Inskip, C., MacFarlane, A. & Rafferty, P. (2010). Creative professional users musical relevance criteria. Journal of Information Science, 36(4), 517 - 529. doi: 10.1177/0165551510374006 <http://dx.doi.org/10.1177/0165551510374006 >

Creative Professional Users' Musical Relevance Criteria

Charlie Inskip

Department of Information Science, City University London, UK

Andy MacFarlane

Department of Information Science, City University London, UK

Pauline Rafferty

Department of Information Studies, University of Aberystwyth, Wales

Abstract

Although known item searching for music can be dealt with by searching metadata using existing text search techniques, human subjectivity and variability within the music itself make it very difficult to search for unknown items. This paper examines these problems within the context of text retrieval and music information retrieval. The focus is on ascertaining a relationship between music relevance criteria and those relating to relevance judgements in text retrieval. A data-rich collection of relevance judgements by creative professionals searching for unknown musical items to accompany moving images using real world queries is analysed. The participants in our observations are found to take a socio-cognitive approach and use a range of content and context based criteria. These criteria correlate strongly with those arising from previous text retrieval studies despite the many differences between music and text in their actual content.

Keywords: creative music search, user relevance criteria.

1. Introduction

Music is a complex concept and, although there are some similarities between music and text in terms of search, there are also some significant differences caused by this complexity [1]. Although known item searching for music can be dealt with by searching metadata using existing text search techniques, human subjectivity and variability within the music itself make it very difficult to search for unknown items. This paper examines these problems within the context of text retrieval and music information retrieval (MIR). The focus is on examining the relationship between music relevance criteria and those relating to relevance judgements in text retrieval.

The next section discusses the concept of relevance and, in particular, relevance criteria arising from studies of users of text retrieval systems. This is followed by a discussion of the contribution of relevance to evaluation, concentrating on music information retrieval systems. The approach and methodology are then presented, which involved analysing relevance judgements of music experts in the context of choosing music for television and cinema commercials. Finally the findings are presented and discussed in relation to other work on relevance criteria. This leads to the conclusion that relevance judgement categories in music appear to relate strongly to earlier findings in those relating to text, despite the many differences between music and text in their actual content.

2. Relevance

The purpose of evaluation is to measure the performance of an information retrieval (IR) system and help determine how effective it is at meeting the information needs of the users. The established measures for systems evaluation are precision (the extent to which the system is able to leave behind non-relevant items) and recall (the extent to which the system is able to find relevant items). There is a range of user-oriented evaluation measures [2] which attempt to reflect the real world, but relevance and its related measures of precision and recall are often key to the user experience.

In order to determine precision and recall, or

> "the probability of agreement between what the system retrieved or constructed as relevant (systems relevance) and what a user or user surrogate assessed or derived as relevant (user relevance)" [3]

an agreement on a definition of relevance is required. Even in the case of text this is not clear-cut, although a relevant document may be described as being one that satisfies a users Anomalous State of Knowledge (ASK) [4]. In his comprehensive review of the relevance literature, Saracevic discusses how

> 'relevance is a, if not even the, key notion in information in information science in general and information retrieval in particular" [5:2126].

Indeed, much information retrieval work relies on broad agreement on the concept of relevance. This 'broad agreement' is important. Despite much research and reflection on the concept it has been difficult for the community to agree on either a definition or a theory of relevance. The problems in agreeing on the *manifestations*, *behaviour* and *influences* [3] are profound, and centre on the personal, subjective and intuitive nature of this experience. In other words, when presented with a set of documents, humans seem to be able to choose which of those documents suit their purpose, and which do not, but there are no strict rules for these decisions, although the attributes of relevance may be summarised to include relation, intention, context (internal and external), inference, selection, interaction and measurement [3]. Mizzaro also discusses how

> "there are many kinds of relevance, not just one" [6:811].

His framework, derived from his comprehensive review, is based on a relationship between, on one side, *document*, *surrogate*, or *information* and, on the other, *problem*, *information need*, *request* and *query* [6:811] or, again, a measure of the match between aspects of the user experience and an output of the system.

Judgements of relevance compound these difficulties. Mizzaro's analysis of relevance judgements identifies

> "the kind of relevance,… the kind of judge, … what the judge can use to judge,… what the judge can use to express judgement,… the time at which judgement is expressed" [6:812].

He identifies the period 1977 onward as being a key period for user studies, the mid-1990s being a defining moment for investigation into relevance criteria.

