tasks, MIR evaluation can include the measure on task accomplishment in the evaluation as well. This measure can include several points:

- (How) does the user find the desired music?
- How many steps does the user take to find the music?
- How long does the user spend in finding the music?

This is often referred as *task completion time* or *time lag* and has long been recognized as one of the most important criteria that could be used to evaluate IR systems [8][18][26].

Measure on system support of the music information seeking process. Information seeking is a process instead of just a search outcome [9]. It is often a continuous process instead of a single query-result activity. This is true in the music domain as well. For example, during a music search process, users may modify and refine their queries after listening to part of a retrieved song. Specific points of this measure can include:

- How does the system support identification of and sequence of sub-tasks (if any) toward the completion of the general music information seeking task?
- How useful is the system in querying support (e.g., query reformulation suggestions)?
- How does the system support displaying and playing search results?
- How does the system support saving search results?

The last two points are very different from text IR. Displaying search results in the music domain is an on-going research question in and of itself. Unlike a text document, a music document (or part of it) has to be played and listened to by the user before he or she can make judgments on its value to the music information needs. Also, due to intellectual property laws, saving a music document is much more complicated than saving a text document, which raises an additional challenge to MIR systems.

Measure on system support of user experience. This aspect includes general rules suggested by experts from Human-Computer Interaction (HCI) and IR areas, as well as our suggested measures that are unique to MIR systems.

- How is the user satisfied with his music finding experience?
- How easy does the user feel the system is to use?
- How much does the system help the user to avoid confusion or "getting lost"?

These points are frequently seen in system interface evaluation, for example in [25].

- How is the system easy to learn?
- How quickly can a user familiarize himself/herself with the system again after not using it for a while?

The above two points are on the *learnability* of MIR systems. Researchers have realized the influence of learning in the IR process. For example, Borlund [6] argued that IR evaluation should pay attention to the learning process of users and gauge the *learnability* of an IR system. The issue of *learnability* is even more important in MIR because MIR systems often include multimedia interfaces that are novel and unfamiliar to most users. For instance, McPherson and Bainbridge [21] studied usage patterns of the MELDEX digital music library and found that although the system supported melodic querying, users still preferred issuing textual queries. The overhead of learning how to issue a melodic query had impeded users from taking full advantage of the system.

- How well does the system support entertainment?
- How well does the system support social life?

The above two points are unique to the MIR domain. Usually, people search music for entertainment purposes, and thus the MIR systems should attempt to support users' general purpose of entertainment. In addition, music is also an aspect of social life. Users often share their favorite music with friends and carefully select certain kinds of music for specific social events and occasions such as weddings, parties, trips or romantic dates. Therefore, it should be a goal of mature MIR systems to support users' music-related social life.

The above proposed measures focus on different aspects of users' interactions with MIR systems. It is desirable to combine multiple measures in evaluation, but the adoption of any of these measures would complement current system-centered approaches.

5. CONCLUSIONS

As the ultimate goal of MIR systems is to help users in seeking music information, the evaluation of MIR systems should take users into consideration. This paper advocates a paradigm shift from system-centered evaluation to user-centered evaluation in

MIR. Interactive IR in the text domain has been an active research area for decades and has plenty to offer to MIR. By applying IIR and HCI evaluation measures to MIR and proposing measures unique to the music domain, this paper aims to elicit more work and attention on user-centered MIR evaluation.

6. REFERENCES

- [1] Bainbridge, D., Cunningham, S.J., and Downie, J.S. 2003. 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).
- [2] Beaulieu, M., Robertson, S. and Rasmussen, E. 1996. Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1): 85-94.
- [3] Belkin, N.J. 2008. Some(what) grand challenges for information retrieval. *ACM SIGIR. Forum Archive*, 42 (1): 47-54.
- [4] Belkin, N., Cole, M., and Liu, J. 2009. A model for evaluation of interactive information retrieval. In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, pp.7-8.
- [5] Belkin, N.J. and Vickery, A. 1985. *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-Based Systems*. Library and Information Research Report 35: The British Library, University Press, Cambridge.
- [6] Borlund, P. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), Paper No.152.
- [7] Borlund, P. and Ingwersen, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3): 225-250.
- [8] Cleverdon, C. W. and Keen, E.M. 1966. *Factors determining the performance of indexing systems*. Vol. 1: Design. Vol 2: Results. Cranfield, U.K: Aslib Cranfield Research Project.
- [9] Cole, M., Liu, J., Belkin, N.J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., and Zhang, X. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. In Proceedings of the 3rd Workshop on Human-Computer Interaction and Information Retrieval (HCIR) (pp. 1-4).
- [10] Cooper, W.S. 1973. On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation," *Journal of the American Society for Information Science*, 24, 87-100.
- [11] Cunningham, S. J., Bainbridge, D. and Falconer, A. 2006. "More of an art than a science": supporting the creation of playlists and mixes. In Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR).
- [12] Cunningham, S.J., Jones, M., and Jones, S. 2004. Organizing digital music for use: an examination of personal music collections. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR), pp. 447-454.
- [13] Downie, J. S. 2008. The Music Information Retrieval Evaluation eXchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4), 247-255.
- [14] Downie, J. S. and Cunningham, S. J. 2002. Toward a theory of music information retrieval queries: System design implications. In Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR).
- [15] Geva, S., Kamps, J., Peters, C., Sakai, T., Trotman, A., and Voorhees, E. (Eds.). 2009. Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. Workshop held at SIGIR 2009, Boston, MA, July 23, 2009.
- [16] Hearst, M. 2009. *Search User Interfaces*. Cambridge University Press.
- [17] Itoh, M. 2000. Subject search for music: quantitative analysis of access point selection. In Proceedings of the 1st International Conference on Music Information Retrieval (ISMIR).
- [18] Kelly, D. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2): 1-224.
- [19] Lee, J. H., 2010. Analysis of user needs and information features in natural language queries seeking music information. *Journal of the American Society for Information Science and Technology*, 61 (5): 1025-1045.
- [20] Lee, J. H. and Downie, J. S. 2004. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR), pp. 441-446.
- [21] McPherson, J.R., and Bainbridge, D. 2001. Usage of the MELDEX digital music library. In Proceedings of the 2nd International Conference on Music Information Retrieval (ISMIR).
- [22] Pauws, S. and Eggen, B. 2002. PATS: Realization and evaluation of an automatic playlist generator. In Proceedings of the 3th International Conference on Music Information Retrieval (ISMIR).
- [23] Pauws, S. and van de Wijdeven, S. 2005. User evaluation of a new interactive playlist generation concept. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR).
- [24] Saracevic, T. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26, 321-343.
- [25] Sharp, H., Rogers, Y., & Preece, J. 2007. *Interaction Design: Beyond Human-computer Interaction*. John Wiley & Sons.
- [26] Su, L.T. 1992. Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28, 503-516.
- [27] Vignoli, F. 2004. Digital music interaction concepts: a user study. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR).
- [28] Vignoli, F. and Pauws, S. 2005. A Music Retrieval System Based on User Driven Similarity and Its Evaluation. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR).