POLITECNICO DI MILANO
Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria e Design del suono
Dipartimento di Elettronica e Informazione

A music search engine based on semantic text-based queries

Supervisor: Prof. Augusto Sarti
Assistant supervisor: Dr. Massimiliano Zanoni

Master graduation thesis by:
Michele Buccoli, ID 770967

Academic Year 2011-2012

POLITECNICO DI MILANO
Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria e Design del suono
Dipartimento di Elettronica e Informazione

Motore di ricerca di brani musicali basato su query testuali semantiche

Relatore: Prof. Augusto Sarti
Correlatore: Dr. Massimiliano Zanoni

Tesi di Laurea di:
Michele Buccoli, matricola 770967

Anno Accademico 2011-2012

A Roberta
perché mi ha salvato senza averne intenzione

Abstract

The advent of digital era has dramatically increased the amount of music that people can accede to. Millions of songs are just "a click away" from the user. This comes with several side-effects: the absence of mediators for music suggesting, the issues on browsing such a huge amount of music and retrieving of songs. Such large collection of songs makes them difficult to be navigated through classical meta-information such as title, artist or musical genre. Listening to songs without knowing any information about them raises the issue of retrieving meta-information of a track by describing its content.

Music Information Retrieval is the multidisciplinary research field that deals with extraction and processing of information from music. Information can be related to emotional or non emotional aspects of music, can provide an objective or subjective description, can be automatically computed or manually annotated by human listeners. In the latest years, Music Information Retrieval research community has proposed solutions for the issues mentioned above.

In this work we propose a music search engine that deals with queries by natural language semantic description and text-based semantic example. The former involves a description of songs by means of words. The latter concerns the retrieval of songs with a semantic description similar to some proposed examples. Both emotional and non emotional-related semantic description are allowed. We exploit the Valence-Arousal mapping to model affective words and songs. We introduce a set of semantic non-emotional high-level descriptors and we model them in a space we defined *semantic equalizer*. Songs are mapped in the semantic equalizer as well. Song similarity based on high-level similarity (emotional and non emotional) has been implemented. A natural language processing module is present in order to capture the distinctions of meaning that human language currently adopts.

This system can solve the problem of retrieving music among large music libraries. It is also suitable for music recommendation and music browsing

purposes. Indeed, we implement a prototype able to list results for a semantic query or to organize them in a playlist fashion.

The system has been tested with a proven data set composed by 240 excerpts of 15 seconds each. A questionnaire about the system's performances and utility has been proposed to 30 subjects. We obtained good rates on system's performances and the subjects positively evaluate the usefulness of this kind of system. These are promising results for future progresses.

Sommario

L'avvento dell'era digitale ha aumentato drasticamente la quantitá di musica accessibile online. Milioni di canzoni sono a portata di click. Ció ha portato dei risvolti negativi: stanno scomparendo i mediatori che consiglino nuova musica agli utenti ed é sempre piú difficile gestire le proprie collezioni musicali.

Il Music Information Retrieval (MIR) é un campo di ricerca multidisciplinare che studia l'estrazione e l'utilizzo di informazioni musicali. Tali informazioni possono essere di natura emozionale o non emozionale, possono descrivere la musica ad alto o basso livello, possono essere calcolate automaticamente o annotate manualmente. Una grande quantitá di musica ne rende difficile la navigazione, in quanto le classiche meta-informazioni come titolo del brano, artista o album sono insufficienti per descriverne il contenuto musicale e le sue caratteristiche. L'ascolto casuale di canzoni senza conoscerne alcuna informazione (sempre piú frequente, nei nuovi scenari aperti dall'aumento di musica disponibile), hanno introdotto il problema di ritrovare meta-informazioni di un brano cercando di descriverne il contenuto. Negli ultimi anni la ricerca scientifica MIR ha proposto diverse soluzioni per i problemi sopraccitati.

In questa tesi proponiamo un motore di ricerca musicale che gestisce query semantiche in linguaggio naturale ed esempi musicali basati su testo. La descrizione semantica puó essere sia emozionale che non emozionale. Abbiamo utilizzato un mapping nel piano di Valence-Arousal per modellare parole affettive e canzoni. Abbiamo poi introdotto un insieme di descrittori non emozionali di alto livello e li abbiamo modellati in uno spazio che abbiamo chiamato *equalizzatore semantico*. Le canzoni sono state analogamente modellate su questo spazio. É stato implementato anche un sistema di similaritá di alto livello tra canzoni (sia emozionale che non emozionale). Abbiamo inoltre inserito un modulo di processamento del linguaggio naturale per catturare le sfumature di significato tipicamente adottate nel linguaggio umano.

Questo sistema puó risolvere il problema di ritrovare musica in librerie musicali molto grandi. Il sistema é anche adatto a suggerire nuova musica e navigare tra le canzoni. Abbiamo infatti implementato un prototipo che, data una query semantica, restituisce i risultati sotto forma di lista di canzoni o di playlist musicale.

Il sistema é stato testato con un data set composto da 240 segmenti di 15 secondi ciascuno e proposto a 30 soggetti per una valutazione soggettiva. Abbiamo ottenuto buoni risultati sul funzionamento del sistema e risposte incoraggianti sull'utilitá percepita di un sistema come il nostro che apre scenari ottimisti per sviluppi futuri.

Acknowledgements

I would like to thank:

Professor Augusto Sarti for making possible for me to realize such an exciting work and to study such interesting subjects on my master degree.

Dott. Massimiliano Zanoni, for the topic of the thesis, for suggestions, for ideas, for the pacience (expecially for the pacience) and the long-term support.

The ISPG laboratory, an amazing ambient where develop anything. Paolo, Dejan, Antonio, Eliana, Ambra, Simone, Bruno and Stefano have been the best colleagues I could hope to have.

All my colleagues from the Sound and Music Engineering Master Degree, in particular my mates Matteo, Mario, Michele, Giorgio, Giuseppe, Saranya, Giammarco, Giorgia the outsider and many more. You help me, support me and make me discover the beer.

All of my friends who support and stand me during this work, fill my tests and read my thesis for suggestions: Valentina for providing the *porceddu*, Marianna for taking vacation to partecipate at the big day and Francesca for her *wonderfulness*.

My family who always supported me in what I was doing even if they did not realize what is it about.

All the people who help me filling my surveys and my test.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Music has always had an important role in the human life. The advent of the digital era has considerably increased the amount of music content that users can accede to. Most part of the music published in the last century is on sale online. Thousands or even millions of songs can be stored in media storage. The amount of music currently available is more than a person can listen to in an entire life. This information overload leads to an evolution in music listening experience. Until today mediators, such as music dealers and music magazines, have played an important role in collecting, organizing, retrieving and suggesting music for people. Users can now directly access to music via Internet and mediators risk to disappear. The music organization reflects the taxonomy used by mediators to classify music and is based on meta information such as title, album name, artist, year and so on. Nevertheless, people have recently been using to reach and listen to music without knowing any information about it. Users habits are changing: on one side people keep listening to music by a certain artist or belonging to a certain genre; on the onther side, they listen to music that inspire a certain mood or that has a certain sound.

New applications and paradigms are needed to collect, suggest, organize and retrieve music for people. Scientific community and industries are working to build automatic mediators that can address these issues. In order to realize this, it is crucial to investigate about music content, how to model it and which aspects are significant for representing it. On the other side, it is important to investigate how users understand music content, which description they use for it and how they would like to access it. For example, one of the main prerogative of music is to inspire feelings to listeners. People use to choose their music according to the mood they are feeling or in order to be brought to a certain emotion. Nevertheless, other non-emotional factors are taken into account when listening to music, such as musical genre (rock, pop), rhythmic aspects (slow, fast, dynamic) or other generic descriptors used in common language (dull, easy, catchy, noisy).

The gap between users and music content must be filled in order to build novel mediators. *Music Information Retrieval* (MIR) community studies and investigates elements involved in users description and music content. MIR is a multidisciplinary research field that deals with the retrieval and processing of information from music. MIR disciplines include: musicology, psychology, psychoacoustic, academic music study, signal processing, computer science and machine learning. Music information can be formalized and described hierarchically from lower level, which is related to sound content, to higher level, which is related to perception of sound. These information are referred to as *descriptors* or *features*. *Low-level features* (LLF) are directly extracted and computed from the audio signal and describe information related with *Spectral* or *Energy* components. They are extremely objective, but they poorly describe music to users. *High-level features* (HLF) carry a great significance for human listeners, hence they are the most feasible for wide-diffusion application. They are very subjective, since they give the higher level of abstraction from the audio signal they refer to. They can represent emotional-related descriptors of music (ED) or to non emotional-related (NED). *Mid-level features* (MLF) represent the middle layer between low- and high-level features. MLFs introduce a first level of semantics and combine LLFs with musical and musicological knowledge. In MIR literature, the gap between users description and music content is referrered to as *gap between low-level and high-level features*.

New paradigms need an accurate high-level music description. HLFs can be manually annotated by human listeners (context-based), but this approach is impractical, due to the large amount of music availability and the high subjectivity of the annotation. HLFs can also be based on a set of LLFs (content-based), but this is a hard task that involves machine learning prediction techniques. In the current situation, context-based approaches are mainly used in applications that interact with users, hence via HLFs, whereas content-based approaches have usually been limited to applications dealing with the audio content, hence via LLFs. In chapter 2 we will give an overview of these applications. An application that interacts with users and addresses the issue of wide music availability should be content-based and describe music by means of HLFs.

The representation of HLFs has two approaches: *cathegorical* and *dimensional*. The cathegorical approach tends to assign binary value descriptors to music. That is, a song can be either descripted or not descripted by a feature, such as *rock* or *not rock*. In the dimensional approach, it is possible to quantifies how much a feature describes a song, e.g., *in a 9-point scale from 1 to 9, this song has a value of roughness equal to 7.*

With particular respect to emotional-related descriptors, *Music Emotion Recognition* (MER) is a research field that aims at investigating how to conceptualize and model emotions perceived from music. Dimensional approach to MER aims at representing emotions in a continuous (dimensional) space. The most referred space in the literature is the Valence and Arousal 2-dimensional space (AV). The Valence concerns how much a mood is positive or negative, whereas the Arousal is related to the energy of the emotion (exciting or calming). Sometimes an additional dimension, called Dominance, is considered. It represents the sense of control or freedom to act of an emotion [1] (AVD space). For the purposes of Music Emotion Recognition, songs are mapped as points in the AV or AVD space. The points represent the feelings inspired by the songs.

Technology is experiencing another kind of evolution, related with user interaction. Inputs to the computer changed from command-line terminal to windows environment and are moving to touch-screen interaction and gesture commands. Nowadays an user-friendly interface is a basic design requirement for any application. New systems seem to be oriented to the comprehension of users' requests by allowing people to communicate as most intuitively as possible. *Natural Language Processing* (NLP) is a discipline that concerns making machines able to understand human natural language. Through NLP, users might communicate their requests to a machine as they were talking to a human. One of the most relevant NLP application[1] is Siri[2], the Apple's voice assistant. It is able to receive commands in natural spoken language and translate them into directives for the device operative system.

Music search is one of the challenge that MIR community is facing. A first approach for music search involves with taking meta information (such as title or artist) from the user and returning the correspondent music content. This kind of music indexing does not consider any information about actual music content, neither at high level nor at low level. In the new music scenario, users may want to retrieve music without having any information about it, but only an idea on what to retrieve. In order to invert the process, some kind of description for music must be provided. Here are some examples of type of query for MIR applications:

[1] How innovative is Apple's new voice assistant Siri, http://www.newscientist.com/article/mg21228365.300-how-innovative-is-apples-new-voice-assistant-siri.html

[2] Apple Inc., http://www.apple.com/uk/ios/siri/

- by humming: the user sings or hums the song;
- by tapping: the user taps the rhythm pattern of the song;
- by beatboxing: the user emulates the rhythm pattern of the song by beatboxing it;
- by example: songs are retrieved by similarity to one provided;
- by semantic description: the user describes the song by text.

In this thesis we address the problem of song retrieval in a content-based manner by text-based natural language query. The purpose of this work is to create a music search system using query by semantic description. We aim at processing natural language query in order to exploit the richness of language and to capture the significant concepts of the query and qualifiers that specify the intensity desired. In this thesis we will use *words* and *concepts* as synonyms. We analyze emotional and non emotional-related description by means of semantic high and mid-level features. We use dimensional approach both for EDs (by the Valence-Arousal space) and NEDs. Song similarity can also be specified in the semantic description. The similarity among songs is intended as similarity among songs' semantic descriptions. The system combine content-based approach for the annotation and high-level features for the descriptions, hence we propose a intuitive system for users and scalable for large amount of music. We named the system *Janas*, from an ancient Sardinian word that means *fairies*. With this word, we refer to the ancient era when music and magic were considered deeply bound.