If relevance is ‘subjective, cognitive, situational, multidimensional, dynamic and measurable [7] then it is central to its understanding that users are investigated, a paradigm shift called by Park [8] amongst others. A number of qualitative research projects (including [9], [10], [11]) attempted to identify key criteria in users relevance judgements. Schamber arranges 80 such criteria in a table [7:11]. Harter [12] also comments on the wide range of criteria derived from such studies. Saracevic summarises these criteria, which he calls ‘*clues*’ [5:2130], noting that although there is variety in their labels, they are ‘remarkably similar’ in concept. He generalises them into groups, detailed in Table 1 (below). The arrows indicate the dynamic relationship between these characteristics where relevance judgements are concerned:

Information characteristics	
Content	Topic, quality, depth, scope, currency, treatment, clarity
Object	Characteristics of information objects, eg type, organisation, representation, format, availability, accessibility, costs
Validity	Accuracy of information provided, authority, trustworthiness of sources, verifiability
↕	↕ ↕
Individual characteristics	
Use or situational match	Appropriateness to situation, or tasks, usability, urgency; value in use
Cognitive match	Understanding, novelty, mental effort
Affective match	Emotional responses to information, fun, frustration, uncertainty
Belief match	Personal credence given to information, confidence

Table 1 Adapted from Saracevic [5] Information and individual characteristics

As this research is looking at a relationship between information and the individual, clearly the content of the document and the context of the individual are important contributors. This context is not only cognitive but also social: considering the role of socio-cognitive relevance in interactive information retrieval (IIR) systems [13] then this will include not only the system and the user but also the environment, giving rise to a holistic approach to evaluation. The issue of context in IIR [14], [15] is a key topic in information retrieval in the new century: human behaviour is operating within the larger system of socio-cultural codes and competences which are likely to influence cognitive processes, and relevance judgements are wider than purely relationships between user and information. The content of documents has been central to the development of IIR systems. Algorithmic or system relevance and topical or subject relevance focuses on document (or surrogate) content, while cognitive relevance or pertinence, situational relevance or utility and affective relevance may draw influences from both content and contextual factors [3].

Algorithms attempt to rank a selection of documents drawn from a system in order of relevance by examining such criteria as the frequency of key words appearing in the document and where in the document the key words appear. In known item searching (‘*I am looking for the text of Speaker Michael*

Martin's resignation speech') it can easily be determined whether the documents found by a system are relevant or not. This is a binary (yes/no) decision and a system with high precision will successfully retrieve a number of documents containing this information. In unknown item or question answering search (*'I want to know whether tadpoles eat mosquito larvae'*) this can be a more difficult decision and the user needs to review each document to establish its relevance. This is normally determined by whether it resolves their information need. A system which has high recall will offer a large number of results, probably containing the words 'tadpole', 'eat', 'mosquito' and 'larva'. If one of these documents answers the question, the search is complete, if it does not, the query may be revised by the user and a new set of results reviewed.

It is possible to see a link between text and music retrieval. In known item search for music (*'I want to hear David Bowie's 'Heroes' sung in German'*) it is easy to determine whether a system has successfully met this request by the user listening to a short extract of the piece (as long as the piece includes sung vocals), and it is likely that the system will find this using metadata text search, if the recording is in the system. All other versions of the song ('Heroes' in any other language, sung by Bowie, or covers of 'Heroes' sung in German by other artists) will not be relevant. Again, this is a binary decision. This may also be the case when a user is searching for a known piece of music (Bowie's 'Heroes' in German) but cannot remember or does not know the detail, which may give rise to the query *'Find me a song about brave men or supermen sung in German by a male singer'*. This could generate a range of results including extracts from Wagner's Ring Cycle, German covers of Laurie Anderson's 'O Superman' and schlager versions of 'Scotland The Brave', depending on the collection(s) being searched. The user, again, would have to browse through the list and listen to the results until the relevant item was found. There is still only one correct (and many incorrect) answers to this query, however.

Music search engines (MSE) face a very difficult problem if users do not know which song they are seeking *'until I hear it'* – a frequent situation when searching for music to accompany moving images. A clichéd query, mentioned in numerous interviews conducted in researching this subject, would be *'I'm looking for something quirky and uplifting with a bit of a build'*. In MSEs these subjective ('*quirky*', *'uplifting'*) and musical-feature ('*build'*) descriptions can be applied to metadata fields using human indexing. Unfortunately *'quirky'* could be based on rhythmic, lyrical, genre or non-musical cultural features, one listener or indexer's '*uplifting'* may be another's '*inspiring'*, and a '*build'* could be anywhere in the piece (or may be mistaken for a lyric, artist or title element). These difficulties need to be accounted for when determining the precision and recall of a particular system. The fact that there is not one correct answer and that the user wishes to choose rather than be told which is the 'best' choice, suggests that successful systems may need to have higher recall than precision.

