REDUCING BIAS IN THE PROBABILISTIC EVALUATION OF AUDIO MUSIC SIMILARITY

Adriana M. Suárez I.

A thesis submitted for the degree of Master in Sound and Music Computing September 2015, Barcelona, Spain

Advisor:

PhD Julian Urbano

Department of Information and Communication Technologies

Universitat Pompeu Fabra

To my beloved brother Julio, who was my biggest supporter

Acknowledgments

This work would not have been possible without the help, the support and contribution of many people. First, and foremost, I thank God for the strength and blessings in completing this work. I will always be thankful for the chance, power and wisdom provided during this master course.

Second, I would like to thank my supervisor Dr. Julian Urbano for his valuable and generous guidance, advice and encouragement through this study. He is a committed researcher and a master programmer in *R* who has always been willing to transmit all he knows without hesitation. Under his supervision, I learned a lot of statistic and programming tricks in *R*. More importantly, I instilled the spirit of challenge and confidence of going for research of topics considered "different" and still maintain the conviction until the end.

Third, I must express my gratitude to all the crew from the Music and Sound Technology group. To their head, Dr. Xavier Serra, for the opportunity he provided me to take part of this master. Also, to the teachers and classmates whom I had shared this chance. It was a great pleasure to study in a multicultural environment. All the knowledge they transmitted is truly appreciated.

Fourth, I cannot forget the valuable support of my international friends that I had the opportunity to meet in my stay in Barcelona. I appreciate their encouragement throughout the ups and downs of my study, and their presence and help throughout my stay away from home.

Lastly, but by no means the least, I am extremely grateful to my family to whom this thesis is dedicated. I will be always helpful for their immeasurable love, care and support, which they have always given and it has always been helpful; even from distance. Grateful for understanding this journey that made a significant

change in my career, in which I bet a lot but also I know I gained even more, both as a person and as a professional. Also grateful for comprehending my physical absence in meaningful family times. Moltes gràcies. Thank you a lot. Muchas gracias.

“Thank you for the music, the songs I'm singing
Thanks for all the joy they're bringing
Who can live without it, I ask in all honesty
What would life be?
Without a song or a dance what are we?
So I say thank you for the music
For giving it to me”

ABBA

Abstract

This thesis work conduits research toward the estimation of relevance judgments for the task of Audio Music Similarity in the context of MIREX. It is intended to improve and support the evaluation experiments run for this task from the point of view of efficiency, studying different probabilistic models and methods with the aim of reducing the cost of the annotation process. Therefore, by doing better estimations of relevance judgments and using all the tools at hand (research, literature, technology) the time used by people performing this task can be utilized in others activities.

Resumen

Este trabajo de tesis consiste en una investigación acerca de la estimación de juicios de relevancia para la tarea de Audio Music Similarity en el contexto de MIREX. Fue creado para mejorar y apoyar los experimentos de evaluación realizados desde el punto de vista de la eficiencia, estudiando diferentes modelos y métodos probabilísticos con el objeto de reducir el costo del proceso de anotaciones. Realizando mejores estimaciones de los juicios de relevancia y usando las herramientas disponibles (investigación, estado de la técnica, tecnología) el tiempo usado por las personas que realizan esta tarea actualmente, podría ser mejor utilizado en la ejecución de otras actividades.

Index

Figure List

Table List

Chapter 1
INTRODUCTION

Information retrieval (IR) is the field concerned with representing, searching, and manipulating large collections of electronic text and other human-language data (Buettcher, Cormack and Clarke, 2010). A user has information need and use an IR system in order to retrieve relevant information from a document collection.

IR systems are boundless and even essential nowadays since they facilitate daily life of people supporting activities in business, entertainment, education, medical services, and so on. Web services engines like Google, Yahoo, among others are the most popular web IR services for their great capacity of converging information from different sources. Music IR systems like Shazam, implementing music identification technology are quite popular and useful today.

These systems have been used prior the invention of computers. Before 1940's intelligence and commercial retrieve systems where already implemented and just until the appearance of the first computer-based systems, mechanical and electro-mechanical devices performed the retrieve functions. With the generalization of the computers, IR technics grew up as the increase of storage and processor speed allowed managing bigger datasets (Sanderson, M., & Croft, W. B, 2012).

IR has been widely used through the story from several fields: text and cross-language, image and multimedia, speech and music (Manning, C. D., Raghavan, P., & Schütze, H, 2008). In the case of music, Music Information Retrieval (MIR) is concerned on the extraction and inference of meaningful features of music, it's indexing and the development of different search and retrieval schemes (Downie, J. S, 2003). It started with the analysis of symbolic representations of songs (mostly MIDI scores); with the evolution of computing systems during the early 2000's, signal processing was also included permitting the extraction of features

directly from the audio. (Manning, C. D., Raghavan, P., & Schütze, H, 2008). These features are pitch, temporal, harmonic, timbral, editorial, and textual and bibliographic facets.

1.1 Information Retrieval

Information retrieval (IR) is the field concerned with representing, searching, and manipulating large collections of electronic text and other human-language data (Buettcher, Cormack and Clarke, 2010). A user has information need and use an IR system in order to retrieve relevant information from a document collection.

IR systems are boundless and even essential nowadays since they facilitate daily life of people supporting activities in business, entertainment, education, medical services, and so on. Web services engines like Google, Yahoo, among others are the most popular web IR services for their great capacity of converging information from different sources. Music IR systems like Shazam, implementing music identification technology are quite popular and useful today.

These systems have been used prior the invention of computers. Before 1940's intelligence and commercial retrieve systems where already implemented and just until the appearance of the first computer-based systems, mechanical and electro-mechanical devices performed the retrieve functions. With the generalization of the computers, IR technics grew up as the increase of storage and processor speed allowed managing bigger datasets (Sanderson, M., & Croft, W. B, 2012).

IR has been widely used through the story from several fields: text and cross-language, image and multimedia, speech and music (Manning, C. D., Raghavan, P., & Schütze, H, 2008). In the case of music, Music Information Retrieval (MIR) is concerned on the extraction and inference of meaningful features of music, it's indexing and the development of different search and retrieval schemes (Downie,

J. S, 2003). It started with the analysis of symbolic representations of songs (mostly MIDI scores); with the evolution of computing systems during the early 2000's, signal processing was also included permitting the extraction of features directly from the audio. (Manning, C. D., Raghavan, P., & Schütze, H, 2008). These features are pitch, temporal, harmonic, timbral, editorial, and textual and bibliographic facets.

1.2. Information Retrieval Evaluation

Evaluation has come to play a critical role in information retrieval research (Downie, 2002) as it allows measuring how successfully an information retrieval system meets the goal of assessing users to fulfill their information needs. The IR community has paid a lot of attention to the topic, implementing evaluation standards and experimental rigor on investigations, which have been effective in moving the field forward. Music Information Retrieval initially followed the evaluation practices of text; however, not enough research has been done to properly know when this approach can be fully applied or not because music, unlike text has, a complex nature.

1.2.1 Early Work in Text Information Retrieval Evaluation

Evaluation in Text Information Retrieval has been the focus of a lot of research:

- The *Cranfield Project 2* (1962-1966) was an experiment accomplished by Cyril Cleverdon (Cleverdon, 1991) and considered as the basis that shaped the form that IR evaluation will take for the next years. In this project, experiments were conducted in order to test and compare different search strategies in a controlled laboratory environment (test collection).