We implemented a prototype of the system as a web search engine that outputs a ranked list of songs or a playlist. However, there are several possible applications. Web digital-media store may use it in order to suggest songs from free text-based queries or by content similarity with previous orders. This system may be used as an automatic playlist generator for music player softwares or in portable music devices. Users may also be interested in the possibility of searching among their personal music collection via semantic queries.

This thesis is organized as follows. Chapter 2 presents an overview of the state of the art for: main music search engines systems, music recommendation systems, high-level descriptors used in commercial application or proposed by the MIR community, latest paradigms for music browsing. In Chapter 3 we list and explain tools and theoretical background we needed to develop our project. We cover: Bayesian decision theory, natural language sentence parsing, audio features, emotional-related HLFs, machine learning regressors for generating content-based HLFs. In chapter 4 we discuss details of the implementation of system under discussion. In chapter 5 we describe experimental results and the data set we used to collect them. Chapter 6 analyzes conclusions and possible future applications for the system we present.

Chapter 2

State of the art

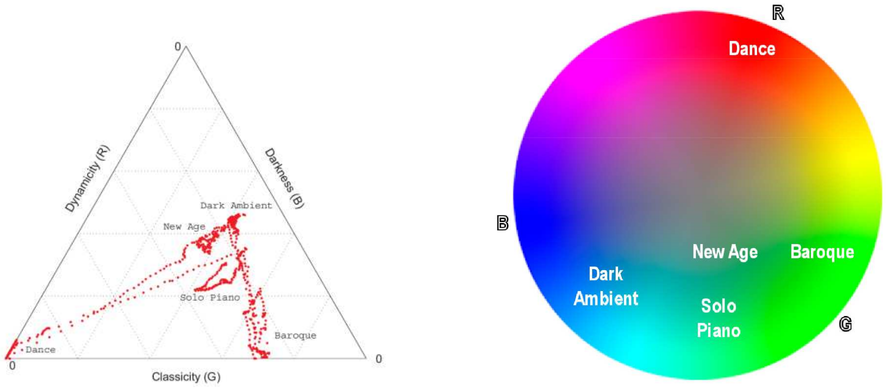
In this chapter we will provide an overview of the state of the art for Music Information Retrieval researches and applications. We will start with an overview of the high level features currently used to describe music. We will then describe the most relevant music recommendation systems currently available, from commercial and research fields. The third section concerns music representation and navigation, i.e. the solutions introduced to overcome the standard classification of songs (name, artist, genre). In the last section we will discuss the current efforts in Music Information Retrieval for the searching and retrieve of music.

2.1 High-level Features

High-level features describe music with a high level of abstraction. High-level features are usually divided in emotion-related and non emotion-related. The latter include descriptors for a wide variety of music characteristics.

In [4] the author introduces a set of bipolar-continuous NEDs for high-level perceptual qualities of textural sound modeling. The descriptors considered are: *high - low*, *ordered - chaotic*, *smooth - coarse*, *tonal - noisy* and *homogeneous - heterogeneous*. Such descriptors are suitable to describe textural sound, i.e., abstract and environmental sounds, but they are not for more complex sounds like songs.

In [5] the authors train a SVM learning machine to classify music genres. They find three high-level features able to represent and visualize genres. Such features are: *darkness*, *dinamicity* and *classicity.* They use these features to map songs' time-varying evolution among the genre space. Although these features seem to have a high descriptive potential, they do not have an intuitive definition for the generic user. A visualization of music genre is shown in figure 2.1.



(a) Resulting triangular plot for mixed genre stream test.

(b) Resulting average genre colors for mixed genre stream test.

Figure 2.1: Visualization of high-level features obtained through the analysis of the heterogeneous music stream

In [6] the authors describe a system for semantic annotation and retrieval of audio content. The annotation and retrieval is based on a vocabulary, descripted in [7] of 159 cathegorical semantic descriptors[1], divided in:

- *emotion*: concerns feelings inspired by the songs;
- *genre*: the musical genre of the songs;
- *instrument*: the instruments played during the song, included *male* and *female lead vocals*
- *song*: some general aspects such as *changing energy level* or *catchy/memorable*;
- *usage*: typical situation for listening that particular song e.g., *at a party*, *going to sleep* and so on);
- *vocals*: the style or features of the singer, such as *duet* or *breathy*.

In [8] the authors present a system that hierarchically classify recordings by genre. They extract 109 musical features divided in seven main cathegories:

- Instrumentation (e.g. whether modern instruments are present);

[1] CAL-500 semantic vocabulary for music analysis, http://cosmal.ucsd.edu/cal/projects/AnnRet/vocab.txt

- Musical Texture (e.g. standard deviation of the average melodic leap of different lines);
- Rhythm (e.g. average time between attacks);
- Dynamics (e.g. average note to note change in loudness);
- Pitch Statistics (e.g. fraction of notes in the bass register);
- Melody (e.g. fraction of melodic intervals comprising a tritone);
- Chords (e.g. prevalence of most common vertical interval).

These HLFs exhibit good performances in genre classification. Nevertheless they have been extracted from MIDI symbolic recordings, hence they have not been proofed in a real-world situation with actual audio signal.

2.2 Music recommendation systems

Music recommendation systems help users to navigate among the large amount of available music by suggesting songs that match their musical tastes. Context-based approach for music recommendation is limited to the comparison of users' music libraries for songs suggestion. Content-based approach can also focus on song actual content for the retrieval of songs similarity.

Genius is an automatic playlist generator inside the software iTunes[2]. Once it has created a playlist, it also suggests songs from the Apple Store that matches similarity with the songs in the user's library. Although no formal description of the algorithm has been provided, it is probably based on context similarity among users' libraries[3].

Last.fm[4] is a website founded in 2002, that builds a profile of user's musical tastes from Internet radio stations or computer's music player. Starting from this profiling, Last.fm provides a service of recommendations for new music, based on context-based similarity with other profiles.

In [9] the author provides a description of a music recommendation system based on context, content and user profiling. The context based information are gathered from music related RSS feeds. The content-based information is extracted from the audio. Finally, profiling information about user's listening habits and user's friends of friends' interests are considered during the music recommendation process.

[2] Apple Inc., http://www.apple.com/itunes/

[3] "How iTunes Genius Really Works", http://www.technologyreview.com/view/419198/how-itunes-genius-really-works/

[4] Last.fm Ltd., http://www.last.fm

All the content-based recommendation systems discussed above do not allow to drive the recommendation and they are mainly based on personal musical tastes. The system we propose in this work allows people to choose personal criteria for music recommendation.

2.3 Music browsing

Music has been traditionally listened to and browsed following classical taxonomy. People used to listen to music by an artist, or from an album, or matching some favorite genre. This was not sufficient and people started to create playlist of different artists, albums or genres with the intent to collect and browse music that matched other aspects of music, such as relaxing songs while studying or positive and fast songs while jogging. Music browsing differs from music recommendation because the former aims at suggesting music similar to users' tastes, whereas the latter provides ways to organize music. In this section we will present some applications that browse music and the features they use to organize it.

Pandora[5] is a website that provides a customized web radio station similar to users' tastes. It is based on the Music Genome Project[6], that aims to *capture the essence of music at the fundamental level* using a set of almost 400 attributes. Since the features are context-based and manually annotated, songs are limited to the ones just included in the Pandora database and not easily scalable.

Stereomood[7] is a website for music streaming depending on the mood or emotions felt. Once the user has chosen that particular feeling, Stereomood generates a playlist of tracks that match that mood. The database is composed by context-based annotations and it contains also annotations not directly related to objective but inspires mood, such as *sunday morning* or *it's raining.*

Mufin[8] is a service that includes music player and cloud storage functionalities. Users' songs are uploaded and analyzed by Mufin, that maps them in a sort of Valence-Arousal space extended with a synthetic-acoustic dimension. Users can have a 3D view of their music into this space and create playlists by mood neighboring. The mapping follows a content-based approach, hence no manual annotation by the user is needed. A screenshot of its 3D view of songs is shown in figure 2.3.

[5] Pandora, http://www.pandora.com

[6] The Music Genome Project, http://www.pandora.com/about/mgp

[7] Stereomood srl., http://www.stereomood.com

[8] mufin GmbH, http://www.mufin.com/us/software/mufinplayer

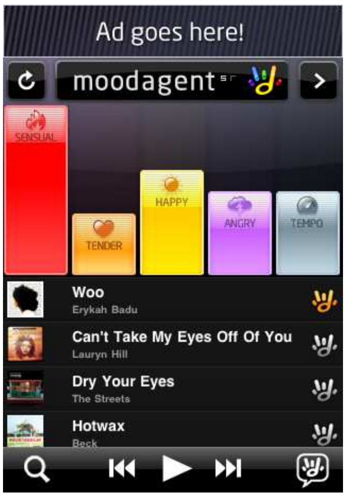


Figure 2.2: The interface of Moodagent for Android devices.

Mood Agent[9] is a music player that uses four high-level emotional-related descriptors (*sensual*, *tender*, *happy* and *angry*) and one mid-level feature (the *tempo*) to create a playlist based on music similarity. The analysis on music is content-based and the playlists are built tuning descriptors as they were sliders on an equalizer. Its interface is shown in 2.2.

Musicovery[10] is a website and mobile application that maps songs in a quantized Valence-Arousal plane. Selecting an area in the VA plane, the user can play a song according to a certain mood. It also provides a tool to play songs starting from an artist and finding similar songs. Its algorithm is based on a set of 40 acoustic features context-based (annotated by *an expert at Musicovery*[11]) that are processesed to find the mood of the song.

In [10] the authors describe an approach to multimedia playlist generator based on prior information about musical preferences of the user. The playlist generator can be driven by environmental or unintentional signals and by intentional control signals. The features selected as control signals are mood, brightness and RMS to specify loudness. It is also possible to tap the desired tempo. Features are content-based and refer to excerpts of song, in order to dynamically build the playlist (*on fly*). Since the system takes into account also the history of the system in order to capture the preferencies of the user, it can be seen as in-between music browsing and recommendation.

Most of these systems use only an emotional description to navigate among songs; those which do not, are based on just one typology of description. In this work, we combine together semantic emotional and non emotional-related description in order to provide a higher degree of freedom for user.

[9] Syntonetic, www.moodagent.com

[10] Musicovery, www.musicovery.com

[11] About us, Musicovery, http://musicovery.com/aboutus/aboutus.html

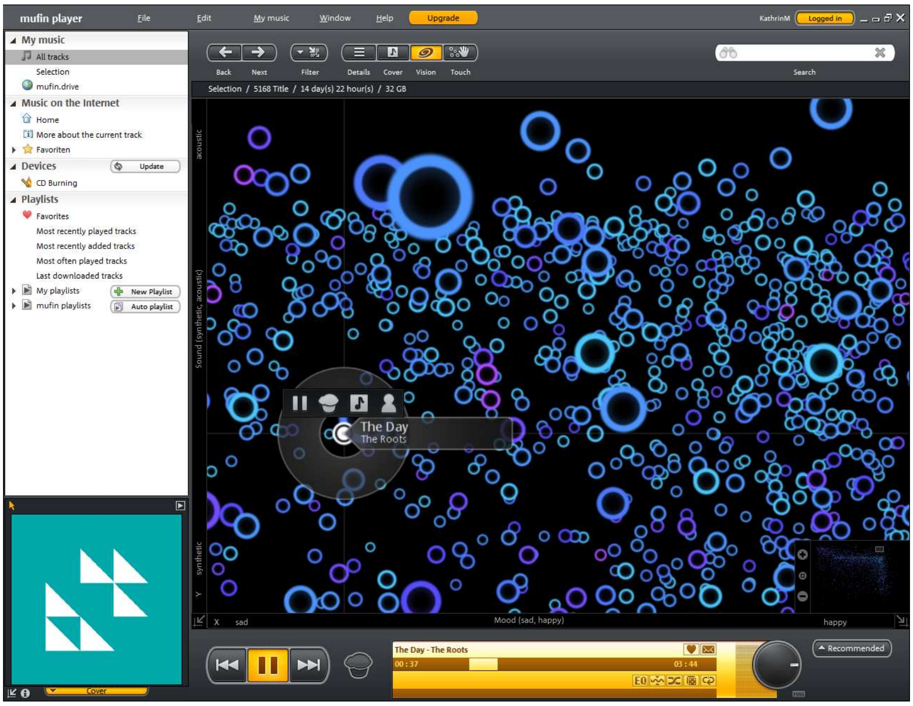


Figure 2.3: The interface of Mufin Player.

2.4 Music search and retrieve

Music search engines offer the possibility to retrieve songs by describing its content. They do not aim at organizing music or recommending it. In the latest decades some applications were created to retrieve songs by an analysis of their actual content.

Soundhound[12] is a mobile phone application that allows to search and retrieve music via query by humming. Soundhound is the rename of Midomi[13], a website for music search via query by humming. In Midomi, researches are made among both original songs and recordings sent by users, using features as pitch, tempo variation, speech content and location of pauses[14].