3. Evaluation of Music Information Retrieval Systems

Rasmussen recommends that relevance should relate to the task context if it is to reflect a user-centred approach:

> Because evaluation is based on query averages for precision and recall, there is little emphasis on the flexibility to match user needs and outcomes. From the user's perspective, process and strategy may be as important as outcomes, and mechanisms to study process and measures to evaluate it are still lacking. [16:48]

This view is supported in image retrieval, for example, where Sormunen et al [17] recommend:

> In outlining an evaluation framework, the first task is to define the function of a system that is to be evaluated. The framework has to include a description of potential users, their needs and the performance criteria relevant to users. [17:4]

Here, image retrieval evaluations consider the users, focusing on the problems they have in formulating queries, noting that they prefer browsing over query, and that their final decision is made using criteria which can be difficult to identify as they are not clearly based on image content or textual descriptions. Although there are many differences between images and music, these considerations frequently arise in MIR, particularly in unknown item searching.

The MIR community have been evaluating algorithms in an annual 'contest', MIREX (Music Information Retrieval Evaluation eXchange) [18], since 2004. This was informed by a rigorous research and consultation process using TREC [19] as a model. MIREX was established to deal with three main issues in MIR evaluations:

> 1. no standard collection of music against which each team could test its techniques;
>
> 2. no standardized sets of performance tasks; and, 3. no standardized evaluation metrics. [19]

These three issues are aspects that TREC have recognised as key to scientific evaluation [19], confirmed by Voorhees [20] during the MIREX consultation process. Downie [20] highly recommends the use of precision and recall as metrics when evaluating systems and algorithms. However the MIR community is not united in this approach and the participants employ a wide range of metrics.

Rüger [21] also discusses how a wide range of tasks in MIR (ad hoc searches, audio identification, classification, feature extraction) should be identified, each requiring specific approaches to evaluation. This reinforces the point that while precision and recall are appropriate system evaluation metrics they are not always the best measures. Downie [22] in effect summarised the key issues: defining relevance, building consensus and structuring reliable and valid tests. He also strongly urged for representation from a wide range of disciplines and that the needs of users be recognised by the evaluations and the community as a whole.

In the light of these and other ongoing discussions, in MIREX teams of researchers agree on performance tasks, such as Genre Classification, Artist Identification and Mood Classification. Each task is evaluated using metrics specific to that task and there is no blanket application of precision and recall. Accuracy was the predominant evaluation measure in 2004 and 2005 [18]. In 2006 a clearer focus was brought onto precision and recall (onset detection, query by singing, melodic similarity) while an additional range of statistical analysis was also used (p-score, f-measure, ANOVA) [23], which was no doubt applauded by Flexer [24] who had called for a more rigorous approach to statistical evaluation in MIR.

If precision and recall are to be more widely used in MIR evaluation then determining relevance is paramount. As it can be problematic to determine the 'meaning' of music when there is no 'ground truth', relevance decisions are difficult. In an attempt to resolve this, Downie proposes that:

> there should be enough information contained within the query records that reasonable persons would concur as to whether or not a given returned item satisfied the **intention** of the query. [1]

This approach has been taken on board by a number of MIREX participants, who gather ground truth data from human volunteers who participate in evaluations (including, for example, [25], [26], [27], [28]), generating subjective ground truth sets which are required when there is no objective ground truth available. Although this can be time consuming and expensive, and can be inaccurate and corrupted [23], it does indicate a worthy attempt to solve this problem and indicates a move towards user-oriented evaluations.

4. Research question

Part of this research involves evaluation of a selection of 6 Music Search Engines, operated by major companies in the music industry, which have been developed for the purpose of disintermediating the process of music synchronisation. A set of 27 real written queries ('briefs') have been collected from creative music searchers. These verbose and subjective queries relate to the information needs of the makers of TV and cinema commercials and TV (the Users). Their purpose is to communicate their need to the rights holders (Owners) who then use this information to generate search result sets. The question is 'how do these creative professionals relevance criteria relate to users' relevance criteria identified in text retrieval studies?' It is hoped that these findings will contribute towards the under-researched area of defining music relevance and thus towards appropriate systems development and evaluation.

5. Methodology

Keywords and concepts were extracted from the briefs and applied to each Music Search Engine. This generated sets of results. These results were then evaluated by expert intermediaries to determine whether or not a given returned item satisfied the **intention** of the query [1].