- The *MEDLARS* (Medical Literature Analysis and Retrieval System) Demand Search Service (1966-1967) was one of the early operational computer-based retrieval systems. It considered the evaluation of a complete system from a user perspective, taking into consideration the user requirements (Lancaster, 1968).

- The *SMART project* (1961-1995) (System for the Mechanical Analysis and Retrieval of Text) was created both as a retrieval tool and as a vehicle for evaluating the effectiveness of a large variety of automatic search and analysis techniques, where the main evaluation viewpoint taken was the user (Kent, Lancour, Daily, 1980).

- *TREC* [2] (Text Retrieval Conference) started (1992) as an annual venue to support research within the information retrieval community by providing the necessary infrastructure for large-scale evaluation of text retrieval methodologies.

- *NTCIR* (National Institute of Informatics- Test beds and Community for Information access Research) (1999) provided almost the same infrastructure than TREC but for Asian languages.

- *CLEF* [2] (Conference and Labs of the Evaluation Forum) (2000) was created to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal.

- *INEX* (Evaluation of XML retrieval) (2002), which focuses on structured information.

[1] http://trec.nist.gov/overview.html

[2] http://www.clef-initiative.eu/web/clef-initiative/home

1.2.2 Early Work in Music Information Retrieval Evaluation

Some initiatives towards the development of Music Information Retrieval evaluation frameworks took place. The organization of the first International Symposium on Music Information Retrieval *(ISMIR)* in 2000, with the intention of bringing together the MIR research community into one location to treat among other topics, the creation of formal evaluation standards for MIR (Downie, 2000) was one of them. As a consequence, some workshops on the creation of standardized test collections, tasks and metrics for music digital library (MDL) and Music Information Retrieval (MIR) Evaluation, were placed in July 2002 at the ACM/IEEE Joint Conference on Digital Libraries. The outcome of these workshops was the recognition by the Music IR community's of the creation of a periodic evaluation forum for Music Information Retrieval systems. The story of MIR evaluation has been shaped since then:

- During the 5th edition of the ISMIR in 2004, placed in Barcelona, Spain, an Audio Description Contest (ADC) [3] was accomplished. It proposed some tasks in order to define evaluation and statistical methods to compare systems.

- In 2005, the *MIREX* [4] *(Music Information Retrieval Evaluation eXchange)* run for the first time as the community-based framework for the formal evaluation of Music Information Retrieval (MIR) systems and algorithms. MIREX is coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana- Champaign.

- *MusiClef*, which run from 2011-2013 covered multimodal music tagging (Orio, Liem, Peeters, & Schedl, 2012) and focus evaluation on professional application scenarios.

[3] http://ismir2004.ismir.net/ISMIR_Contest.html

[4] http://www.music-ir.org/mirex/wiki/

- The *Million Song Dataset Challenge* (*MSD*, 2012) was created to overcome music dataset sharing limitations (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011). With this approach, researchers could grant access to a number of features but not to the algorithm that performs the process, neither to the audio.

- *Quaero-Eval*, inspired by NIST and MIREX evaluations, since 2012 focuses on audio and music processing. In this venue the tasks are agreed first with the participants and then a common repository is shared. The algorithms are run in a test sets with evaluation frameworks by an independent body that does not participate in the evaluation process.

- *MediaEval* [5], 2010, is a initiative focuses in multimodal approaches involving human and social aspects of multimedia e.g., speech recognition, multimedia content analysis, music and audio analysis, user-contributed information (tags, tweets), etc.

Fig 1 graphically encompasses both text and music evaluation initiatives.

[5] http://www.multimediaeval.org/about/

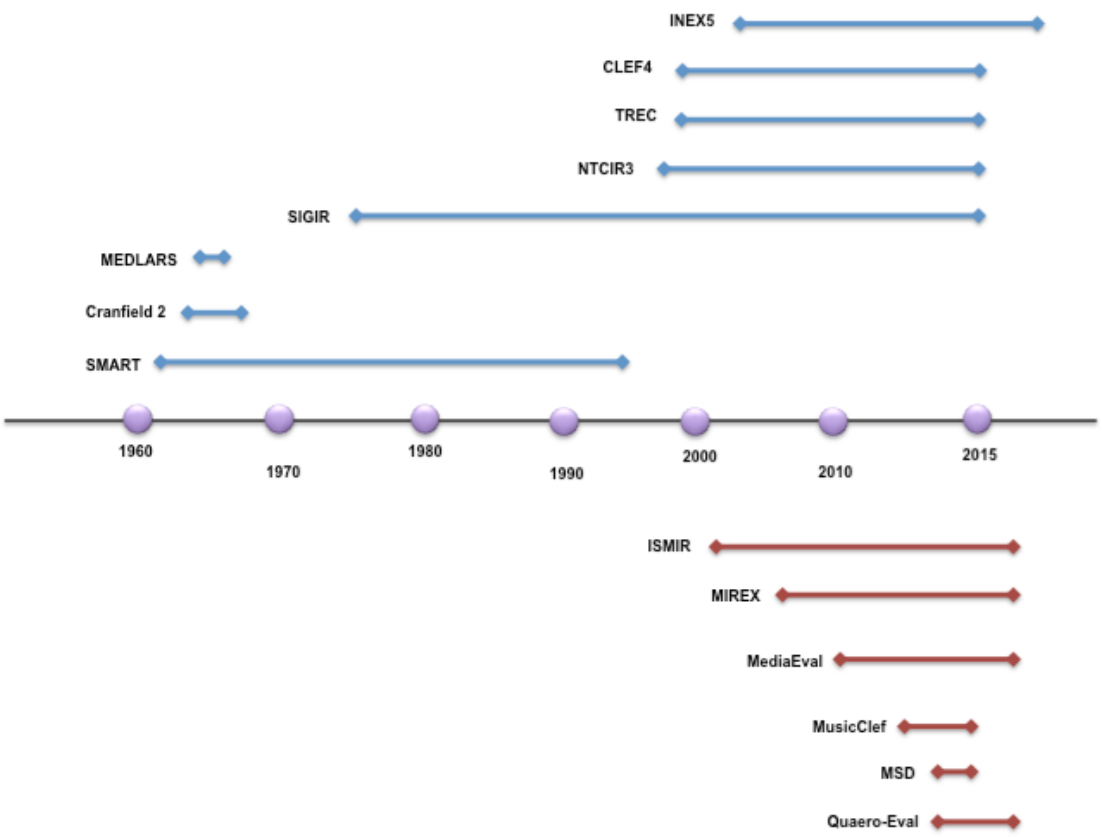

Fig 1. Timeline of Evaluation in Text IR (top) and Music IR (bottom).

1.3. Audio Music Similarity

Audio Music Similarity (AMS) deals with the challenge of discovering similar songs. It is generally used in MIR task such as music recommendation, playlist generation or plagiarism detection. AMS is one of the most important tasks in MIR and has participated in MIREX since 2006, evaluating so far 85 systems. Furthermore, the same document collection with 7,000 audio documents has been used since 2007.

In the context of MIREX this task resembles the Text IR scenario: for a given audio clip (the query), a system returns a list of songs from a corpus (candidate songs), sorted by their musical similarity to the query.