Shazam[15] is a popular mobile application that accepts music excerpts recorded by the microphone of the mobile device and retrieve the song recorded. The algorithm faces several problems, such as low quality record-

[12] Soundhound Inc. http://www.soundhound.com

[13] Midomi, http://www.midomi.com

[14] "This Website can name that tune", http://news.cnet.com/This-Web-site-can-name-that-tune/2100-1027_3-6153657.html

[15] Shazam Entertainment Ltd, http://www.shazam.com

ings or ambient noise. In [11] the author gives a description of the algorithm that extract a robust fingerprint by building a so-called costellation map from the spectogram of the recorded audio.

In [12] the author builds a semantic space for queries and an acoustic space for audio signals. The semantic space uses a hierarchical set of multinomial models to represent and cluster a collection of semantic documents. The acoustic space uses a signal processing chain composed by Mel-frequency Cepstral Coefficient (MFCC) extraction, stacked together through frames, analyzed by linear discriminant analysis (LDA) and finally feed to a Gaussian mixture model recognizer. The two spaces are linked together by another gaussian mixture model. This approach has good performance for the experiment proposed by the author, that relies on short audio fragments and simple semantic queries. It is tailored on analyzing the objective content of an audio signal (*what is recorded*) rather than qualities of a song (*how it sounds*).

In [6] the authors create a system of music information retrieval based on semantic description queries. To overcome the lack of data set semantically labeled, they collect a dataset of 500 songs from humans' listenings and annotations. The data set, named Computer Auditory Lab 500 (CAL500) is currently available online[16]. The songs have been modeled as GMM distributions by an Expectation-Maximization algorithm (EM). Using the models found, they also realize an automatic semantic annotator for songs.

Queries by humming or by example are useful to retrieve a certain song, but they cannot (and are not intended to) be used for music recommendation. The semantic information retrieval systems discussed are more similar to our work. In addition, we exploit the richness of language using a natural language processing module in order to accept complex queries. Moreover, our mood vocabulary considers about 2000 emotions.

[16] CAL500, http://cosmal.ucsd.edu/cal

Chapter 3

Theoretical Background

In this chapter we will present the theoretical background needed for our work. In the first section we will introduce some Music Information Retrieval methods. We will first analyze the audio features and the regressors we used in building a content-based data set. We will also give an overview the Music Emotion Recognition field. In the second section we will provide the fundamentals of Bayes decision theory we based our work on. In the last section we will present Natural Language Processing definitions and we will focus on the problem of sentence parsing and the solution we chose.

3.1 Music Information Retrieval

Music Information Retrieval is a multidisciplinary research field that deals with music information. Music information is expressed by *features* or *descriptors*. Music features are classified according to their level of abstraction: low-level features are the most objective, whereas high-level features carry the greatest semantic significance.

3.1.1 Audio Features

Low-level features, also referred to as audio features, can be classified on the acoustic cues they are capturing. LLFs can measure: the energy in the audio signal, its distribution and its related features (such as loudness and volume); the temporal aspects related with tempo and rhythm; some attributes related with the spectrum. In the following, we will illustrate the features we employed in this work, as descripted in [13]. The list of the features is shown in table 3.1.

	Low-level
Spectral	MFCC, Spectral Centroid, Spectral Flux, Spectral Rolloff, Spectral Flatness, Spectral Contrast
	Mid-level
Rhythmic	Tempo

Table 3.1: Low and mid-level features used in this work

Mel-Frequency Cepstrum Coefficients

Mel-Frequency Cepstrum Coefficients (MFCCs) are spectral LLFs that are based on Mel-Frequency scale. Mel-Frequency scale models the human auditory system's perception of frequencies. MFCCs are obtained from the coeffiecients of the discrete cosine transform (DCT) applied on a reduced Power Spectrum. The reduced Power Spectrum is computed from the log-energy of the spectrum pass-band filtered by a mel-filter bank. The mathematical formulation is:

$$c_i = \textstyle\sum_{k=1}^{K_c}\{log(E_k)cos[i(k-\frac{1}{2})\frac{\pi}{K_c}]\} \quad \text{with} \quad 1 \leq i \leq N_c, \tag{3.1}$$

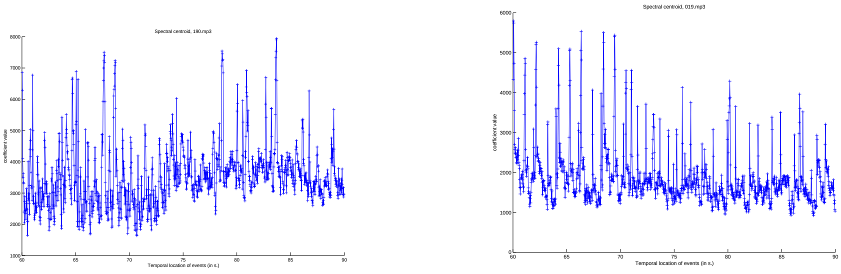
where c_i is the $i-th$ MFCC component, E_k is the spectral energy measured in the critical band of the $i-th$ mel-filter, N_c is the number of mel-filters and K_c is the amount of cepstral coefficients c_i extracted from each frame.

Spectral Centroid

Spectral Centroid is the center of gravity of the magnitude spectrum. Given a frame decomposition of the audio signal, Spectral Centroid is computed as:

$$F_{SC} = \frac{\sum_{k=1}^{K} f(k)S_l(k)}{\sum_{k=1}^{K} S_l(k)}, \tag{3.2}$$

where $S_l(k)$ is the Magnitude Spectrum at the $l-th$ frame and the $k-th$ frequency bin, $f(k)$ is the frequency corresponding to $k-th$ bin and K is the total number of frequency bins. Spectral Centroid can be used to check whether the magnitude spectrum is dominated by low or high frequency components. It is often associated with the brightness of the sound. Spectral Centroids for two songs are shown in figure 3.1.



(a) Disturbed - "Down with the Sickness"

(b) Henya - "Orinoco Flow"

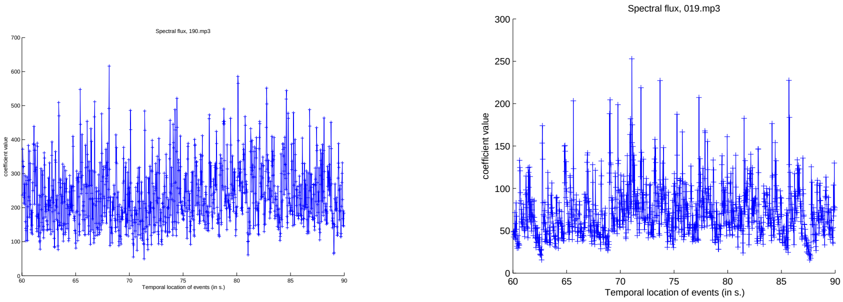
Figure 3.1: Spectral Centroid for two songs.

Spectral Flux

Spectral Flux captures the spectrum variations, computing the distance between the amplitudes of the magnitude spectrum of two successive frames. We consider the Euclidean distance:

$$F_{SF} = \frac{1}{K}\sum_{k=1}^{K}[\log(|S_l(k)| + \delta) - \log(|S_{l+1} + \delta|))]^2, \qquad (3.3)$$

where $S_l(k)$ is the Magnitude Spectrum at the $l-th$ frame and at the $k-th$ frequency bin and δ is a small parameter to avoid $\log(0)$. A representation of Spectral Flux for two songs is shown in figure 3.2.



(a) Disturbed - "Down with the Sickness"

(b) Henya - "Orinoco Flow"

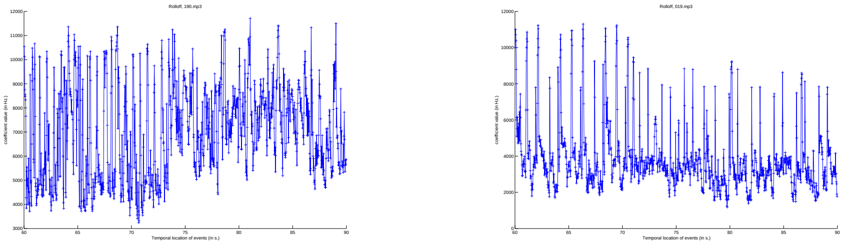
Figure 3.2: Spectral Flux for two songs

Spectral Rolloff

Spectral Rolloff represents the lowest frequency F_{SR} at which the value of the sum of the power spectrum of lower frequencies till F_{SR} reaches a certain amount of the total sum of the magnitude spectrum. Spectral Rolloff is formalized as:

$$F_{SR} = \min\{f_{K_{roll}} | \sum_{k=1}^{K_{roll}} (S_l(k)) \geq R \sum_{k=1}^{K} (S_l(k))\}, \tag{3.4}$$

where $S_l(k)$ is the Magnitude Spectrum at the $l-th$ frame and the $k-th$ frequency bin, K is the total number of frequency bins, K_{roll} is the frequency bin index corresponding to the estimated rolloff frequency $f_{K_{roll}}$ and R is the frequency ratio. In [14] authors consider R at 85% whereas in [15] R is fixed at 95%. Spectral Rolloff is shown in figure 3.3.



(a) Disturbed - "Down with the Sickness"

(b) Henya - "Orinoco Flow"

Figure 3.3: Spectral Rolloff for two songs with value $R = 85\%$

Spectral Flatness

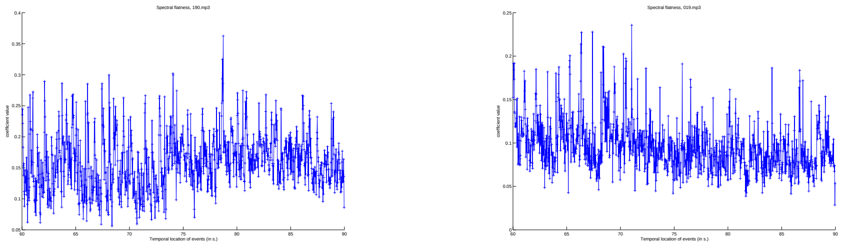
Spectral Flatness gives a measure of how much an audio signal is noisy, estimating the similarity between the magnitude spectrum of the signal frame and the flat shape inside a predefined frequency band. It is computed as:

$$F_{SF} = \frac{\sqrt[K]{\prod_{k=0}^{K-1} S_l(k)}}{\sum_{k=1}^{K} S_l(k)}, \tag{3.5}$$

where $S_l(k)$ is the Magnitude Spectrum at the $l-th$ frame and the $k-th$ frequency bin, K is the total number of frequency bins. A representation for two songs is shown in figure 3.4.

Spectral Contrast

Spectral Contrast captures the relative distributions of the harmonic and non-harmonic components in the spectrum. They have been introduced in order to compensate the disadvantage of spectral information reducing. It is defined as spectral peak, spectral valley, and their dynamics separated into different frequency sub-bands.

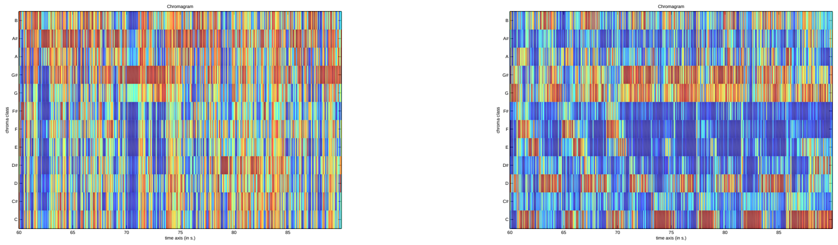


(a) Disturbed - "Down with the Sickness" *(b) Henya - "Orinoco Flow"*

Figure 3.4: Spectral Flatness for two songs

Chroma features

Chroma features attempt to capture information about the musical notes present in the audio from its spectrum. The log-magnitude spectrum is mapped into a log-frequency scale, that corresponds to a linear scale for the music temperate scale. Given the frequencies of each note in a twelve-tone scale, regardless of the original octaves, a histogram of the notes is built. The result of such processing is called *chromagram*. Each bin represents one semitone in the chroma musical octave. A representation is shown in figure 3.5.



(a) Disturbed - "Down with the Sickness" *(b) Henya - "Orinoco Flow"*

Figure 3.5: Chromagram for two songs

Tempo

The tempo is a mid-level features that represents the speed of a given piece. Tempo is specified in beats per minute (BPM), i.e., how many beats must be played in a minute. Beat is defined as *the temporal unit of a composition, as indicated by the (real or imaginary) up and down movements of a conductor's hand*[16].

3.1.2 Regressor and machine learning algorithms

A learning machine is a system that deals with learning from data and predicting new data. Given $(\mathbf{x_i}, y_i)$, $i \in \{1, ..., N\}$ a set of N pairs, where $\mathbf{x_i}$ is a $1 \times M$ feature vector and y_i is the real value to predict, a regressor $r(\cdot)$ is defined as the function that minimize the mean squared error (MSE) :

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}(r(\mathbf{x_i}) - y_i)^2. \tag{3.6}$$

Features used for learning are called *predictors*. The set of pairs referred to as *training set*[17]. Given a training set, a regressor is estimated by two steps: the training phase and the test phase. In the training phase, the training set is used to estimate a regression function. In the test phase, a set of predictors with outcome available, called *test set* is used to estimate the performances of the regressor by comparison of the correct outcome and the output of the regressor. The block diagram of training and test phases is shown in figure 3.6.