A range of facets used within the search engines and identified by Inskip et al [29] were used in coding the queries. If a term arose which did not match existing facet codes a new code was generated. This iterative approach led to a comprehensive set of facet codes which were applied to a set of written briefs, as in the example below:

Query 009 (coded):

009.

We are looking for a <MOOD> cool </MOOD> <MOOD> fun </MOOD> <TEMPO > jaunty </TEMPO> and <MOOD><TEMPO> upbeat </MOOD> </TEMPO> track with a <MOOD> happy vibe </MOOD> and a certain <MOOD> feel good factor </MOOD> <MOOD> it shouldn't take itself too seriously </MOOD>. Ideally it should be from a <DATE> new </DATE> and <DATE> <CHART> up-and-coming </DATE> </CHART> <ARTIST> artist </ARTIST> ; [client] <MUSIC FUNCTION> would like to be associated with a <MOOD> fresh </MOOD> <DATE> new sound </DATE> </MUSIC FUNCTION>, and not with something

<DATE> old </DATE> or <DATE> dated </DATE>.

<MUSIC FUNCTION> The music should guide us through the story and mirror the

<VISUALS SUBJECT> positive journey the main character is taking </VISUALS SUBJECT> </MUSIC FUNCTION>. <VISUALS SUBJECT> He is in his own little world of fun, which contrasts with the busy urban surroundings </VISUALS SUBJECT>. The music should be <MOOD> positive, </MOOD> <MOOD> easy going <MOOD> and <MUSIC FUNCTION> make the listener smile. </MUSIC FUNCTION>

Although the overall tempo of the song should be <TEMPO> upbeat </TEMPO> to

<MUSIC FUNCTION> reflect the gliding motion of the journey, </MUSIC FUNCTION> <TEMPO> the pace should be varied, </TEMPO> and the track <MUSIC STRUCTURE>

should have some <VOLUME> quieter </VOLUME> moments and enough space to accommodate <EXTRA-MUSICAL> sound effects <EXTRA-MUSICAL> </MUSIC STRUCTURE> –

<VISUALS SUBJECT> the character will be going down the slide at different speeds at different points, occasionally slowing down or even stopping. </VISUALS SUBJECT> <EXCLUDE> Please avoid </EXCLUDE> anything too <GENRE> folky </GENRE> or <MOOD> dreamy. </MOOD> Any <LYRICS> lyrics should relate loosely to the story of the ad, which conveys a positive journey. </LYRICS>

<EXCLUDE> Please avoid any songs with <LYRICS> specific lyrics, e.g. to do with driving a car.</LYRICS> </EXCLUDE>

Each coded query was to be applied to each of the six search tools, giving a maximum of 162 results sets. However it was found that a number of queries from one source (a TV company 'trailers' department) were not suitable for coding as they contained no direct information about the music required for the trailer, this decision being left entirely to the creative making the trailer. These entirely contextual briefs were discarded from the coding exercise as none of the services had the facility to construct a query using the information contained within the brief. After removing duplicates the set of briefs was reduced to 19. Queries were applied step-by-step. In some cases applying all facets led to a return of no results. In these cases terms were removed from the query until a manageable set of results (preferred number of 10) were returned. If a larger set was returned from using all the terms then the first ten results were kept as a results set.

The results sets of up to 60 results for each brief were made into numbered playlists using Spotify software [30]. Spotify is a web-based application which allows users to listen to a large selection of streamed music without infringing copyright regulations. This application was chosen because it is widely used by the participants in their work and would allow the playlists to only be accessible to all the members of the panel of experts during their observations by researcher log-in, without infringing copyright by downloading material. Although not all of the material was available on this service it was felt that sufficient songs were included to make the use of the application valid. For example, in observation 024SPOT, 49 songs were generated, and 41 of these tracks were available on Spotify. These were randomised to make the source rights holder less obvious in case of bias.

Seven of these playlists were then evaluated by seven creative music search experts, each listening to and commenting on the songs on one list. The experts were drawn from a pool of previous interview participants and others who had not previously been interviewed but had been recommended as possible participants by interviewees. This satisfies the snowball sampling approach used to select research participants throughout this project. Each expert participant was presented with a written brief and asked to read it. They were advised that they could write on the brief if they wished. They were then asked to listen to tracks that had been generated by the briefs being applied to the search engines and comment on '*whether or not the track meets the brief*'. Each observation session lasted for around 45 minutes and most participants completed their allocated playlist within this period. Creative music searchers' time budgets are limited and we were careful not to make them unwilling to take part in future research by burdening them. It was felt that it was more important to generate rich and detailed data than to elicit relevance judgements on every song on the list. The sessions were recorded using a digital voice recorder and transcribed (including repetitions and pauses) within 48 hours. The data was then analysed to gain insight into the relevance judgements of this group of music searchers.