1.4 Importance of Evaluation in Music Information and Retrieval and Motivation

The Roadmap for Music Information Research, created for the expansion of the context of research from the perspectives of technological advances, stated as one of the main challenges: “vi) promoting best practice evaluation methodology, defining meaningful evaluation methodologies and targeting long-term sustainability of MIR (Serra, Magas, Benetos, Chudy, Dixon, Flexer, Widmer, 2013).

In spite of all initiatives created to widen the scope of evaluation, MIR community is still concern on the way that systems are evaluated because current evaluation practices do not fully allow them to improve as much as they wish (Peeters, Urbano, & Jones, 2012). Furthermore, research by (J Urbano, Schedl, & Serra, 2013), demonstrated that evaluation in ISMIR comprised only 6% of research.

MIREX has been a significant venue to convey the study and establishment of MIR evaluation frameworks; although it was created mirroring TREC methodologies,

eventually the Music IR community has realized that not everything from text applies to music. Also their evolution in time have been different; in text for instance, research in evaluation has produced an environment of continuous improvement, which has not been the case in Music IR. It seems that MIR community does not seem to pay as much attention as evaluation as it should.

Particularly, in the case of Audio Music Similarity, few studies about the influence of this TREC-like approach have been done.

The purpose of this thesis is to improve the evaluation process in Audio Music Similarity task in MIREX, studied from the perspective of efficiency with emphasis in the reduction of annotation cost.

The approach to follow is twofold: first, study the literature of low cost evaluation in Audio Music Similarity. Second, study models and methods in order to propose a new or improve the existing framework to estimate relevance judgments in Audio Music Similarity.

Chapter 2
STATE OF THE ART

2.1 MIREX Evaluation Process

2.1.1 The Cranfield Paradigm

MIREX provides an evaluation framework for MIR researches to compare, contrast and discuss the result of their algorithms and techniques in the same way than TREC has done it to the text Information retrieval community (Downie et al., 2014). In general, MIREX and TREC use test collection with evaluation measures in order to assess effectiveness of their systems. Test collection are a resource used to test and compare search strategies in a laboratory environment. They are composed by:

1. Collection of *documents* of significant size.

2. Tasks and/or *queries* to be performed on the test collections; and,

3. *Relevance judgments* (qrels) compose of a list of document/pair describing the relevance of documents to topics.

Test collection along with evaluation measures stipulates a simulation of users in a real searching environment. They are generally used by researchers for instance to asses retrieval systems in isolation helping finding failures inside their applications and comparing effectiveness among them.

In order to asses the performance of systems, both TREC and MIREX follow the Cranfield's paradigm which is a test bed consisting in a set of documents ***D***, a set of Information need statements or queries ***Q*** and a set of relevance judgments ***R***

that is compiled by human assessors ***H,*** which tell what documents should be retrieved for which query (ground truth). In Music Information Retrieval one of the task that emulate this behavior is Audio Music Similarity: for a given audio clip (the query), an AMS system returns a list of music pieces (documents) considered to be similar to it.

2.1.2. MIREX Evaluation in Audio Music Similarity

For the evaluation of system's effectiveness in the task of Audio Music Similarity in MIREX, relevance judgments and effectiveness measures are utilized. The relevance judgments in this context are scores given to each query-candidate, representing their similarity. In a real scenario, the task of collecting these judgments takes several days or weeks (J Urbano & Schedl, 2013)

In general terms, the evaluation process in MIREX runs as follows:

1. ~50 [6] queries ***Q*** are selected randomly and deliver to the participants.
2. The participant systems retrieved a ranked list with the 10^7 most similar pieces of music from a music collection ***D.*** These music pieces are 30-second audio clips of music material.
3. All the results are consolidated and evaluated this time using subjective judgments (ground truth) by human assessor using a software tool called "Evaluatron 6000" (E6K).
4. After listening to each query-candidate pair, graders were asked to rate the degree of similarity of the candidate audio excerpt to the query in two ways: a) By selecting one of the three BROAD categories of

[6] In past editions of MIREX 100 queries were used

[7] In past editions of MIREX, 5 similar musical pieces were retrieved

similarity: Not Similar (NS), Somewhat Similar (SS), and Very Similar (VS); and, b) By assigning a FINE[8] score between 0.0 (Least similar) and 10.0 (Most similar).

In the case of effectiveness measures, the one reported to assess effectiveness in Audio Music Similarity is **CG@10** (Average Gain after 10 audio documents retrieved) (Downie, Ehmann, Bay, Jones, 2010). For an arbitrary system *A*:

$$CG@k = \frac{1}{k} \sum_{i=1}^{k} G_i \qquad (1)$$

Where G_i is the gain of the *i-th* document (song) retrieved - the similarity scored assigned- by graders, using FINE or BROAD scale. After the process of judging is done, the mean score of the gains obtained for every executed query ranks the systems. In order to minimize random effects the Friedman test is run with the Average Gain score of every system, with the Tukey's HSD to correct the experiment-wide Type I error rate. The result of this evaluation is a scale-dependent pairwise comparisons between systems, telling which one is better for the current set of queries *Q*.

2.2 Validity, Reliability and Effectiveness

Validity, reliability and effectiveness are crucial aspects of testing. All IR evaluation experiments need to be guided considering them. This thesis work will be focus from the point of view of efficiency.

Validity is the extent to which the experiment actually determines what the experimenter wishes to determine (Tague-Sutcliffe, 1992). For example, are the

[8] In past editions of MIREX this value was between 0 and 100

selected variables really representatives of the experiment? Or in an evaluation experiment, is system A better than system B?

Reliability is the extent to which the experimental results can be replicated. (Tague-Sutcliffe, 1992). Thus, if an experiment is replicated, will we obtain similar results? There is a close relationship between validity and reliability. For example, if with one sample system A performs better than system B, but with a different sample is the opposite case, our results then cannot be repeatable; hence they will be unreliable.

Efficiency is the extent to which an experiment is effective (valid and reliable) (Tague-Sutcliffe, 1992). For instance, if in an evaluation experiment the ground truth annotation process is inaccurate, the validity of the result can be affected. On the other hand, if this ground truth is not efficient enough (as stated before, this process can be tedious and expensive) the reliability of the results may be impacted as well. Therefore, evaluation experiments must find a balance between validity and reliability and the efficient cost of the annotation process. In this context, do exist others experiments related to low cost annotations process to obtain valid and reliable results?

For this reason, searching in the present literature in response to the latter question is a must.

For example, some studies presents that judgments are affected by *many characteristics of retrieved records and users*, and also by situational factors (Harter, 1996). Therefore, some research shows that *crowdsourcing* is a viable alternative for the creation of relevance judgment; however because of the diversity in the backgrounds of participants, some control methods need to be established (Alonso, Rose, & Stewart, 2008).

Another approach to this matter is to *decrease the necessary number of judgments*. For example, in the *pooling method*, a set of top ***d*** ranked documents

returned by participating systems is selected to create the pool of documents that need to be judged (Spärk Jones & van Rijsbergen, 1975). Next, all the duplicates documents into the pool are eliminated (considered non-relevant) and the remaining ones are evaluated by assessors. TREC was the first event that used these partial relevance judgments. This technique has its drawbacks, for example, the existence of defective systems could affect the pooling methods and assessors can evaluate thousands of irrelevant items.