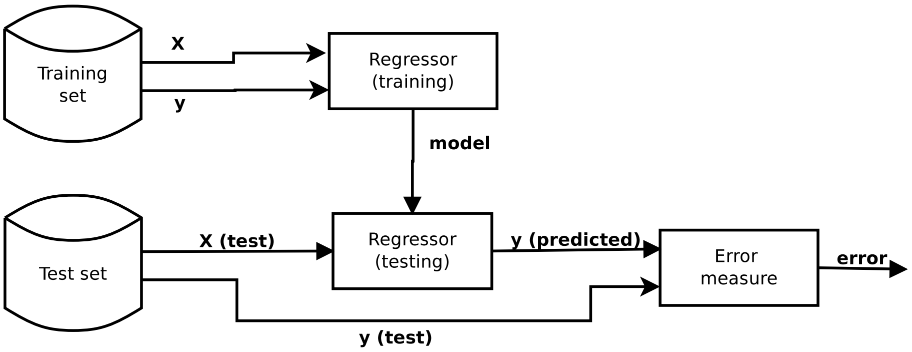


Figure 3.6: Block diagram of training and test phases for a supervised regression problem

In the following we will denote vectors with bold lowercase letters and matrices with bold uppercase letters.

Multiple Linear Regression

Linear regression starts from the assumption that there exists a linear relationship between features and variables that must be predicted. Although this is a rare assumption, this kind of regression exhibits good performances. Multiple Linear Regression (MLR) is formalized as:

$$r(\mathbf{X}) = \mathbf{X}\beta + \xi, \tag{3.7}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$ is the $N \times M$ matrix of features, β is the $M \times 1$ vector of coefficients and ξ is the $N \times 1$ vector of error terms. In order to minimize the MSE function between $r(\mathbf{X})$ and $\mathbf{y}$, where $\mathbf{y}$ is the expected output value, the least square estimator has the form:

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.8}$$

The estimate value for a new $1 \times M$ feature vector $\hat{\mathbf{x}}$ is estimated as:

$$r(\hat{\mathbf{x}}) = \hat{\mathbf{x}}\beta. \tag{3.9}$$

MLR is strictly dependent on the assumption that errors in the observed responses are normally distributed. A robust version has been developed to make MLR reliable in case errors are prone to outliers. The robust MLR method we used in our work is based on an iterative computation of weights of the regression function. Weights are assigned to each observation depending on their distance from the prediction. Assigning low weights to high distances leads to a lower regard to outliers.

3.1.3 Music Emotion Recognition

Music has always been connected to emotions. In fact, composers used to annotate mood markings on music sheets in addition to tempo indications (e.g., *in a loving manner*[1]). This is a great help to provide the music players additional information on the execution. The emotional description is one of the most intuitive for music. Indeed, as discussed in chapter 2, it is one of the most used in applications. Music Emotion Recognition (MER) is the field in MIR that studies the relationship between music and emotions. As mentioned before, two approaches are available for HLFs: the categorical and the dimensional. The former describes music with features that have a binary value, depending on whether a certain feature describes a song. The latter identifies how much a feature describes a song. In this study we focus only on dimensional approach. A dimensional approach for emotional-related descriptors involves the mapping of feelings and songs in a 2-dimensional plane, called Valence-Arousal (VA) plane. The Valence indicates how much the feeling is positive or negative, whereas the Arousal quantifies the energy of an emotion[1]. Mapping songs in the VA plane gives an immediate feedback about their emotional content. An approximated mapping of a few moods is shown in figure 3.7.

[1] *amorevole*. Music Dictionary, Virginia Tech,
http://www.music.vt.edu/musicdictionary/texta/amorevole.html

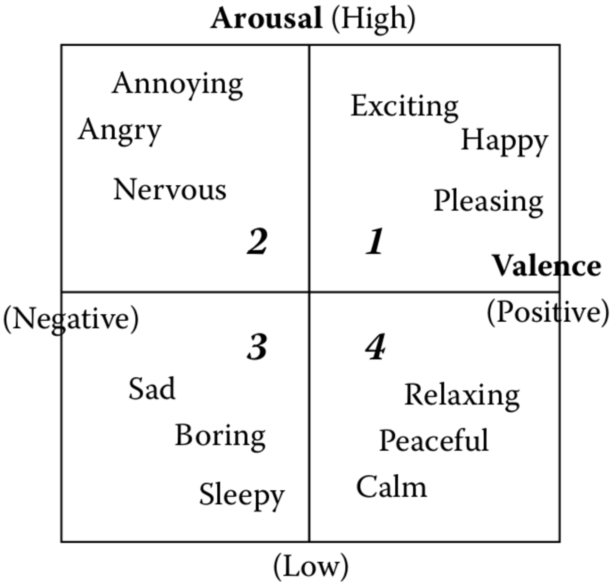


Figure 3.7: The 2D valence-arousal emotion plane, with some moods approximately mapped [1]

In [18] the authors mapped 2476 affective words in the Valence - Arousal - Dominance space. Most part of the semantic emotional-related description is based on their work.

3.2 Bayesian Decision Theory

A classifier is a learning machine that attempts to estimate from predictors a value in a discrete range of possible values [17]. Bayesian decision theory is a statistical approach for the problem of classification. It starts from the assumptions that the decision problem is posed in probabilistic terms and the probability values needed for classification are known. The Bayesian decision theory[19] explains how to use such probabilities to build a classifier.

3.2.1 Prior probability

Given an object s to be classified in one category s_i with $i \in [1, N]$, the *a priori probability* $P(s_j)$ is the probability that the object is s_j. If no further information was available, a logical decision rule for classification is:

$$\text{classify } s \text{ as } s_k \text{ if } P(s_k) = max_i(P(s_i)). \tag{3.10}$$

Given a further information q depending on the state of s, the *class conditional probability density function* $p(q|s_i)$ is the probability[2] that q has a certain value given an object known as s_i.

Given q and s, the *posterior probability* $P(s_i|q)$ is the probability that s is classified as s_i given the value q.

The *joint probability density* $p(q, s_j)$ is the probability that an object is s_j and has a certain value q, and it can be written as $p(q, s_j) = p(q|s_j)P(s_j) = p(s_j|q)P(q)$. Rearranging these leads to the *Bayes formula*:

$$P(s_j|q) = \frac{p(q|s_j)P(s_j)}{P(q)} \tag{3.11}$$

where $P(q)$ can be found as:

$$P(q) = \sum_{k=1}^{N} p(s_k)P(q|s_k) \tag{3.12}$$

Bayes formula states that posterior probability is computable as the prior probability times the class-conditional density function. The class-conditional density function $p(q|s_j)$ is the *likelihood* of s_j with respect to q. The factor in the denominator is a scale factor that ensures that all posterior probabilities sum to one. The *Bayes decision rule* for classification states:

$$\text{classify } s \text{ as} \quad s_j \quad \text{if} \quad P(s_j|q) = max(P(s_i|q)) \quad \text{with} \quad i \in [1, ..., N] \tag{3.13}$$

3.2.2 Modeling the Likelihood function

Given a set of information $\mathbf{q} = [q_1, q_2, ...q_m]^T$, we introduce a set of discriminant functions $g_i(\mathbf{q})$ with $i \in [1, N]$ such that the classification rule becomes:

$$\text{classify } s \text{ as } s_j \text{ if } g_j(\mathbf{q}) = \max(g_i(\mathbf{q})). \tag{3.14}$$

The discriminant function can be the posterior probability or some other measure dependent on the posterior probability such as:

$$g_i(\mathbf{q}) = p(\mathbf{q}|s_i)P(s_i), \tag{3.15}$$

$$g_i(\mathbf{q}) = ln(p(\mathbf{q}|s_i)) + ln(P(s_i)). \tag{3.16}$$

[2] We will use an uppercase $P(\cdot)$ to denote a probability mass function and a lowercase $p(\cdot)$ to denote a probability density function.

The conditional densities and prior probabilities are usually modeled as Gaussian densities or multivariate normal densities. A general multivariate normal density in d dimensions is written as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp[-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)], \tag{3.17}$$

where $\mathbf{x}$ is a d-component column vector, μ is the d-component *mean vector*, $\mathbf{\Sigma}$ is the d-by-d *covariance matrix* and $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ its determinant and its inverse. We can model the conditional densities and prior probabilities as multivariate normal densities:

$$p(\mathbf{q}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp[-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{q}})^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_{\mathbf{q}})], \tag{3.18}$$

$$P(\mathbf{s_i}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma_i}|^{1/2}} exp[-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \mathbf{\Sigma_i}^{-1}(\mathbf{q} - \mu)]. \tag{3.19}$$

With such modeling, the 3.16 becomes:

$$g_i(\mathbf{q}) = -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^T \mathbf{\Sigma_i}^{-1}(\mathbf{q} - \mu) - \frac{d}{2}ln(2\pi) - \frac{1}{2}ln|\mathbf{\Sigma_i}| + lnP(\mathbf{s_i}). \tag{3.20}$$

3.3 Natural Language Processing

The discipline of Natural Language Processing (NLP) deals with *the design and implementation of computational machinery that communicates with humans using natural language* [20, Preface]. NLP includes a wide variety of researched tasks, such as automatic summarization, discourse analysis, natural language generation, question answering. In this section we will focus on parsing, i.e., determining the grammar analysis of a given sentence. We will first introduce the Part-of-Speech tagging and the Context Free Grammars[21]. We will then review the Probabilistic Context-Free Grammars and probabilistic sentence parsing [22].

3.3.1 Part-of-Speech tagging

Part-of-speech (POS) is a linguistic category of words, that is generally defined by its grammar role in a sentence. POS's major categories are verbs and nouns. POS can be divided into two supercategories: *closed class* types and *open class* types. The former include those categories whose members' amount can unlikely increase, like prepositions. The latter include categories like nouns or verbs, where new words often occur. Part-of-speech tagging is the process of assigning part-of-speech categories to word in a corpus. POS tagging faces several issues, like disambiguation (is *book* a noun or a verb?) and open class terms that may be unknown by the tagger. POS tagging can employ two main approaches: *rule-based* and *stochastic*. Rule-based approach involve using disambiguation rules to infer the POS tag for a term. Stochastic approach computes probabilities:

$$P(\text{word}|\text{tag}) \times P(\text{tag}|\text{previous } n \text{ tags}) \tag{3.21}$$

to classify a term with a certain tag. POS's rules and probabilities are computed or inferred from previously annotated sentence corpus. POS tagging is useful in order to analyze the grammar of a sentence and its meaning. In order to represent a sentence, some kind of organization of POS is needed.

3.3.2 Context-Free Grammar

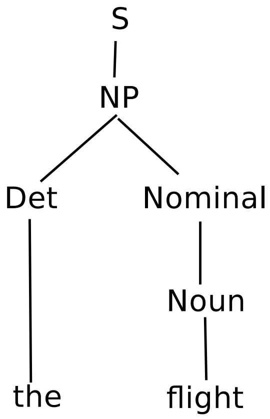


Figure 3.8: An example of parse tree representation of a Context-Free grammar derivation.

A group of words may behave as a single unit or phrase, called a *constituent*. For example, a *noun phrase* is a group of words linked to a single noun. A context-free grammar (CFG) consists of a set of rules, each of which expresses the ways that symbols of the language can be grouped and ordered together, and a lexicon of words and symbols. For example, a noun phrase (NP) can be defined as:

$$NP \rightarrow Det\ Nominal. \tag{3.22}$$

$Nominal$ can be defined as:

$$Nominal \rightarrow Noun|Noun\ Nominal, \tag{3.23}$$

i.e., a nominal can be one or more nouns. *Det* and $Noun$ can be defined as well as:

$$Det \rightarrow a; \tag{3.24}$$

$$Det \rightarrow the; \tag{3.25}$$

$$Noun \rightarrow flight. \tag{3.26}$$

The symbols that are used in a CFG are called *terminal* if they corresponds to words (like *the* or $flight$) and *non-terminal* if they express clusters. We say that a terminal or non-terminal symbol is derived by a non-terminal symbol if it belongs to its group. A set of derivation in a CFG is commonly represented by a parse tree, where the root is called *start symbol* (see figure 3.8). A CFG is defined by four parameters:

1. a set of non-terminal symbols N
2. a set of terminal symbols Σ
3. a set of productions P of the form $A \to \alpha$ where A is nonterminal and $\alpha \in N \cup \Sigma$
4. a start symbol S

The CFG is suitable to parse a sentence, i.e., to represent a sentence as a parse tree that groups the constituents and explains the underlying grammar and words' POS tags. A parse tree can be generated by means of a Probabilistic Context-Free Grammar.