6. Findings and Discussion

The relevance judgements, anonymised numerically here ([35]-[41]) were coded in detail, using codes extracted from the Music Search Engines but iteratively generating new codes when terms were introduced that did not fit into the existing framework. These codes were then quantified, ranked by frequency, and categorised according to Saracevic's [5] information and individual characteristics (Table 1 above). These characteristics were also categorised as to whether they were context- or content-based. This gave rise to Table 2, below:

Context or content	Characteristics (Information)	Code	Quantity
Content	**Content**	Mood	327
		Genre	97
		Lyrics	81
		Date	76
		Production	64
		Instrument	60
		Tempo	47
		Music structure	45
		Music function	40
		Vocal	39
		Artist	24
		Music style	22
		Instrumental	18
		Build	12
		Version	11
		Song title	3
		Feel	3
		Volume	1
		Song subject	1
	Characteristics (Individual)		
Context	**Use / situational**	Visuals subject	33
		Extra-musical	24
		Audience	14
		Visuals	13

		Brand	9
		Time availability	4
	Object	Budget	16
		Clear	14
		Syncability	9
		Territory	4
		Format	3
		Owners	1
	Content	Similar	7
	Cognitive	Novelty	84
		Message	8
	Belief	Would not pitch	13
		Would pitch	2
	Affective	Subjective	25

Table 2 Coded criteria and characteristics

6.1. Musical content

In the context of previous findings [29][31] it did not come as a surprise that 'Mood' criteria clearly outweighed all other content aspects. References were predominantly drawn from the briefs themselves, although there was noticeable addition of mood criteria which did not appear in the briefs but were raised by the participants.

Mood criteria exceed others in frequency in the briefs themselves [31]. The affective nature of music description has previously been identified as being popular amongst unknown item music searchers [32]. Affective aspects of content, however, are not often discussed as relevant criteria in the text-retrieval literature. Text users' relevance criteria appear to be more related to the 'topic' of the document [33] or the 'goal' of the user [8], hence the development of IR systems which determine the 'subject' of the document from its word content. Determining the 'meaning' of music, its 'subject' or 'topic' is only possible from the perspective of the listener, not from the content alone [34]. Lyrics can be used to elicit perceived meaning but this can vary widely between listeners and is very context-sensitive. Determining 'mood' is equally problematic:

> …it doesn't sound 'fresh' in any way whatsoever to me… [35];
>
> But it doesn't have enough of that sort of the playful sort of nature [37];
>
> So essentially the first element, for me, could mean absolutely any piece of music whatsoever. [38].

However its wide useage as a content-based relevance criterion means more work has to be done on determining the meaning of mood descriptors if music IR systems are to truly reflect human information behaviour.

Some musical criteria (lyrics, date, artist, song title) are factual and can be presented via textual metadata. As relevance criteria they are easy for a system to resolve. They are used frequently in these observations, indicating their importance as expressions of musical relevance:

> Lyrically it's not very appropriate [37];
>
> lyrically, I mean it would work [37];
>
> Obviously I think lyrics are a massive thing here[38];
>
> I don't think the lyrics are really going to work [39].

Other aspects, such as 'production', 'song subject', are not so easy to define using metadata, either because of multiple sub-criteria (production can be defined by a wide range of factors ranging from the name of the producer, the period or 'feel' of the recording to the density or number of instruments) or because of the multi-layered meanings of the successful pop song:

> more polished production more – more commercial sounding [41];
>
> Could work, maybe a bit in your face [39];
>
> it's quite clean and precise [38];
>
> the production does sound sort of karaoke [37].

Despite the problems in defining many of these criteria they are still widely used in making relevance decisions. Feedback between the Users and the intermediaries and Owners helps clarify briefs and establish mutual agreement on the meaning or relevance of the music in question.

6.2. Contextual

The most frequently used contextual criterion was that of 'novelty':

> it sounds very clichéd [35];
>
> heard this a million times [37];
>
> It could fall into standard ad music category [38];
>
> It's been heard before [40].