Some research is focus in how *evaluate systems with incomplete judgments* and still be confident with the results of the experiments. The idea is to use random variables to represent relevance judgments; the estimation of these values though, can have some degree of error and uncertainty, but also, for most documents they work pretty well.

Let G_i being a Random Variable representing the relevance level assigned to document *d*. It presents a multinomial distribution and depends of the scale used by human assessors.

The expectation and variances can be defined as random variables as well:

$$E\,[G_i] = \sum_{l\,\epsilon L} (G_i = l).\ l$$

$$\mathrm{Var}\,[G_i] = \sum_{l \epsilon L} (\,G_i = l)\,.\ l^2 - E[G_i] \qquad (2)$$

Every time a human assessor makes an annotation G_i then $E\,[G_i] \leftarrow g_i$ and $Var\,[G_i] = 0$; it means there is no uncertainty of G_i.

Research about incomplete judgments can be described as follows:

- (Buckley and Voorhees, 2004) investigated about *evaluation measures robust enough to cater for incomplete judgments;* this research introduced the need of a proper evaluation measure for large collections *bpref,* which

calculated system's scores having into account top non-relevant judgments rejected by the traditional pooling method.

- (Carterette, Allan, & Sitaraman, 2006) conduit an investigation about *Minimal Test collections for retrieval evaluation* which has lead into an algorithm that in minimal time evaluate retrieval systems with high degree of confidence and using a minimal number of judgments.
- (Aslam, Yilmaz, 2007) have shown that giving the average precision of a minimal fraction of judge documents using a small number of relevance judgments, the *relevance of the remaining unjudged documents can be inferred*.
- (Carterette, 2007) studied *Robust Test Collections for Retrieval Evaluation*, where a model able to achieve reusability with very small sets of relevance judgments per topic was presented.
- (Carterette & Allan, 2007) proposed the use of *inter-document similarity,* in which document similarity is the key to evaluate retrieval systems with more accurate and robust results, using 99% less relevance judgments than TREC conferences.

As stated before, research in text information retrieval has been meaningful for the creation of continuous improvement in evaluation techniques. In music, this topic has received about half of the attention but still the little research conducted so far, has been significant. For example, in order to create large datasets and reduce the number of annotations needed, low-cost evaluation alternatives have been explored. For instance, (J. Urbano & Schedl, 2013) applied Minimal Test Collection (MTC) algorithms to the evaluation of the Audio Music Similarity task in MIREX, which reduced the annotation cost to less than 5%. Therefore, the researches investigated how to compare systems when incompletes judgments are available and still be confident about the results. The idea is to model probabilistically the relevance judgments provided by human assessors using the same concept of random variables. Then, they created models to estimate these relevance

judgments as accurately as possible and obtain good estimates of systems effectiveness even with few available judgments.

Let G_i being a Random Variable representing the relevance level assigned to document *d*. If the scale is *Fine*, G_i can take one of three values and if this is Broad, it can take one of 11 values. To estimate the relevance of a document with (2) the *P* ($G_i = l$) needs to be known for each relevance level of *L* (the possible value giving by a human annotator using the scale *L*). It means, the distribution of G_i has to be calculated. The followed approach was the estimation of the relevance of each document individually, creating two models fitted with features about every query-document.

These features are:

i) ***Output-based:*** used when there are no judgments available; represents aspects of the system outputs. (See Table 1). For an arbitrary document *d (song)* and query *q (looking for similarity among songs).*

Feature	Description
fSYS	% of systems that retrieved d for *q*. If many systems return *d*, It's expected that *d* is more similar to *q*.
OV	Degree of overlap between systems
aRANK	Average rank in which systems retrieved *d* for *q*. Documents at the top are expected to be more similar to *q*
sGEN	Whether the musical genre of *d* is the same as *q*
fGEN	% of all documents retrieved for *q* that belong to the same musical genre than *d* does
fART	% of documents retrieved for *q* that belong to the same artist as *d* does

Table 1. Output-based features

ii) **Judgment-based features:** Utilizes known judgments (See Table 2)

Feature	Description
aSYS	Average relevance of documents retrieved by the system
aDOC	Average relevance of all the other documents retrieved for *q*
aGEN	Average relevance of all the documents retrieved for *q* that belong to the same genre as *d* does
aART	Average relevance of all the documents retrieved for *q* performed by the same artist as *d*
aART	% of documents retrieved for *q* that belong to the same artist as *d* does
aSYS	Average relevance of documents retrieved by the system
aDOC	Average relevance of all the other documents retrieved for *q*

Table 2. Judgment-based features

These models were created and fitted with data from the task of AMS in MIREX 2007,2009,2010 and 2011. Only those features that improve the model were selected. R^2 (coefficient of determination) was used to measure the variability of the predicted outputs, where a value of 1 means a perfect fit of the data by the model. Table 3 introduces these results.

Feature	Description	R^2 Broad	R^2 Fine
M_{jud}	fTEAM, OV, aSYS, aART	0.9156	0.9002
M_{Out}	fTEAM, OV, sGEN, fART, fGEN	0.3627	0.3439

Table 3. Features for the two models

Table 3 shows that M_{jud} presents good estimates. However, the estimation of G_i has to be calculated after judging some documents to obtain aSYS and aART. For this reason M_{Out} was created in order to estimate G_i even when there is no available judgments. As expected, the latter model performed worst than the former.

Table 4 presents statistics of all features for each model. Models for year *Y* are fitted to exclude all judgments for that year.

M_{out}										
Parameter	Broad					Fine				
	All	2007	2009	2010	1011	All	2007	2009	2010	1011
fSYS	123	91	53	121	93	140	107	70	137	97
OV	213	172	102	91	147	306	251	125	150	207
fSYS:OV	76	57	18	263	73	78	61	23	11	67
fART	295	191	257	319	133	283	174	276	290	125
sGEN	708	561	470	620	459	792	613	557	672	517
fGEN	2141	1428	1169	1888	2034	2313	1548	1250	2090	2148
sGEN:fGEN	279	174	92	263	328	478	321	183	447	496
R^2	0.3627	0.3459	0.3296	0.3780	0.4032	0.3439	0.3280	0.3175	0.3569	0.3786
RMSE	0.3254	0.3188	0.313	0.352	0.345	0.2412	0.2432	0.2341	0.2619	0.2501
Avg. *Var*	0.1054	0.1088	0.1121	0.0995	0.0989	0.0569	0.0577	0.0596	0.0538	0.0545

M_{jud}										
Parameter	Broad					Fine				
	All	2007	2009	2010	1011	All	2007	2009	2010	1011
fSYS	6	4	4	21	1	15	10	11	37	3
aSYS	144	103	104	115	106	109	74	88	89	75
aART	30810	23058	20753	26864	21705	41552	31337	28147	35913	29164
R^2	0.9156	0.9122	0.9089	0.9166	0.9245	0.9002	0.8987	0.8980	0.8991	0.9051
RMSE	0.1376	0.1301	0.1272	0.1427	0.1518	0.0922	0.091	0.0899	0.0936	0.0957
Avg. *Var*	0.0178	0.0167	0.0172	0.0175	0.0196	0.0069	0.0067	0.0069	0.0071	0.007

Table 4. Likehood-ratio Chi-squared (under the name of All) statistic of all features for each model, with R^2 scores, RMSE (Rooted Mean Squared Error) between predicted and actual scores, and average variance of estimates for M_{out} and M_{jud}. Adapted from (Julián Urbano, 2013)

Table 4 presents results that show that:

1) In the case of M_{Out} the best results come from features fART, fGEN and sGEN, in other words, from data related to artist and genre confirming that they are good features to estimate similarity between two music excerpts (Flexer and Schnitzer 2010). For M_{jud} the best results are originated by aART demonstrating that if two songs from two artists are similar, other songs from them tend to be similar as well. This case represents the decision of MIREX to filter out songs that share the same artist than the query because they are likely to be similar.