3.3.3 Probabilistic Context-Free Grammar

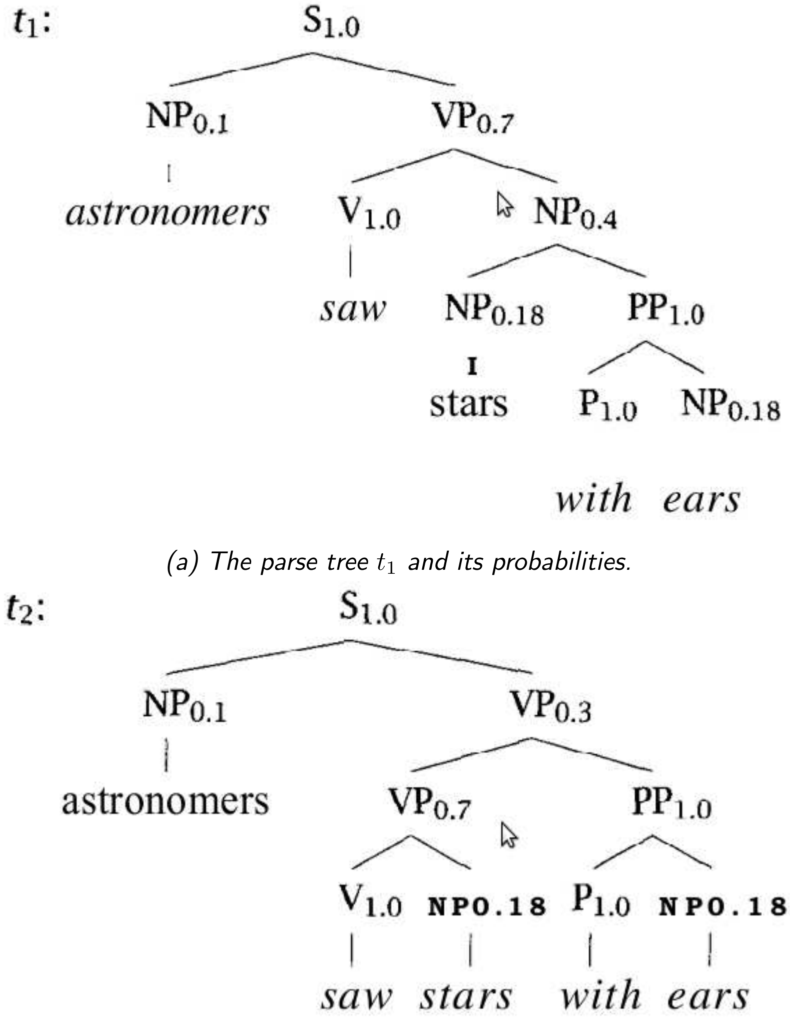
A *Probabilistic Context-Free Grammar* (PCFG) is a probabilist model that builds a tree from a sentence using probabilities to choose among possible structures. It is formalized as a CFG where probabilities are considered in productions:

1. a set of non-terminal symbols N
2. a set of terminal symbols Σ
3. a set of productions P of the form $A \to \beta[p]$ where p is the probability that A will be expanded to β
4. a start symbol S

Such grammar can be used to parse sentences of language. Probabilities of expansions are inferred by means of machine learning techniques on previously annotated sentences. PCFGs have been proofed to be a robust model, because implausible expansions have low probability. They also give a good probabilistic language model for English. In the example in figure 3.9 we show two probabilistic parse trees for the sentence *astronomers saw stars with ears*. The values on the nodes refer to the probabilities for that node to be derived from his father. We can see the starting point probability is equal to 1. We can compute the parse tree probabilities as:

$$\begin{aligned} P(t_1) &= 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\ &= 0.0009072 \\ P(t_2) &= 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\ &= 0.0006804 \end{aligned} \tag{3.27}$$

The most probable parsing, and the one correct, is t_1.



(a) The parse tree t_1 and its probabilities.

(b) The parse tree t_2 and its probabilities.

Figure 3.9: Two parse trees for the sentence astronomers saw stars with ears.

Chapter 4

Implementation of the system

In this chapter we will describe the architecture of the system. The general scheme of the system is shown in figure 4.1. The system is composed by four main elements: i) a semantic model of songs; ii) a semantic model ofs concepts; iii) the computational core; iv) the visualization module.

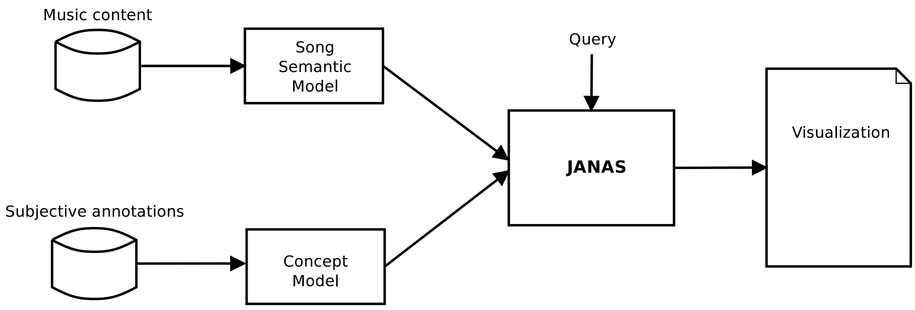


Figure 4.1: Block diagram for the general system descripted.

For each song, two semantic models are derived by its music content. The first model refers to emotional-related description of the song, whereas the second refers to non emotional-related description.

A similar formalization is also used for concepts. A word is modeled either emotionally or non-emotionally, depending on its meaning.

Given the system is based on free-text query, the computational core parses the query to capture the key-words that are relevant for the research. The query can express: semantic emotional-related description; semantic non emotional-related description or song similarity. Semantic descriptions are mapped in the emotional- and non emotional-related concepts model. The song similarity is computed as similarity among songs' semantic models. This kind of research is defined *query by semantic example* (QBSE[23]) since it considers similarity among semantic descriptions of the songs. The computational core uses the mapping to compute similarity scores that represent how much a song match the query. In this chapter we will refer to the computational core as *Janas*.

The visualization module shows the results of computational core.

4.1 Music content semantic modeling

The system we propose deals with semantic text-based query based on emotional- and non emotional-related description. We used the data set proposed in [24], called MsLite. This data set is only annotated for ED, for this reason we ran a survey to annotate it for NEDs. From the survey we collected NED annotations only for a part of the data set. Because of this, we implemented an automatic annotation system in order to annotate the non-annotated songs. The NED conceptualization modeling system is based on regression functions explained in Chapter 3.

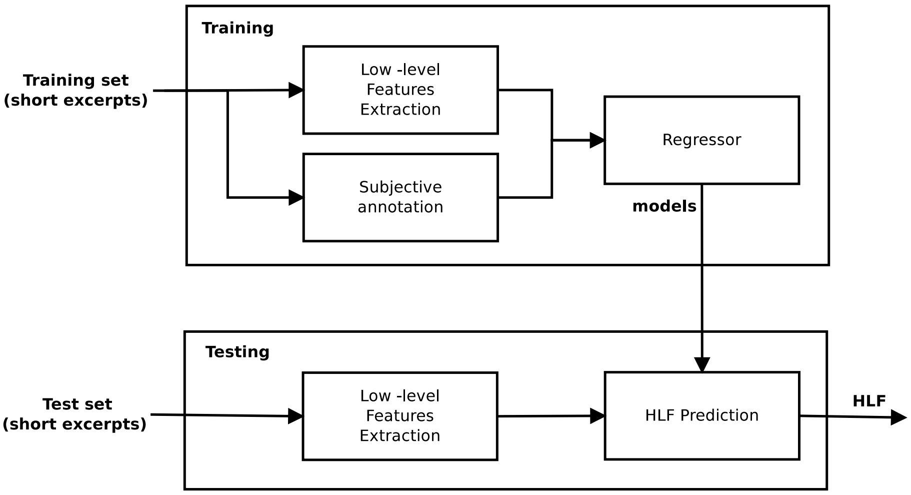


Figure 4.2: Block diagram for a regressor problem.

The general scheme for a regression procedure is shown in figure 4.2. For the purpose of regressors' training, we built a training set. The training set was composed by LLFs extracted from the audio excerpts (as mentioned in chapter 3) as predictors and subjective annotations as outcome variables. We trained two regressors, a Linear Regressor and its Robust version. Since it exhibited the best performances in the training set, we use the Linear Regressor to predict non-emotional HLFs value for those excerpts that had not been annotated. We modeled the non-emotional music content as normal distributions, that are defined by the mean and the standard deviations. For annotated excerpts we used the means and the standard deviations of annotations, whereas for non annotated excerpts we used predicted values as means and root mean square errors as standard deviations. The root mean square errors for the two regressors are listed in table 5.2. The whole procedure of annotation and machine learning prediction is presented in chapter 5. The block diagram for the music content semantic modeling is depicted in figure 4.3. Each song is modeled, from its content, by a emotional-related model, we named VA, and by a non emotional-related model, we named EQ.

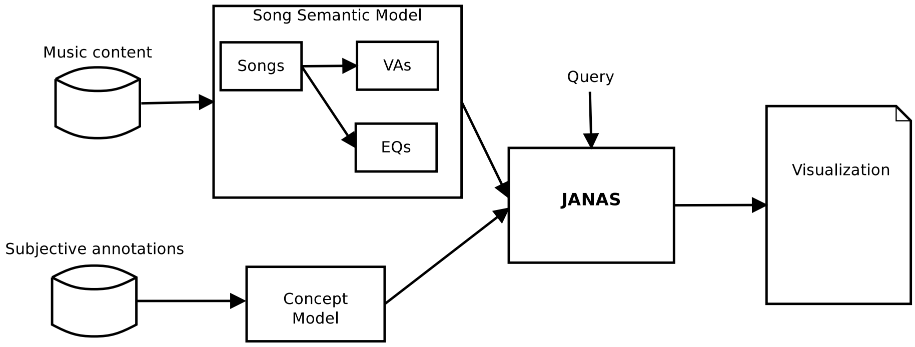


Figure 4.3: The block diagram with detail view of Song Semantic Model.

4.1.1 Emotional Descriptors

Emotions can be mapped in the 2-dimensional Valence-Arousal plane, whose axis are Valence (negative-positive) and Arousal (low-high energy). Songs are annotated in a 9-point scale from 1 to 9. For Valence, 1 is related to a very negative sensation and 9 to a very positive sensation, whereas for Arousal 1 is related to a sensation with no energy and 9 to an extremely energic sensation. The annotation provided in the MsLite depends on people musical tastes and personal perception. Given a set of K_i annotations from testers $\{(v_{1,i}, a_{1,i}), ..., (v_{K_i,i}, a_{K_i,i})\}$ for a song S^i, where $i = 1, ..., N$ is the index of the song and N is the amount of songs in the data set, we obtained the mean $\mu^{\mathbf{i}}_{\mathbf{VA}}$ and standard deviations $\sigma^{\mathbf{i}}_{\mathbf{VA}}$ of annotation as:

$$\begin{aligned}\mu^{\mathbf{i}}_{\mathbf{VA}} &= \begin{bmatrix}\mu^i_V\\ \mu^i_A\end{bmatrix} = \begin{bmatrix}\frac{1}{K_i}\sum_{k=1}^{K_i} v_{k,i}\\ \frac{1}{K_i}\sum_{k=1}^{K_i} a_{k,i}\end{bmatrix},\\ \sigma^{\mathbf{i}}_{\mathbf{VA}} &= \begin{bmatrix}\sigma^i_V\\ \sigma^i_A\end{bmatrix} = \begin{bmatrix}\sqrt[2]{\frac{1}{K_i-1}\sum_{k=1}^{K_i}(v_{k,i}-\mu^i_V)^2}\\ \sqrt[2]{\frac{1}{K_i-1}\sum_{k=1}^{K_i}(a_{k,i}-\mu^i_A)^2}\end{bmatrix}.\end{aligned} \tag{4.1}$$

We modeled the songs in the database as normal distributions in the Valence Arousal plane:

$$S^i_{VA}(\mathbf{n_{VA}}) \sim \mathcal{N}(\mu^{\mathbf{i}}_{\mathbf{VA}}, \mathbf{\Sigma^i_{VA}}), \tag{4.2}$$

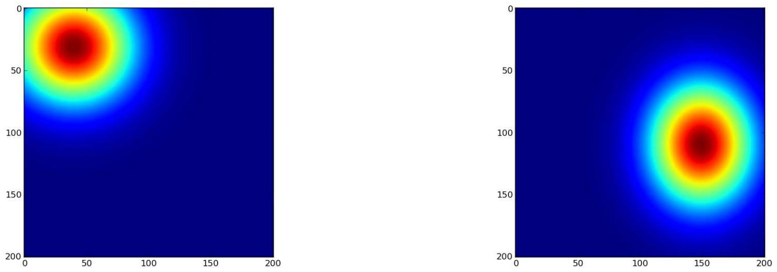
where $\mathcal{N}(\cdot)$ denotes a normal distribution,

$$\mathbf{\Sigma^i_{VA}} = diag(\sigma^i_{VA}) = \begin{bmatrix} \sigma^i_V & 0 \\ 0 & \sigma^i_A \end{bmatrix} \tag{4.3}$$

is the covariance matrix and $\mathbf{n_{VA}} = [n_V, n_A]^T$ represents a point in the Valence-Arousal plane. The emotional-related semantic distribution of the song is normalized as:

$$\int_1^9 \int_1^9 S^i_{VA}(n_V, n_A) dv\ da = 1, \tag{4.4}$$

in order the songs have the same probability. In figure 4.4 we show a representation of two songs modeled as the normal distribution. In the following we will refer to the emotional-related semantic model of a song as the song's VA.