This is a cultural phenomenon based on the fact that this music use is within the context of a highly competitive industry determined to get the attention of the distracted television viewer. It is widely believed in the industry that a novel or previously unused piece or style of music is more likely to gain their attention, hence the need to find 'previously unused' pieces of music, although these must not be so far out of the viewers experience as to alienate them from the product being presented in the commercial. This cognitive criterion cannot be found within the music itself but only through an analysis of the history of uses of that piece of music (and music of a similar type) in synchronisation and in wider culture. It features as a keyword and also as a criterion not featured in the brief and is frequently discussed in interviews on this subject as a key 'unknown item' relevance criterion.

The 'characteristics of the information object', or, here especially, the availability and cost issues, are also brought in by the participants as relevance criteria. The budget, likelihood of successful clearance, territory, and ownership factors are indeed relevant to whether an intermediary will present a piece of music for consideration.

> even if they could afford it, which is unlikely [36];
>
> it's certainly going to be cheap [37];
>
> would the client want to pay for a Vangelis piece of music that people wouldn't recognise [41];
>
> Mercury Rev are not going to allow their track to be used for a washing powder ad [36];

Not going to get cleared for an advert. [38].

If they present something that will lead to difficulties or unexpected expenses in business negotiation this is likely to cause problems between them and their clients. The 'syncability' of a piece of music, or, more likely the syncability of a genre, specifically rap and hip hop, American R&B and various forms of heavy rock are frequently discussed:

hip hop stuff's really difficult to to sync [35];

hip hop, r&b that kind of thing steer clear of [37];

I've never licensed a piece of R&B [38];

hip hop's always a tricky one [39].

It is agreed at all levels that these genres are 'unsyncable' and were often dismissed from these evaluations even without being played, purely from the Artist information. It is acknowledged that these observations took place in London and are specific to British advertising and likely that cultural differences will make significant difference to this particular judgement if one was looking at, say, the American market.

Use and situational aspects are also brought to bear in relevance decisions. The subject of the visuals, which were not available to the participants apart from as described in the brief, is a key factor determining the relevance of the material chosen. If a visual is described as '*cold and clinical*' (Brief 026) then the intermediary makes a relevance decision based on whether the music is '*cold and clinical*', or whether it could be described as such. Matching musical elements to visuals is a highly subjective and contextual problem, although it does vary in complexity according to the concepts used in the description and the weights ascribed to the criteria in the brief.

it's something that I think I'd probably put to picture, just to just to try it [37];

it's not going to slow down the visuals [38];

clearly you have to see it first [40].

Extra-musical criteria, such as voiceovers or sound effects are often added to the soundtrack of a commercial and frequently the participants discussed how the music needed to leave space for a voiceover or sound effects:

I don't think it would have any space to accommodate sound effects [35];

is probably going to clash with a voiceover [36];

it doesn't actually mention anything here about voiceover [37].

The intended audience also impact on the choice, and matching the music to the demographic is often thought of as being extremely important to the eventual choice.

it's not going to appeal to most purchasers of washing powders, which is going to be woman ages 25-45 [36];

it depends on who they're trying to sell it to, what the demographic is [38];

to appeal to the kind of the middle class older generation and also the younger demographic as well. [39].

This is even tested by the Users when the commercial is completed, as an 'end-user' relevance evaluation. Other important situational aspects include whether the music would suit the brand, and the time available to the creatives involved in the process.

Finally, the belief aspect, or 'personal credence given to information, confidence' [5:2130] indicates whether (or not) they would pitch a track, determined by their own tastes and gut instincts. If a track meets all other aspects of a brief but does not match the taste or world view of the intermediary it is dropped:

that works to a certain degree. I wouldn't pick it though. Haha [38];

> Not something I'd particularly like, wouldn't pitch it, but it's probably the kind of thing that they might want to hear. [35];
>
> Gary Glitter, I just wouldn't put him forward because he's a bloody paedo. But let's listen to it. [40].

The belief systems of the creatives are, with some, the bottom line. They are using their choices in music to identify themselves amongst a number of competitors and expressing their expertise through presenting a set of results which will be particular to them and not replicable by another searcher.

6.3. Summary

These observations show that the relevance judgements of these participants are framed within Saracevic's schedule of Information and Individual criteria. This reinforces and builds on the work of the text-retrieval user studies of the mid-1990s. The relevance judgements of these creative music professionals are situated in a socio-cognitive paradigm. They do not only make judgements based on whether a piece 'sounds' right (content criteria), whether it includes elements that appear to match the brief (content and context), but also whether the piece would be used by the end User (contextual criteria).