2) RMSE and Average Variance demonstrate how well these models estimate relevance judgments. For a better comparison across scales, they were normalized between 0 and 1, resulting in *Broad* = {0; 0.5; 1} and *Fine* = {0.05; 0.15; ... 0.95}. It can be noticed that Fine scale makes better estimation of relevant judgments.

3) Although M_{jud} performs better than M_{out}, this one can still be used because its estimation's error can be compared to the differences expected when human assessors performs relevance judgments.

Then, after creating the probabilistic estimation of relevance judgments using random variables, effectiveness scores used to rank systems according to their performance in the evaluation of AMS, needed to be predicted using random variables as well. Therefore, three possible scenarios to use according to the evaluation needs were set. In the implementation of this scenarios data from MIREX 2007, 2009, 2010 and 2011 was used. The results demonstrated that:

i) In the first scenario, when there are not relevance judgments available, M_{out} can be used and the order of systems is estimated with an average accuracy of 92% and with an average confidence in the rankings of 94%.
ii) In the second scenario, when the goal is estimate system's differences, it showed that just using 2% of the judgments (estimating the other 98%) the differences could be correctly estimate in 93% of the cases.

iii) In the third scenario, when the focus is the estimation of absolute scores, just with 25% of the relevance judgments they can estimate with an error of +-0.05. In this last scenario, effectiveness in the ranking of systems is highly overestimated. One approach to correct this issue was the use of a threshold of variance as a practical correction factor to use in the stopping condition. As a consequence, the error was reduced but at the expense of making several judgments, (between 15% and 35%). Fig 2 present this situation:

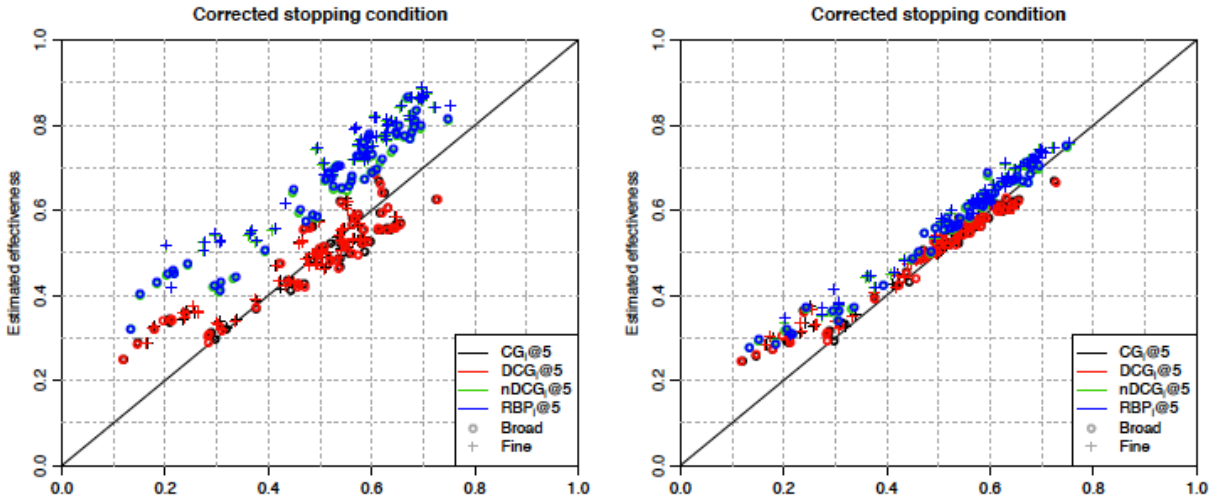

Fig 2. Estimated vs. actual absolute effectiveness scores in MIREX 2007, 2009, 2010 and 2011 when judging documents until expected error is +-0.05 with an uncorrected (left) or corrected (right) stopping condition. Adapted from (Urbano, 2013)

The objective of this thesis project is to improve the existing probabilistic framework in order to get better estimates of relevance, which is intended to improve the predictions of the ranking of systems, reducing the amount of needed judgments observed in Figure 2.

Chapter 3
IMPROVING THE ESTIMATION OF RELEVANCE

Estimating relevance judgments will reduce the annotation cost yet achieving better predictions of effectiveness measures of systems. After reviewing the literature and the available models for prediction, several approaches have been considered to obtain better estimations:

1. Using others configurations of Ordinal Logistic Regression models.

2. Implementing others probabilistic models.

3. Improving model's attributes.

4. Implementing new attributes for models.

Each approach and its corresponding results would be described as follows; data from past edition of MIREX was used:

1. Using others configurations of Ordinal Logistic Regression models

From the literature, (Urbano, 2013) used the regression framework with ordinal logistic regression as the main approach to predict relevance since it takes into account the order of relevance level. Using the statistical language R [9], two distinct configurations of ordinal models were tried inside the aforementioned

[9] http://www.r-project.org

framework: packages *rms* and *MASS.* The results of this implementation are depicted in Table 5.

ORDINAL LOGISTIC REGRESSION

Packet	Model	Scale	R^2	Orig. R^2	%	RMSE	Orig. RMSE	%	Var	Orig. Var	%
rms	Mout	Broad	0,3630	0,3627	**0,08%**	0,3254	0,3254	**0,00%**	0,1054	0,1054	**0,00%**
		Fine	0,3430	0,3439	**-0,26%**	0,2833	0,2412	**17,45%**	0,0178	0,0178	**0,00%**
	Mjud	Broad	0,9160	0,9156	**0,04%**	0,1375	0,1376	**-0,07%**	0,0177	0,0178	**-0,84%**
		Fine	0,9060	0,9000	**0,67%**	0,0900	0,0922	**-2,39%**	0,0069	0,0069	**0,00%**
MASS	Mout	Broad	**This packet does not show R**			0,3258	0,3254	**0,12%**	0,1054	0,1054	**0,00%**
		Fine				0,2408	0,2412	**-0,17%**	0,0177	0,0178	**-0,56%**
	Mjud	Broad				0,1375	0,1376	**-0,07%**	0,0177	0,0178	**-0,56%**
		Fine				0,0900	0,0922	**-2,39%**	0,0069	0,0069	**0,00%**

Table 5. Implementation of rms and MASS packages for Ordinal Logistic Regression in R. Columns Orig. R^2, Orig. RMSE and Orig. Var. represent the values obtained from (Urbano, 2013). MASS package does not show the value of the coefficient of determination, R^2.