(a) Valence-Arousal representation for "Down with the sickness" by Disturbed.

(b) Valence-Arousal representation for "Orinoco Flow" by Henya.

Figure 4.4: Valence-Arousal representation for two songs in the MsLite data set

4.1.2 Non-Emotional Descriptors

Emotional features cannot describe music exhaustively. Indeed, non emotional-related features define a wide range of music qualities and, together with ED, may provide a more complete description. In [2] the authors proposed 27 semantic descriptors divided in *affective/emotive*, *structural*, *kinaesthetic* and *judgement*. For our study we have chosen to model a subset of their entire set of concepts. Specifically, we chose to model all the structural and one judgement bipolar descriptors and one kinaesthetic descriptor, as shown in table 4.1.

In [2] the authors define the *gesture* descriptor as:

> *some aspect that makes a person start to move spontaneously.*

A similar definition is in [25], referred to the term *grooviness*:

> *a groove starts up and people stop whatever they are doing and begin to pay attention to the music; they either put their bodies in motion or adapt ongoing motion to follow the pull of the groove*

In [25], the author discusses the concept of grooviness. He attempts to find a clear definition for this descriptor, whereas in musicologist literature as among musicians no formal definition is provided[1]. The capability of making people move is an important factor while choosing music, hence we considered *grooviness* descriptor in our NEDs' set in substitution of *gesture*.

Semantic Descriptors	
Structural	Kinaesthetic
Soft/hard	Gesture
Clear/dull	
Rough/harmonious	Judgement
Void/compact	Easy/Difficult
Flowing/stuttering	
Dynamic/static	

Table 4.1: List of high-level semantic descriptors chosen from the annotation experiment in [2]

The tempo indicates the speed of a song and it can affect general definition of a music piece. We have chosen to insert the tempo descriptors in the NEDs' set. We indicated the tempo in beats-per-minute (BPM). For the tempo evaluation we used a VAMP plugin for the Sonic Annotator[2] that is based on [26]. In [26] the authors define a beat tracker using a two state model. The first state performs tempo induction and tracks tempo changes, while the second maintains contextual continuity within a single tempo hypothesis. This is similar to the human tapping process. We manually corrected wrong-estimated tempi. Since we model songs as normal distributions, we needed

[1] The usual definition seems to be *You know it when you hear it*

[2] Sonic Annotator, Queen Mary University, http://www.omras2.org/sonicannotator

standard deviations. We computed standard deviation of a song's tempo as an amount of tempo:

$$\sigma^i_{EQ_{BPM}} = 0.125\mu^i_{EQ_{BPM}}, \tag{4.5}$$

where $\mu^i_{EQ_{BPM}}$ is the computed tempo for the song S^i and $\sigma^i_{EQ_{BPM}}$ is the computed standard deviation. We have experimentally chosen 0.125 as the amount of tempo. The complete list of descriptors and their range of values can be found in table 4.2.

High-level descriptors	
soft - hard	(soft) 1 - 9 (hard)
dull - clear	(clear) 1 - 9 (dull)
harmonic - rough	(harmonic) 1 - 9 (rough)
void - compact	(void) 1 - 9 (compact)
static - dynamic	(static) 1 - 9 (dynamic)
flowing - stuttering	(flowing) 1 - 9 (stuttering)
easy - difficult	(easy) 1 - 9 (difficult)
grooviness	(not groovy) 1 - 9 (completely groovy)
Mid-level descriptors	
tempo (BPM)	30-250

Table 4.2: List of NEDs in this work

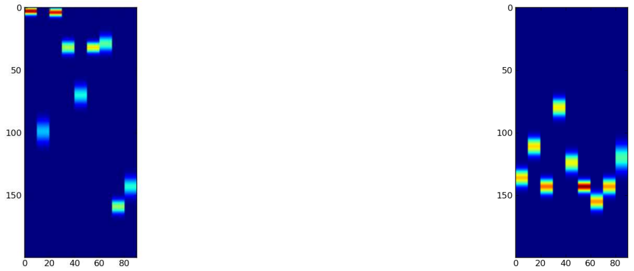
We computed means and standard deviations for the NEDs and we modeled the songs in the data set as normal distributions in 1-dimensional spaces, similarly to the adopted song model in the Valence-Arousal plane:

$$S^i_{EQ_d}(n_d) \sim \mathcal{N}(\mu^i_{EQ_d}, \sigma^i_{EQ_d}) \tag{4.6}$$

Where n_d is a point in the mono dimensional space, $d \in \mathcal{D}$ is the index of the descriptor and $\mathcal{D} = \{hard, clear, rough, comp, dyn, stutt, diff, groovy, BPM\}$ is the set of NEDs. We represented the NEDs as nine juxtaposed bars, one for each descriptor. We called this representation *semantic equalizer*. Each descriptor represents a slider that semantically equalizes the song (see figure 4.5). We define the whole non emotional-related semantic model for the song S^i as the set:

$$S^i_{EQ} = \{S^i_{EQ_d}(n_d)\} \text{ with } d \in \mathcal{D}. \tag{4.7}$$

In figure 4.5 we show a representation of two songs modeled in the normal distribution. In the following we will refer to the non emotional-related model of a song as the song's EQ.



(a) Semantic Equalizer representation for "Down with the sickness" by Disturbed.

(b) Semantic Equalizer representation for "Orinoco Flow" by Henya.

Figure 4.5: Semantic Equalizer representation for two songs in the MsLite data set

In table 4.3 we review the notation for the music content semantic models. Notice that given a song S^i, $S^i_{VA}(\mathbf{n_{VA}})$ is the model for the song's VA, whereas S^i_{EQ} is a set of models for each dimension in song's EQ.

4.2 Concept modeling

We modeled the songs in a semantic space. In order to retrieve them by means of semantic emotional- and non emotional-related descriptors, concepts need to be modeled in a similar semantic space. The block diagram for concept model is shown in figure 4.6. As we can see, the scheme for concept modeling is the dual of song modeling in 4.3.

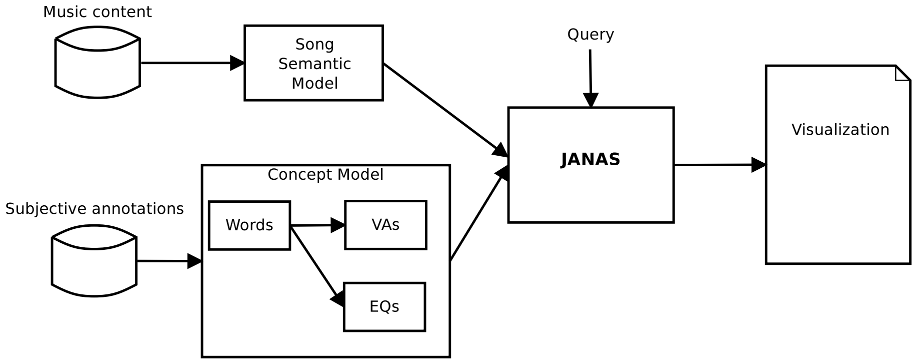


Figure 4.6: System block diagram with detail view for Concept Modeling.

Symbol		Description
$\mathbf{n_{VA}}$	$[n_V, n_A]^T$	point in the Valence-Arousal plane
$S^i_{VA}(\mathbf{n_{VA}})$	$\sim \mathcal{N}(\mu^{\mathbf{i}}_{\mathbf{VA}}, \mathbf{\Sigma^{i}_{VA}})$	Model of S^i in the VA plane
$\mu^{\mathbf{i}}_{\mathbf{VA}}$	$[\mu^i_V, \mu^i_A]^T$	Mean of the distribution of S^i in the VA plane
$\mathbf{\Sigma^{i}_{VA}}$	$diag(\sigma^{\mathbf{i}}_{\mathbf{VA}})$	Covariance matrix for distribution of S^i in the VA plane
$\sigma^{\mathbf{i}}_{\mathbf{VA}}$	$[\sigma^i_V, \sigma^i_A]^T$	Standard deviation of annotations for S^i in the VA plane
$\mathbf{n_d}$		point in the semantic equalizer along the $d-th$ space
d		index of descriptors with $d \in \mathcal{D}$
$\mathcal{D}$		EQ dimensions set. $\mathcal{D} = \{hard$, $clear$, $rough$, $comp$, dyn, $stutt$, $diff$, $groovy$, $BPM\}$
S^i_{EQ}	$\{S^i_{EQ_d}\}\forall d \in \mathcal{D}$	Set of models of S^i in EQ dimensions
$S^i_{EQ_d}(n_d)$	$\mathcal{N}(\mu^i_{EQ_d}, \sigma^i_{EQ_d})$	Model of S^i in the EQ $d{-}th$ dimension
$\mu^i_{EQ_d}$		Mean of the distribution of S^i in the EQ $d-th$ dimension
$\sigma^i_{EQ_d}$		Standard deviation of the model of S^i in the EQ $d-th$ dimension

Table 4.3: Notation used for song semantic models

4.2.1 The emotional-related semantic model

Feelings and emotions can be mapped in the VA plane as well as songs. In [18] the authors introduce a set of affective norms for English words and build a data set called ANEW. The authors collect a set of emotional-related words manually tagged by human annotators and computed means and standard deviations for Valence, Arousal and Dominance. The amount of words in ANEW is too sparse for the purposes of our work, since it includes general words such as *song*. Wordnet[3] is a lexical database of English language. English words are grouped into sets of nouns, verbs, adjectives and adverbs and interlinked by means of conceptual-semantic and lexical relations. In [27] the authors manually selected affective words from Wordnet, version 1.6, creating the Wordnet-affect database. We chose to model only the affective words that are present in both ANEW and Wordnet databases. The emotional-

[3] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu

related semantic model is the dual of the emotional-related song model in 4.1.1. For a word W^w with w index of the word $w = \{happy, sad, angry, ...\}$:

$$W_{VA}^{w}(\mathbf{n_{VA}}) \sim \mathcal{N}(\mu_{\mathbf{VA}}^{\mathbf{w}}, \mathbf{\Sigma}_{\mathbf{VA}}^{\mathbf{w}}) \tag{4.8}$$

where $\mu_{\mathbf{VA}}^{\mathbf{w}} = [\mu_V^w, \mu_A^w]^T$ are the values of Valence and Arousal from the ANEW dataset, $\mathbf{\Sigma}_{\mathbf{VA}}^{\mathbf{w}} = diag(\sigma_{\mathbf{VA}}^{\mathbf{w}}) = diag([\sigma_V^w, \sigma_A^w]^T)$ is the covariance matrix. The distribution is normalized as:

$$\int_1^9 \int_1^9 W_{VA}^{w}(n_V, n_A) dvda = 1. \tag{4.9}$$

in order the words to have the same probability.

4.2.2 The non emotional-related semantic model

For the semantic EDs' modeling, we used large existent data set Wordnet Affect and ANEW. The research studies about NEDs have not been so deeply developed and we could not rely on previous researches or data set. Filling the gap between emotional- and non emotional-related semantic descriptors is beyond the purposes of this work. We modeled only the bipolar concepts defined for each concept-dimension chosen. We modeled the bipolar non-affective words in a linear manner, assigning the maximum (1) and minimum (-1) probability values to the bounds of the dimension. For example, the words *soft* and *hard* are modeled as:

$$\begin{aligned} W_{EQ_{hard}}^{soft}(n_{hard}) &= \frac{5}{4} - \frac{2}{8} n_{hard} \\ W_{EQ_{hard}}^{hard}(n_{hard}) &= \frac{2}{8} n_{hard} - \frac{5}{4} \end{aligned} \tag{4.10}$$

where $W_{EQ_{hard}}^{soft}$ indicates the distribution of the word *soft* along the *hard*-dimension indicated by the variable n_{hard}. Values of the word *soft* range from 1 (at 1: *completely soft*) to -1 (at 9: *completely hard*) and vice versa. A representation of this modeling is shown in figure 4.7.

On the other side, tempo has been accurately descripted in the latest centuries. Composers use to indicate tempo to executors by means of many kind of markings. The tempo markings are not directly related to a certain BPM. Moreover, tempo markings vary during the centuries. In [28] the authors build a model for emotional responses to rhythm features. They introduce a table (shown in table 4.4) with Italian tempo markings and the correspondent ranges of BPM. We chose these correspondences for the tempo semantic model.