Many criteria which are not explicit in the query, some content-based but mainly contextual, are added by the experts, using their own codes and competences, when reviewing the material. For example there is often no mention in the brief of business issues such as whether the artist would allow their music to be used to advertise this type of product or whether the available budget would cover the music suggested for use, but these issues come up frequently in the comments relating to this query when the participant is making their relevance decisions:

> Someone like Barbara Streisand's just going to be – she's going to be expensive. That's obviously something that comes in down the line. But it's still, you know, especially you know, when you're looking at something – well looking at anything, you need to take that into consideration. So that obviously it's not a creative thing, but it is something that limits us. … [37];
>
> All I'd say is good luck with licensing Jethro Tull. For a mobile phone commercial. [haha really?] yes [they're tricky are they?] yes. incredibly expensive. In my experience [41].

These and other contextual criteria rarely appear in the briefs but are used extensively in the relevance decisions of these participants and are based on their extensive experience working in this area.

7. Conclusion

Saracevic's observations on relevance criteria have been discussed in reference to findings using a data-rich collection of relevance judgements by creative professionals searching for unknown musical items to accompany moving images using real world queries. It has been shown that the criteria synthesised by Saracevic from his thorough review of user evaluation literature in text retrieval correlate strongly with those arising from a close analysis of expert music user observations, particularly how a range of measures may be used depending on criteria and approach. The participants in our observations use a range of content- and context-based criteria, taking a socio-cognitive approach. Their information criteria are predominantly Content based, although Object aspects are also important. Validity is not a key concept and does not appear to often arise. Individual characteristics are predominantly Use and Cognitive, although Affective and Belief aspects also have an involvement.

Overall relevance judgement categories in music, therefore, appear to relate strongly to earlier findings in those relating to text, despite the many differences between music and text in their actual content. However the importance of the highly subjective nature of musical features such as 'mood' and the wide ranging technical difficulties in extracting features from music mean that the development of systems that are able to comprehensively reflect the users' situation is a highly complex problem. It is hoped that, as those systems are built and evaluated, approaches are made to incorporate increasingly higher levels of relevance criteria into their development in order to produce better performing, more useful and more usable services.

8. Future work

In this ongoing PhD research there will be more investigation into the search process of creative professionals working in music synchronisation, analysing their discourses in more detail with a view to identifying interpretive repertoires that may shed more light not only on their relevance judgements but also on their wider information behaviour. It is planned to present a holistic review of this process in the hope that this work will inform systems development not only in this area but in the wider area of recreational use.

9. Acknowledgements

The researchers would like to thank all the participants for taking part in these observations and being so free with their comments and valuable time. Charlie Inskip gratefully acknowledges financial support from AHRC for this PhD research.

10. References

[1] J.S. Downie, Toward the Scientific Evaluation of Music Information Retrieval Systems, Proceedings of International Symposium on Music Information Retrieval, Baltimore, USA, Oct 27-30 2003.

[2] D. Bawden, User-Oriented Evaluation of Information Systems and Services (Gower, Aldershot, 1990)

[3] T. Saracevic, Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. Journal of The American Society for Information Science and Technology 58(13) (2007) 1915-1933

[4] N. Belkin, R. Oddy, & H. Brooks. ASK for Information Retrieval. Journal of Documentation, 38 (1982) 61-71 (part 1) & 145-164 (part 2).

[5] T. Saracevic, Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behaviour and effects of relevance. Journal of The American Society for Information Science and Technology 58(13) (2007) 2126-2144

[6] S. Mizzaro, Relevance: the whole history. Journal of The American Society for Information Science 48(9) (1997) 810-832

[7] L. Schamber, Relevance and information behaviour. Annual Review of Information Science and Technology 29 (1994) 3-48

[8] T.P. Park, Toward a Theory of User-based Relevance: a call for a new paradigm of inquiry, Journal of The American Society for Information Science 45(3) (1994) 135-141

[9] C. Barry, User-defined relevance criteria: an exploratory study. Journal of The American Society for Information Science 45(3) (1994) 149-159

[10] C. Barry, & L. Schamber, Users' criteria for relevance evaluation: a cross-situational comparison. Information Processing and Management 34(2/3) (1998) 219-236

[11] T.P. Park. The Nature of relevance in Information Retrieval: an empirical study. Library Quarterly 63(3) (1993) 318-351

[12] S. Harter, Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of The American Society for Information Science 47(1) (1996) 37-49

[13] E. Cosijn, & P. Ingwersen, Dimensions of relevance. Information Processing and Management 36 (2000) 533-550

[14] T.D. Anderson, Studying human judgments of relevance: interactions in context. Information Interaction in Context Proceedings of 1st IIiX Symposium on Information Interaction in Context, Royal School of Library and Information Science, Copenhagen, Denmark, Oct 18-2-, 2006

[15] I. Ruthven, Interactive Information Retrieval, Annual Review of Information Science and Technology 42(1) (2008) 43-91

[16] E. Rasmussen, Evaluation in Information Retrieval, The MIR/MDL Evaluation Project White Paper Collection Edition #3 (2003) 45-49 Available at: http://www.music-ir.org/evaluation/wp.html (accessed 23 March 2010).