Table 5 presents that using these configurations of ordinal models the improvement in the results were minimum. For example, proving *rms* for M_{out} the coefficient of determination R^2 just increased in a 0,08% for the Broad scale and decrease in -0,26% in the case of Fine. For RMSE in Broad, no improvement was achieved and for Fine, the error increased 17%. Respect to the variance, any scale presented an improvement. For M_{jud} the results were minimum: For R^2 it got 0,04% for Broad and 0,60% for Fine scales; for RMSE and variance, a minimum improvement was achieved. Using MASS package the results enhanced in a minimal amount as well. Hence, using other configurations of Ordinal Logistic Regression did not achieve significant improvements for the prediction of relevance.

2. Implementing others probabilistic models in order to obtain better results

The reviewed literature considered that linear regression was not an appropriate approach because the predicted relevance could be outside the [0,*nL*−1] limits (Urbano, 2013). However if the results can be truncated inside the possible values of Broad and Fine scales, this issue can be addressed. To prove this hypothesis models from the Generalized Linear Models, which can represent categorical, binary and other response types were tested: linear, probit and logit regressions. For probit and logit, the estimated relevance values need to be first mapped inside the range [0-1] and in order to interpret the results, these values need to be transformed back to the original scales; Table 6 presents the results of the evaluation using these models:

REGRESSION

Model	Model Fitted	Scale	RMSE	Orig. RMSE	%	Var	Orig. Last	%
Logit Regression	Mout	Broad	0,3720	0,3254	**-12,53%**	0,0031	0,1054	**-97,06%**
		Fine	0,2416	0,2412	**-0,17%**	0,0150	0,0177	**-15,25%**
	Mjud	Broad	0,2997	0,1376	**-54,09%**	0,0148	0,0177	**-16,38%**
		Fine	0,0961	0,0922	**-4,06%**	0,0148	0,0069	**114,49%**
Probit Regression	Mout	Broad	0,3715	0,3254	**-12,41%**	0,0009	0,1054	**-99,15%**
		Fine	0,2415	0,2412	**-0,12%**	0,0006	0,0177	**-96,61%**
	Mjud	Broad	0,2950	0,1376	**-53,36%**	0,0028	0,0177	**-84,18%**
		Fine	0,0952	0,0922	**-3,15%**	0,0004	0,0069	**-94,20%**

Table 6. Implementation of Logit and Probit regression using VGAM package in R. Columns Orig. RMSE and Orig. Var. represents the values obtained from (Urbano, 2013). This package does not present R^2

Logit and Probit regression did not improve the prediction of relevance as it is demonstrated by RMSE and Variance results.

In the case of Multiple Linear Regression, the predicted values sometimes do not

fall in a range within [0-2] for Broad or [0-100] for Fine scale; in all these cases they need to be truncated inside the corresponding scale in order to obtain the correct mapping with the estimates values. Table 7 presents these results:

MULTINOMIAL REGRESSION

Model	Model Fitted	Scale	R2	Orig. R2	%
Multiple Linear Regression	Mout	Broad	0,3320	0,3627	**-1,11%**
		Fine	0,3557	0,3439	**0,41%**
	Mjud	Broad	0,8824	0,9156	**-3,04%**
		Fine	0,9114	0,9002	**1,01%**

Table 7. Implementation of Multinomial Linear Regression Columns Ant. R^2, Ant. represents the values obtained from (Urbano, 2013).

For the case of M_{out} and Mj_{ud} using the Fine scale, an improvement in the coefficient of determination $R^{2,}$ of 0,4% and 1% was achieved. The rest of predictions did not get any improvement.

3. Improving model's attributes

To improve the prediction power of independent variables some techniques can be applied. For example, implementing a selection method, which is intended to choose the best subset of predictors (Faraway, 2004). For both models M_{out} and M_{jud}, backward elimination approach was applied. This method start testing the interaction of all predictors (features, attributes) and then removes the predictors with the less or the highest value of some parameter, depending of the model (higher p-value, R^2, lowest deviance or AIC, etc). In this case, *rms* packet in R was used with ordinal logistic regression, starting with the interaction of all the

predictors; therefore, in order to decrease the number of permutations, the selection of variables made by (Urbano, 2013) was followed for M_{out} (fSYS, OV, sGEN, fGEN, fART). Using the Deviance as an indicator of quality of good or bad fit for the model, the results are presented in *Table 8*. It shows that even thought the best fit was achieved by the Interaction number 1 and 5, the latter can be selected since it does not used as many parameters as the former. Furthermore, if this interaction is compared with the research from Urbano, the result is almost the same, so this last configuration can be chosen since is less complex than the one presented in interaction 5. Similar results were obtained for $M_{jud.}$

MODELS TRIALS			
Model	**Predictors**	**Deviance**	**AIC**
1	fSYS * OV + fSYS * sGEN + fSYS * fGEN + fSYS * fART + OV * sGEN + OV * fGEN + OV * fART + sGEN * fGEN + sGEN * fART + fGEN * fART	36.342	36.376
2	fSYS + OV + sGEN + fGEN + fART	37.203	37.217
3b	fSYS+OV	44.532	44.540
3c	fSYS+sGEN	40.483	40.491
3d	fSYS+fGEN	38.477	38.485
3e	fSYS+fART	43.346	43.354
3j	OV+sGEN	40.693	40.701
3k	OV+fGEN	38.416	38.424
3l	OV+fART	43.534	43.542
3m	sGEN+fGEN	38.049	38.057
3n	sGEN+fART	39.426	39.434
3o	fGEN+fART	38.031	38.039
4a	fGEN	38.384	38.394
4b	sGEN	40.424	40.434
4c	fSYS	44.536	44.542
4e	OV	44.910	44.916
4f	fART	43.536	47.685
5	fSYS*fGEN + OV*fGEN + sGEN*fGEN + fGEN*fART	**36.401**	**36.423**
Urbano, 2013			
fSYS*OV + fART + sGEN*fGEN		**36.837**	**36.855**

Table 8. Implementation of backward elimination of predictors for $M_{out.}$

4. Implementing new attributes

4.1 Cluster of Genres

Another considered approach was the use of a new independent variable called *Cluster;* it was intended to improve the results of relevance's predictions by clustering genres according to subjective criteria of similarity. The genres used in MIREX are (10):

Baroque, Blues, Classical, Country, Edance, Jazz, Metal, RapHiphop, Rock-and-roll, Romantic. After listening to several songs of each genre from the provided MIREX dataset, the proposed clustering of genres is described in Table 9:

Genres	Cluster
Baroque-Classical-Romantic	***Cluster 1:*** Classical
RapHiphop - Edance	***Cluster 2:*** Electronic
Blues-Rockandroll-Country	***Cluster 3:*** Romantic
Jazz	***Cluster 4:*** Jazz
Metal	***Cluster 5:*** Metal

Table 9. Proposed clustering of genres of MIREX's data.