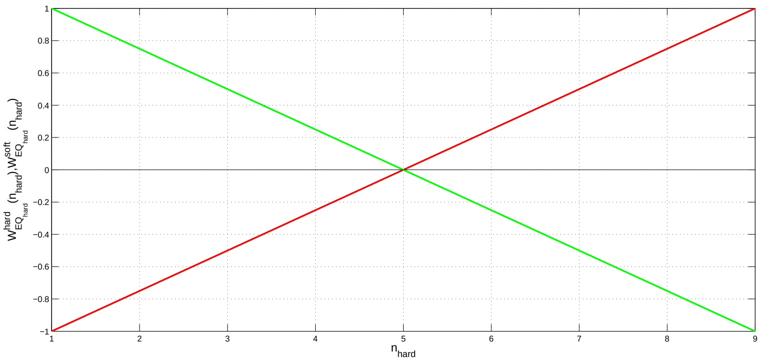


Figure 4.7: Concept modeling for words "hard" (in red) and "soft" (in green) on the hard-dimension.

Tempo Markings	BPM
Adagio	66-76
Andante	76-108
Moderato	108-120
Allegro	120-168
Presto	168-200

Table 4.4: Tempo markings and correspondent ranges of BPM

We modeled the tempo markings partially as normal distributions and partially as uniform distributions. We selected the standard deviation as:

$$\sigma^{w}_{EQ_{BPM}} = 0.25 * (BPM_{w,R} - BPM_{w,L}) \tag{4.11}$$

Where $BPM_{w,R}$ and $BPM_{w,L}$ are the right and the left bounds for the BPM range of the word W^w, with $w = \{adagio, andante, moderato, allegro, presto\}$. The modeling is formalized as:

$$W^{w}_{EQ_{BPM}}(n_{BPM}) \begin{cases} \sim \mathcal{N}(BPM_{w,L}, \sigma^{w}_{EQ_{BPM}}) & \text{if } n_{BPM} \leq BPM_{w,L} \\ = 1 & \text{if } n_{BPM} \in (BPM_{w,L}, BPM_{w,R}) \\ \sim \mathcal{N}(BPM_{w,R}, \sigma^{w}_{EQ_{BPM}}) & \text{if } n_{BPM} \geq BPM_{w,R} \end{cases} \tag{4.12}$$

The normal distributions are normalized such that

$$W^{w}_{EQ_{BPM}}(BPM_{w,L}) = W^{w}_{EQ_{BPM}}(BPM_{w,R}) = 1 \tag{4.13}$$

in order to keep the distribution continuous. A representation of this modeling is shown in figure 4.8. From this figure, we can see that the tempo marking models have a uniform distribution in the middle (between the bounds) and an exponential decrease outside the bounds, with standard deviation related with the width of tempo marking's range. We also add two bipolar adjectives, *fast* and *soft*, and we modeled them as in 4.10, appropriately rescaled for the BMP-dimension range of values.

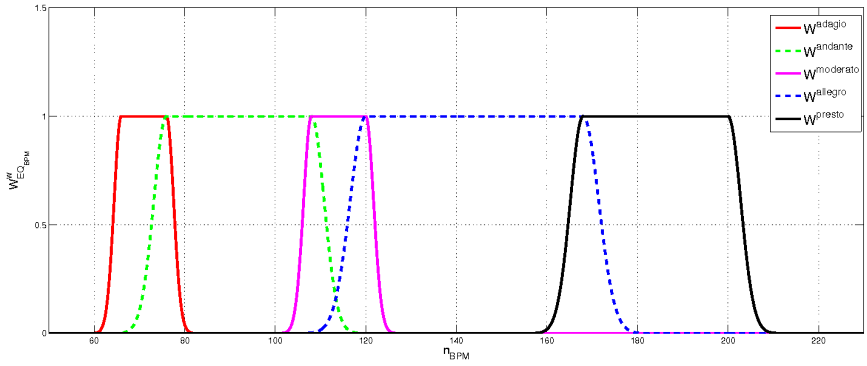


Figure 4.8: Concept modeling for tempo markings words as listed in table 4.4. In the x-axis there are BPM, in the y-axis the tempo markings modeled in the BPM mono dimensional space.

The new notation we introduced is resumed in table 4.5.

4.3 The computational core

As mentioned above, the system we propose deals with natural language semantic queries. The user can search songs by emotional-related description, non emotional-related description and by semantic example, referring to one or several songs. Once the query has been modeled, songs are ranked on its similarity to the query request content. The computational core is composed by two main modules: the first interpret the query and model it, the second compute scores from similarity between query and songs models. The block diagram of computational core is shown in figure 4.9.

4.4 The query model

At first, the query is parsed in order to retrieve titles and authors of songs, if any. If all the songs have been correctly specified or if no songs has been inserted, the query is processed by a semantic parser. If the songs are not well specified, the system asks the user to refine the research among the songs. If this is not the case, the query is parsed. The user is free to refine the research

Symbol		Description
$W^w_{VA}(\mathbf{n_{VA}})$	$\sim \mathcal{N}(\mu^{\mathbf{w}}_{\mathbf{VA}}, \mathbf{\Sigma}^{\mathbf{w}}_{\mathbf{VA}})$	Model of W^w in the VA plane
$\mu^{\mathbf{w}}_{\mathbf{VA}}$	$[\mu^w_V, \mu^w_A]^T$	Mean of the distribution of W^w in the VA plane
$\mathbf{\Sigma}^{\mathbf{w}}_{\mathbf{VA}}$	$diag(\sigma^{\mathbf{w}}_{\mathbf{VA}})$	Covariance matrix for distribution of W^w in the VA plane
$\sigma^{\mathbf{w}}_{\mathbf{VA}}$	$[\sigma^w_V, \sigma^w_A]^T$	Standard deviation of annotations for W^w in the VA plane
W^w_{EQ}	$\{W^w_{EQ_d}\}\forall d \in \mathcal{D}$	Set of models of W^w in EQ dimensions
$W^w_{EQ_d}(n_d)$	$\mathcal{N}(\mu^w_{EQ_d}, \sigma^w_{EQ_d})$	Model of W^w in the EQ $d-th$ dimension, except for $d = BPM$
$\mu^w_{EQ_d}$		Mean of the distribution of W^w in the EQ $d - th$ dimension
$\sigma^w_{EQ_d}$		Standard deviation of the model of W^w in the EQ $d - th$ dimension, except for $d = BPM$

Table 4.5: Notation used for concept models

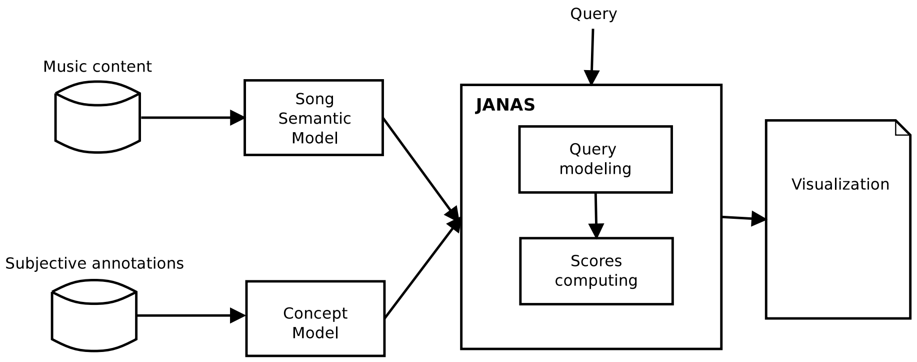


Figure 4.9: System block diagram with detail view of computational core.

specifying the songs or to start a new research. Since the query is not parsed until all the songs are specified, not necessary computational-heavy parsing is avoided. The specified songs' VAs and EQs are built. The query is then parsed through NLP, all the concepts are retrieved and words' VAs and EQs are built as well After the query modeling, three sets of concept words or semantic song models are composed, one for song similarity $\mathcal{Z}_S$ and two for emotional- and non emotional-related description: $\mathcal{Z}_{VA,D}$ and $\mathcal{Z}_{EQ,D}$, and their respective models or set of models Q_S, $Q_{VA,D}$ and $Q_{EQ,D}$. The three sets $\mathcal{Z}_S$, $\mathcal{Z}_{VA,D}$ and $\mathcal{Z}_{EQ,D}$ collect the songs, the emotional- and the non emotional-related descriptors that have been specified in the query. For example, in the query *I'd like a happy, joyful, dynamic and groovy song like "Calling you"* we will have:

$$\begin{aligned} \mathcal{Z}_S = \{S^i\} \text{ with } S^i \text{ is "Calling you ;} \\ \mathcal{Z}_{VA,D} = \{W^{happy}, W^{joyful}\}; \\ \mathcal{Z}_{EQ,D} = \{W^{groovy}, W^{dynamic}\}. \end{aligned} \tag{4.14}$$

Q_S, $Q_{VA,D}$ and $Q_{EQ,D}$ represent the results of mixing the models of the elements in $\mathcal{Z}_S$, $\mathcal{Z}_{VA,D}$ and $\mathcal{Z}_{EQ,D}$ into query's VAs and EQs. The block diagram is shown in figure 4.10.

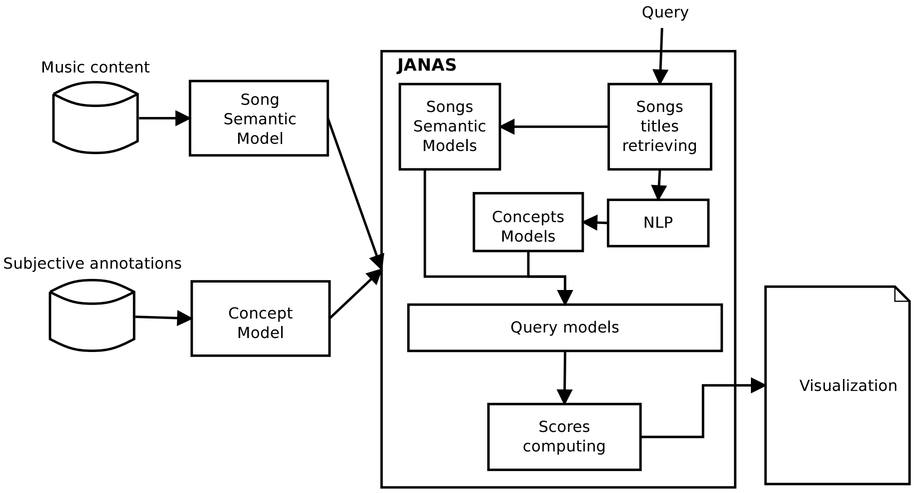


Figure 4.10: System block diagram with detail view of Query Modeling module.

4.4.1 Query by semantic example

In the current version of the system, the titles of songs and their authors must be written in quotation marks, the authors must be introduced by the key word **by** and placed after the song they refer to. Quotation marks are needed to identify the chunks of the query that may be misinterpreted by the natural language processing module. Once titles and authors have been identified, two sets of songs are composed, one with song that perfectly match a song title in the database, identified by $\mathcal{C}$ and one, referred to as $\mathcal{M}$, with titles that do not perfectly match. Songs do not match song title in database due to a spelling mistake or if that particular song not present in the database at all. If only the author is specified, songs by the author are inserted in $\mathcal{M}$. If $\mathcal{M} \neq \{0\}$, the user is asked to specify which songs the retrieved songs must be similar to. The similarity metric we used is the Jaccard similarity[29]:

$$J(t^i, q^k) = \frac{t^i \wedge q^k}{t^i \vee q^k} \tag{4.15}$$

where t^i is the title of the song S^i and q^k is the title asked by the user. Perfect matching is given by unique similarity value equal to 1, uncertainty is given by similarity value higher than the thresholds 0.2 (for songs and authors) and 0.8 (for authors only). Once all songs S^i specified in the query have been correctly defined, the are inserted in the set $\mathcal{Z}_S$. A flux diagram of the procedure is shown in figure 4.11.

4.4.2 The natural language semantic parser

We used the Stanford parser[30] to parse the query in a semantic tree. The Stanford parser is based on PCFG. The parsing has two main reasons: capture the words' grammar roles in the sentence and the dependencies among words. Through scanning the role of nodes, only adjectives, *-ing* verb and foreign words are considered and inserted in a list of word candidates. The dependencies motivation will be discussed in 4.4.5.

4.4.3 Query by semantic non emotional-related description

Word candidates are searched in the non emotional-related semantic database. Words found in non emotional-related semantic database are deleted from the word candidates and inserted in $\mathcal{Z}_{EQ,D}$ for successive EQ modeling.