[17] E. Sormunen, M. Markkula, K. Järvelin, The Perceived Similarity of Photos - A Test-Collection Based Evaluation Framework for the Content-Based Image Retrieval Algorithms, Final Mira Conference, 1999. In: S. Draper et al. (eds) Mira 99: Evaluating interactive information retrieval. Electronic Workshops in Computing (eWic).

[18] MIREX (2009) MIREX 2005-2008 Wikis, Available at http://www.music-ir.org/mirex/2009/index.php/Main_Page#MIREX_2005_-_2008_Wikis (accessed Tuesday, 19 January 2010)

[19] TREC (2010) Text REtrieval Conference, Available at http://trec.nist.gov/ (last accessed Tuesday, 19 January 2010)

[20] E. Voorhees, Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC, The MIR/MDL Evaluation Project White Paper Collection Edition #3 (2003) 7-13 Available at: http://www.music-ir.org/evaluation/wp.html (accessed 23 March 2010).

[21] S. Rüger, (2003) A Framework for the Evaluation of Content-Based Music Information Retrieval using the TREC Paradigm, The MIR/MDL Evaluation Project White Paper Collection Edition #3 (2003) 68-70 Available at: http://www.music-ir.org/evaluation/wp.html (accessed 23 March 2010).

[22] J.S. Downie, Interim Report on Establishing MIR/MDL Evaluation Frameworks: Commentary on Consensus Building, The MIR/MDL Evaluation Project White Paper Collection Edition #3 (2003) 43-44 Available at: http://www.music-ir.org/evaluation/wp.html (accessed 23 March 2010).

[23] J.S. Downie, The music information retrieval exchange (2005-2007): a window into music information retrieval research. Acoustic Science and Technology 29 (4) (2008) 247-255

[24] A. Flexer, Statistical evaluation of music information retrieval experiments, Journal of New Music Research 35 (2) (2006) 113-120

[25] A. Gruzd, J.S. Downie, M. Cameron Jones, J.H. Lee, Evalutron 6000: Collecting Music Relevance Judgments, Proceedings of Joint Conference on Digital Libraries '07, June 17–22, 2007, Vancouver, British Columbia, Canada.

[26] M. Mandel, & D. Ellis, A Web-Based Game For Collecting Music Metadata, Proceedings of 8th International Conference on Music Information Retrieval, Vienna, Austria, Sep 23-27, 2007.

[27] L. Barrington, R. Oda,, & G. Lanckreit, Smarter than Genius? Human Evaluation of Music Recommender Systems, Proceedings of 10th International Society for Music Information Retrieval Kobe, Japan October 26-30, 2009

[28] E. Law, & L. von Ahn, Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games, Proceedings of CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

[29] C. Inskip, A. Macfarlane, & P. Rafferty, Organizing Music for Movies, Proceedings of International Society for Knowledge Organization (UK) Content Architecture conference, London, UK, 22-23 Jun 2009

[30] Spotify (2009) A World of Music, Available at; http://www.spotify.com/en/ (accessed Monday, 18 January 2010)

[31] C. Inskip, A. Macfarlane, & P. Rafferty, Towards the Disintermediation of Creative Music Search: Analysing Queries To Determine Important Facets. Proceedings of ECDL Workshop on Exploring Musical Information Spaces, Corfu, Greece, 1-2 Oct 2009

[32] J. Kim, & N. Belkin, Categories of Music Description and Search Terms and Phrases Used by Non-Music Experts. Proceedings of 3rd International Conference on Music Information Retrieval IRCAM – Centre Pompidou, Paris, France October 13-17, 2002

[33] P. Borlund, The concept of relevance in IR, Journal of The American Society for Information Science and Technology 54(10) (2003) 913-925

[34] C. Inskip, A. Macfarlane, & P. Rafferty, Meaning, communication, music: towards a revised communication model, Journal of Documentation 64(5) (2008) 687-706

[35]-[41] Anonymised texts from relevance judgement observations