Adding this new attribute to the model using a dichotomous binary variable where the value of 1 was assigned if the query had the same genre as the document or 0 otherwise, the results were depicted in Table 10:

Model Fitted	Scale	R^2 Genre Clustering	Original R^2	% Difference
Mout	Broad	0,4030	0,3620	**11,3%**
	Fine	0,3840	0,3430	**12,0%**
Mjud	Broad	0,9150	0,9156	**-0,07%**
	Fine	0,9050	0,9000	**0,6%**

Model Fitted	Scale	RMSE Genre Clustering	Original RMSE	% Difference
Mout	Broad	0,3170	0,3254	**-2,6%**
	Fine	0,2480	0,2412	**2,8%**
Mjud	Broad	0,1370	0,1376	**-0,4%**
	Fine	0,0900	0,0922	**-2,4%**

Model Fitted	Scale	Var. Genre Clustering	Original Var.	% Difference
Mout	Broad	0,1000	0,1054	**-5,1%**
	Fine	0,0630	0,0569	**10,7%**
Mjud	Broad	0,0170	0,0179	**-4,8%**
	Fine	0,0070	0,0069	**1,4%**

Table 10. Implementation of the new attribute Cluster into a Logistic Regression Model. Columns Original R^2, Original RMSE and Original Var. represent the values obtained from (Urbano, 2013).

Table 10 presents that using the new attribute *Cluster* for M_{out} the results were improved for the Broad scale in an 11% and in a 12% for Fine. In the case of M_{jud} there were not improvements.

4.2 Using the distances' media of similarity between genres

Therefore, adding new attributes in order to improve the prediction of relevance was a good choice to obtain better results. For this reason, another attribute formed using the media of the similarity's distances between the genre of the query and the genre of the document (song) called *Distance* was implemented; with this new feature, one is expected to get better results. Table 11 presents the

aforementioned distances between genres and Table 12 introduces results using this new attribute.

genreq	genred	Distance Similarity
Jazz	Jazz	62.918
Metal	Metal	61.092
Classical	Classical	58.618
Electronic	Electronic	56.681
Romantic	Romantic	48.745
Romantic	Jazz	37.284
Metal	Romantic	37.084
Romantic	Metal	36.849
Jazz	Romantic	36.288
Metal	Electronic	28.107
Electronic	Metal	23.881
Electronic	Romantic	19.956
Jazz	Electronic	17.156
Classical	Jazz	16.740
Electronic	Jazz	15.982
Jazz	Classical	15.919
Romantic	Electronic	13.430
Metal	Jazz	13.000

Table 11. Media of distances between genres of queries and songs.

		R^2 with distances (similarity)			
Model Fitted	**Scale**	**R^2 no genre clustering**	**R^2 with genre clustering**	**Original R^2**	**% between highest and original**
Mout	Broad	**0,4670**	0,4340	0,3620	**29%**
	Fine	**0,4490**	0,4210	0,3430	**31%**
Mjud	Broad	**0,9150**	0,9150	0,9156	**0%**
	Fine	**0,9050**	0,9050	0,9000	**1%**

RMSE with distances (similarity)

Model Fitted	Scale	RMSE no genre clustering	RMSE with genre Clustering	Original RMSE	% Between highest and original
Mout	Broad	**0,3030**	0,3110	0,3254	**-7%**
	Fine	**0,2210**	0,2270	0,2412	**-8%**
Mjud	Broad	**0,1370**	0,1370	0,1376	**0%**
	Fine	**0,0900**	0,0900	0,0922	**-2%**

Variance with distances (similarity)

Model Fitted	Scale	Var no genre clustering	Var with genre clustering	Original Var	% Between highest and original
Mout	Broad	**0,0910**	0,0970	0,1054	**-14%**
	Fine	**0,0500**	0,0530	0,0563	**-11%**
Mjud	Broad	**0,0170**	0,0170	0,0178	**-4%**
	Fine	**0,0070**	0,0070	0,0069	**1%**

Table 12. Implementation of distance as attribute into a Logistic Regression Model. Columns Original R^2, Original RMSE and Original Var. represent the values obtained from (Urbano, 2013). Columns R^2, RMSE and Var. with genre clustering represents values obtained when the cluster of genres of Table 9 was performed. Columns R^2, RMSE and Var. with no genre clustering presents the values without this clustering, using the genres proposed from the original data.

While trying different features' interactions using the new attribute *Distance*, one experiment was conducted where no clustering of genres was implemented and the original classification of genres from MIREX were used instead. It leaded to get even better results: a significant improvement in R^2 of 29% for the Broad scale and of a 31% for the Fine scale for M_{out}. As Table 12 presents, using attribute *Distance*, with clustering of genres the results are still meaningful. Also RMSE and Variance values were improved.

These results proved that adding new attributes as independent variables is an optimal path to take for improving the estimation of relevance.

4.3. Using metadata to deal with artist information

After obtaining good results when treating information related to genres, something similar was performed using information from artist. With the restriction that neither the query, nor the document can belong to the same artist, any similarity's media measurement can be calculated from the data from MIREX; for this reason, the use of an external source was necessary. The idea is to contrast the information provided from MIREX's metadata along with information from a music Internet database for instance, in order to look for similarity between artists. The provided metadata file contained information of track artist, album artist and genre. Then several steps were followed:

1. The selected Internet database was Echo Nest[10] because it allows the access to billion of music data points from media and mobile companies like (MTV, BBC, MOG, Pocket Hipster, etc.). Its API [11] provides methods to return a wide range of data from many artists; for the particular case, similarity information.

2. Then, setting the API and coding with Phyton[12] to compare information from the metadata file versus those artists that exist in Echo Nest, a dataset with a list of artist similarities was obtained.

3. Subsequent, this information was integrated with the data provided from MIREX and using a new attribute called *Similarity*, new interactions of features for the regression model was tested.

The obtained results did not improve existing scores. Table 13 presents these outcomes. Any improvement neither for $R^{2,}$ nor for RMSE or Variance were

[10] http://the.echonest.com
[11] http://developer.echonest.com/docs/v4/artist.html#similar
[12] https://www.python.org

achieved. Therefore, the best scores gained so far were gotten using distance similarity without clustering between genres.

ARTIST SIMILARITY

Model Fitted	Scale	R2	R2 no Genre Clus.	%	RMSE	RMSE no Genre Clus.	%	Var	Var RMSE no Genre Clus.	%
Mout	Broad	0,4350	0,4670	**-6,85%**	0,311	0,3030	**2,64%**	0,0968	0,0910	**6,37%**
	Fine	0,4230	0,4490	**-5,79%**	0,2272	0,2210	**2,81%**	0,0535	0,0500	**7,00%**
Mjud	Broad	0,9150	0,9150	**0,00%**	0,1379	0,1370	**0,66%**	0,0178	0,0170	**4,71%**
	Fine	0,9050	0,9050	**0,00%**	0,0902	0,0900	**0,22%**	0,0070	0,0070	**0,00%**

Table 13. Implementation of the attribute artist similarity into a Logistic Regression Model.

CONCLUSIONS

This thesis work studies models and methods in order to improve the framework to estimate relevance judgments of Audio Music Similarity in the context of MIREX, from the point of view of efficiency. After reviewing the literature and existing models for predictions (M_{out}, that predict gain scores when no judgments are available and Mj_{ud} that improves the predictions when judgments are available), several approaches were considered in order to obtain better results.

First, *others configurations of Ordinal Logistic Regression* models were considered. Two packages or statistical programming language R were used. The results did not achieve significant improvement for the prediction of relevance. Second, the *implementation of others probabilistic models* were performed: probit, logit and linear regressions. In all these cases the estimated values needed to be first mapped inside an specific range and then transformed back to the Broad and Fine scale in order to compare. A slight improvement were achieved in R^2 for both models, using Fine scale, 0,4 % for M_{out} and 1% for Mj_{ud}. Third, *improving model's attributes* with the implementation of backward elimination was applied. After testing many interactions of attributes, a simplest configuration was selected; however, because it was similar and slight complex than the one obtained by (Urbano, 2013), this last one was considered instead. The fourth approach was *implementing new attributes for models*. Several experiments were conducted: i) Clustering subjectively the existing genres from the data of MIREX. Adding a new attribute *Cluster* to the logistic regression model, the predictions improved for M_{out} model and the Broad scale in an 11% and in the Fine scale, in a 12%. ii) Using distance's media of similarity between genres as a new attribute, the results improved better than before; performing this trial ignoring the clustering from i), the results gained an improvement in R^2 of 29% for the Broad scale and of a 31% for

the Fine scale for M_{out} model. Also RMSE and Variance values were superior. i) and ii) proved that adding new attributes as independent variables is an optimal path to take for improving the estimation of relevance. For this reason, the last experiment performed, using this time information from artist was iii) Using metadata to deal with artist information. Because artist information has restrictions inside MIREX contest, it cannot be calculated with normal statistical procedures as genre; an external source had to be cast-off in order to get similarity measures. Echo Nest music Internet Database was chosen to get a similarity database of those artists belonging to a metadata file provided from MIREX. Unfortunately, the results did not improve existing scores.

Overall, the results of this dissertation's experiments indicate that attributes obtained from information such as the outcome of systems or metadata, conduits to improve the prediction of relevance. The best results are obtained for M_{out} over M_{jud}, which in most cases resembles the real scenario, when no judgments or just a minimum amount of them are available to make predictions.

FUTURE WORK

There are some lines of research arising from this work that can be pursued. First, in order to decrease the Variance of the predictions, the models can be trained with different amount of information. It will permit a better fit since at the present time some features are calculated using all the judgments, when in real life, this scenario is not always present. Second, it suggested to keep on improving the prediction of relevance, studying the role of attributes originated from artist or genre information, using others music services and sources of information.

Conclusively, with this dissertation, the author expects to encourage research in methods of low- cost evaluation not just for Audio Music Similarity, but also for the other task in Music Information and Retrieval.

REFERENCES

1. Aslam, J. a, Yilmaz, E., (2007) Inferring Document Relevance from Incomplete Information 633–642.

2. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million-Song Dataset. *Ismir*, (Ismir), 591–596. Doi: 10.1145/2187980.2188222.

3. Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 25–32. Doi: 10.1145/1008992.1009000

4. Buettcher S., Cormack G. V. and Clarke, C. L. (2010). Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press.

5. Cleverdon, C. W. (1991). The Significance of the Cranfield Tests on Index Languages. In International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3-12.

6. Carterette, B. (2007). Robust Test Collections for Retrieval Evaluation. Evaluation, 55–62. Doi: 10.1145/1277741.1277754

7. Carterette, B., & Allan, J. (2007). Semiautomatic evaluation of retrieval systems using document similarities. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management - CIKM ’07, 873. Doi: 10.1145/1321440.1321564

8. Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. Proceedings of the 29th Annual International ACM SIGIRConference on Research and Development in Information Retrieval - SIGIR ’06, 268. Doi: 10.1145/1148170.1148219

9. Downie, J. S. (2002). Interim Report on Establishing MIR/MDL Evaluation Frameworks: Commentary on Consensus Building. ISMIR Panel on Music Information Retrieval Evaluation Frameworks, 43–44.

10. Downie, J.S., Ehmann, A.F., Bay, M., Jones, M.C.: The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. In: W.R. Zbigniew, A.A. Wieczorkowska (eds.) Advances in Music Information Re-trieval, pp. 93{115. Springer (2010)

11. Downie, J. S. (2003). Music information retrieval. Annual review of information science and technology, 37(1), 295-340.

12. Downie, J. S., Hu, X., Lee, J., Ha, C. K., Cunningham, S. J., & Yun, H. (2014). 15th International Society for Music Information Retrieval Conference (ISMIR 2014). Ten years of Reflections, challenges and opportunities. (Ismir), 657–662.

13. Flexer, A., & Schnitzer, D. (2010). Effects of Album and Artist Filters in Audio Similarity Computed for Very Large Music Databases. Computer Music Journal, 34(3), 20–28. Doi: 10.1162/COMJ_a_000028. Cited on pages 104 and 106.

14. Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47(1), 37–49. Doi:10.1002/(SICI)1097-4571(199601)47:1<37::AID-ASI4>3.0.CO;2- 3

15. Faraway, J. (2004). Linear Models with R. Chapman and Hall/CRC.

16. Kent, A., Lancour H., Daily E. Encyclopedia of Library and Information Science: Volume 28 The Smart System to Standards for Libraries. CRC Press (1980). ISBN 9780824720285 - CAT# DK2544. 512 pages. Paperback. Series: Library and Information Science Encyclopedia.

17. Lancaster, F. (1968). Evaluation of the MEDLARS Demand Search Service. Technical report, U.S. Department of Health, Education, and Welfare.

18. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge university press.

19. Moghadasi, S. I., Ravana, S. D., & Raman, S. N. (2013). Low-cost evaluation techniques for information retrieval systems: A review. Journal of Informetrics, 7(2), 301–312. Doi: 10.1016/j.joi.2012.12.001.

20. Orio, N., Liem, C. C. S., Peeters, G., & Schedl, M. (2012). MusiClef: Multimodal music tagging task. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7488 LNCS*, 36–41. Doi: 10.1007/978-3-642-33247-0_5.

21. Peeters, G., Urbano, J., Jones, G.J. (2012). Notes from the ISMIR 2012 late-breaking session on evaluation in music information retrieval. In International society for music information retrieval conference.

22. Sanderson, M., & Croft, W. B. (2012). The history of information retrieval

research. Proceedings of the IEEE, 100(SPL CONTENT), 1444–1451. Doi: 10.1109/JPROC.2012.2189916

23. Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., ... Widmer, G. (2013). *Roadmap for Music Information ReSearch*. Retrieved from http://www.ofai.at/research/impml/projects/MIRES_Roadmap_ver_1.0.0.pdf.

24. Spärk Jones, K., & van Rijsbergen, C. J. (1975). Report on the need for and provision of an 'ideal' information retrieval test collection (British library research and development report no. 5266). Cambridge: Computer Laboratory, University of Cambridge. (p. 43)

25. Urbano, J., Universidad Carlos III de Madrid Tesis Doctoral Evaluation in Audio Music Similarity. (2013).

26. Urbano, J., & Schedl, M. (2013). Minimal test collections for low-cost evaluation of audio music similarity and retrieval systems. International Journal of Multimedia Information.... Retrieved from http://link.springer.com/article/10.1007/s13735-012- 0030-4

27. Urbano, J., & Martín, D. (2013). On the Measurement of Test Collection Reliability Categories and Subject Descriptors. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '13, 393–402.

28. Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval, (November 2012). Retrieved from http://link.springer.com/article 10.1007/s10844-013-0249-4. Cited on page 3.

29. Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, *28*(4), 467–490. Doi: 10.1016/0306-4573(92) 90005-K