4.4.4 Query by semantic emotional description

Word candidates are searched in emotional-related semantic database. If a word is found, it is inserted in $\mathcal{Z}_{VA,D}$. If the word is not in the database, the system generates a list of synonyms of the word and attempts to use them instead. If a synonym is present in the database, it is inserted in

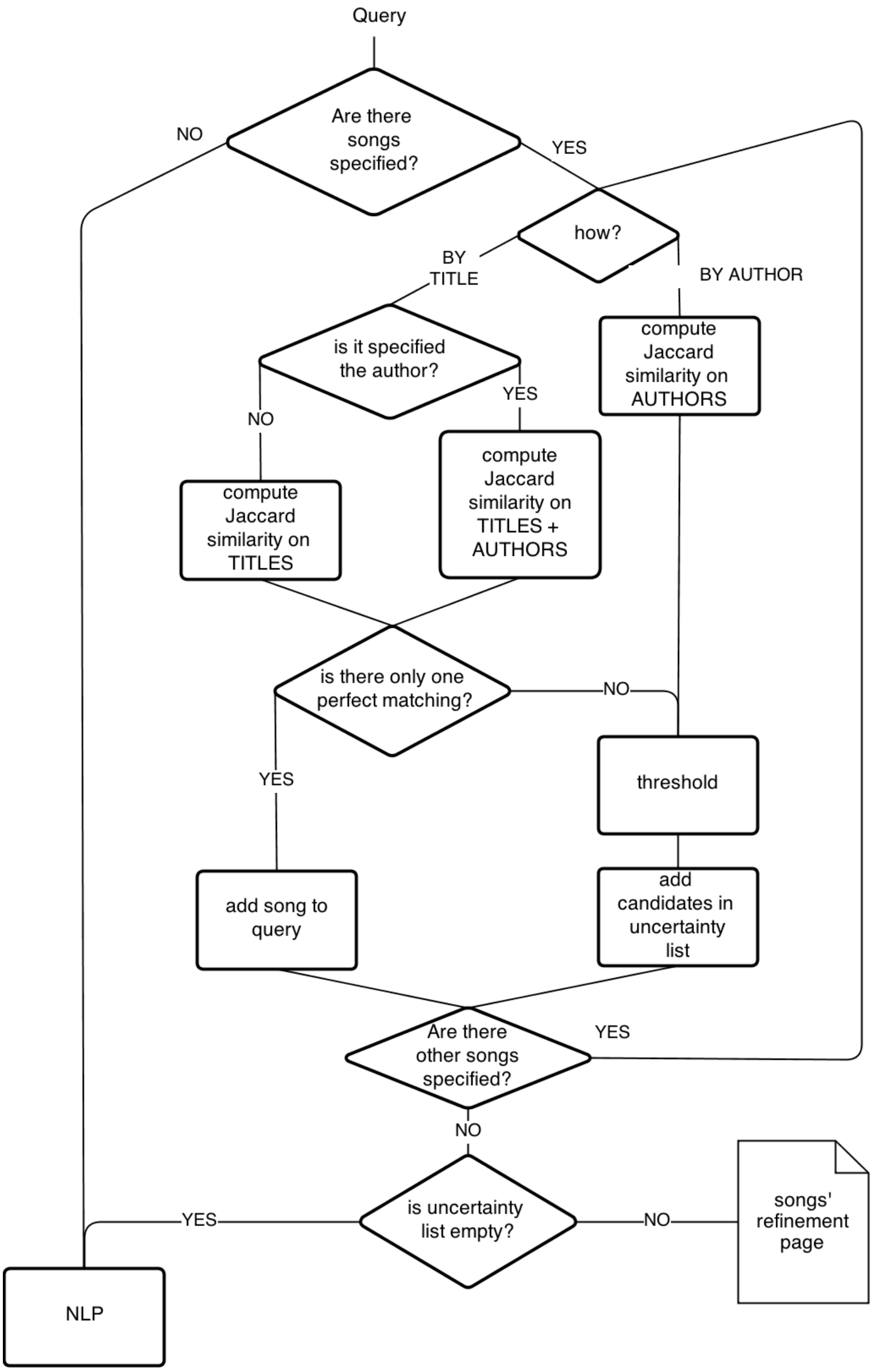

Query
Are there songs specified?
NO
YES
how?
BY TITLE
BY AUTHOR
is it specified the author?
NO
YES
compute Jaccard similarity on TITLES
compute Jaccard similarity on TITLES + AUTHORS
compute Jaccard similarity on AUTHORS
is there only one perfect matching?
NO
YES
threshold
add song to query
add candidates in uncertainty list
Are there other songs specified?
YES
NO
is uncertainty list empty?
YES
NO
NLP
songs' refinement page

Figure 4.11: Flux diagram for songs retrieval in semantic example description.

$\mathcal{Z}_{VA,D}$. The generation of synonyms is made by the Natural Language Toolkit (NLTK)[31], a set of tools for natural language processing that relies on the Wordnet dataset.

4.4.5 The role of qualifiers

In the natural language, people usually add qualifiers to the adjectives to specify the intensity they mean. A song can be defined as *completely happy*, *partly happy*, *not happy at all*, etc. Although all these definitions contain the word *happy*, they represent different concepts. The system we propose deals with qualifiers by means of the semantic natural language parsing. Once a word W^w is found to be relevant, the semantic tree is scanned to retrieve its siblings and their children, as written in the sentence. We observed that this approach is suitable to capture qualifiers, if any. The result of the semantic tree scansion for W^w is the qualifier ψ^w.

In [3] the authors discuss about verbal qualifiers for rating scales. They provide a table of values for intensity qualifiers in a 11-point scale, from 0 to 10. We considered only the mean values. The table with verbal labels and correspondent mean values is presented in table 4.6

Verbal label	Mean Value	Verbal label	Mean Value
a little	2.5	moderately	5.0
average	4.8	not	0.4
completely	9.8	not at all	0.0
considerably	7.6	partly	3.5
extremely	9.6	quite	5.9
fairly	5.3	quite a bit	6.5
fully	9.4	rather	5.8
hardly	1.5	slightly	2.5
highly	8.6	somewhat	4.5
in-between	4.8	very	7.9
mainly	6.8	very much	8.7
medium	4.9		

Table 4.6: Verbal labels and correspondent mean values from [3]

We splitted the 11-point scale in four zones, depending on the meaning of the underlying qualifiers:

0-2.5 negative : the qualifier indicates an opposite meaning with respect to the adjective's one;

2.5-5 far from positive : the qualifier indicates a meaning that is near to the adjective's one;

5-7.5 less than positive : the qualifier indicates a concept less intense than the adjective's one;

7.5.10 more than positive : the qualifier indicates a concept with more intense than the adjective's one.

Each of the four areas have a different modeling for each of the descriptions. If no qualifier is retrieved, the concept model is not modified. In the following we will use ψ^w both as the semantic qualifier for the word W^w and for its value in the likert scale. The context in which we will use will avoid ambiguities.

In addition we modeled the qualifiers *more* and *less* for the non emotional-related description. In order to recognize the *more* qualifier, a scanning of *-er* adjectives is performed.

Qualifier on non emotional-related description

EQ words are modeled as following, depending on the value of the qualifier:

0-2.5 : the word is flipped upside-down and behaves as its opposite with a qualifier value of $10 - \psi^w$ where ψ^w is the actual qualifier value. For example: *not stuttering* corresponds to *extremely flowing*;

2.5-7.5 : the word is identified by a triangle centered in ψ^w, with side 4 and maximum 1;

7.5-10 : the linear function of equation 4.10 is elevated to a power linearly dependent on ψ^w in order to assign more importance to higher value.

In case of *more* or *less* qualifiers, the system checks if any query by semantic example has been specified, i.e., if $\mathcal{Z}_S \neq \{0\}$. In this case, $Q_{EQ,D}$ is modeled as:

$$\begin{gathered}\forall d \in \mathcal{D} : \psi^w = \text{"more"} \ \wedge W^w_{EQ,d} \in \mathcal{Z}_{EQ,D}, \forall S^i \in \mathcal{Z}_S \Rightarrow \\ Q_{EQ_d,D}(n_d) = \begin{cases} Q_{EQ_d,D}(n_d) & \text{if } n_d \leq \mu^i_{EQ_d} \\ Q_{EQ_d,D}(n_d) + 1 & \text{if } n_d > \mu^i_{EQ_d} \end{cases}\end{gathered} \tag{4.16}$$

and

$$\begin{gathered}\forall d \in \mathcal{D} : \psi^w = \text{"less"} \ \wedge W^w_{EQ,d} \in \mathcal{Z}_{EQ,D}, \forall S^i \in Q_S \Rightarrow \\ Q_{EQ_d,D}(n_d) = \begin{cases} Q_{EQ_d,D}(n_d) + 1 & \text{if } n_d < \mu^i_{EQ_d} \\ Q_{EQ_d,D}(n_d) & \text{if } n_d \geq \mu^i_{EQ_d}. \end{cases}\end{gathered} \tag{4.17}$$

Qualifier on emotional-related description

The Valence-Arousal words mapping does not allow an approach like the one discussed for non emotional-related description. There is no *happy*-dimension to move along. Two approaches are possible:

1. for all the words, draw a line that starts from its opposite meaning word and ends to a superlative of the word. E.g.: for *happy*, draw a line that goes from *sad* (for *not happy at all*) to *joyful* (for *completely happy*);

2. tune the standard deviation of the word's distribution in order to assign a higher or lower importance to the points around its mean.

The first approach needs a semantic study that is beyond the purpose of this work. We use the second approach as follows:

0-2.5 : the VA map of the word is flipped upside-down and left-to-right to obtain the opposite meaning of the word. The standard deviation reduction is linearly proportional to the value of ψ^w;

2.5-5 : The word's distribution is subtracted from a normal distribution with the same mean but double standard deviation. This generates a ring around the word's original distribution, that is scaled linearly proportionally to ψ^w;

5-7.5 : the standard deviation enlargement is linearly inverse-proportional to the value of ψ^w;

7.5.10 : the standard deviation shrinking is linearly proportional to the value of ψ^w.

4.4.6 From sets to query modeling

The final distributions for emotional-related sets are multiplied together:

$$Q_{VA,D} = \prod_{W^w_{VA} \in \mathcal{Z}_{VA,D}} W^k_{VA}(\mathbf{n_{VA}}), \tag{4.18}$$

$$Q_{VA,S} = \prod_{S^i \in \mathcal{Z}_S} S^i_{VA}(\mathbf{n_{VA}}). \tag{4.19}$$

The multiplication of normal distributions results in a normal distribution as well, with mean and standard deviation dependent on means and standard deviations of starting distributions:

$$\begin{aligned} \mathcal{N}(\mu^{\mathbf{i}}, \mathbf{\Sigma^i}) \cdot \mathcal{N}(\mu^{\mathbf{j}}, \mathbf{\Sigma^j}) &\propto \mathcal{N}(\mu^{\mathbf{k}}, \mathbf{\Sigma^k}) \\ \text{with} \mathbf{\Sigma^k} &= (\mathbf{\Sigma^i}^{-1} + \mathbf{\Sigma^j}^{-1})^{-1} \\ \text{and} \mu^{\mathbf{k}} &= \mathbf{\Sigma^k} \mathbf{\Sigma^i}^{-1} \mu^{\mathbf{i}} + \mathbf{\Sigma^k} \mathbf{\Sigma^j}^{-1} \mu^{\mathbf{j}}. \end{aligned} \tag{4.20}$$

We multiply words' VAs in order to obtain a VA that is centered in-between the components' means of semantic emotional-related description and weighted by the standard deviations, i.e., taking into account the effect of rescaling by qualifiers.

The final distributions for non emotional-related sets are summed together:

$$Q_{EQ_d,D} = Q_{EQ_d,D} + \sum_{W^w_{EQ_d} \in \mathcal{Z}_{EQ_d,D}} W^k_{EQ_d}(n_d), \tag{4.21}$$

$$Q_{EQ_d,S} = \sum_{S^i \in \mathcal{Z}_{,S}} S^i_{EQ_d}(n_d), \tag{4.22}$$

$\forall d \in \mathcal{D}$. This is done because in non emotional-related description each dimension d represents an independent concept and there is no need to obtain in-between components.

The notation introduced in this chapter is reviewed in table 4.7.

4.5 The retrieval model

Once emotional, non-emotional related and song similarity queries have been modeled, they are all combined together to generate the final query model. Three similarities are computed: the songs' similarity, the EQ similarity and the VA similarity. A detailed view is shown in figure 4.12

We first introduce some notation. The EQ query models for non emotional-related semantic description and song similarity are grouped in two sets:

$$Q_{EQ,D} = \{Q_{EQ_d,D}\} \text{ and} \tag{4.23}$$

$$Q_{EQ,S} = \{Q_{EQ_d,S}\}, \tag{4.24}$$

$\forall d \in \mathcal{D}$. Queries sets and models for total song similarity and semantic description are grouped together:

$$Q_D = \{Q_{EQ,D}, Q_{VA,D}\}; \tag{4.25}$$

$$Q_S = \{Q_{EQ,S}, Q_{VA,S}\}; \tag{4.26}$$

$$Q = \{Q_S, Q_D\}. \tag{4.27}$$

$\mathcal{Z}_S$	Set of songs S^i specified in query for song similarity search
$\mathcal{Z}_{EQ,D}$	Set of non emotional-related words W^w specified in query semantic description
$\mathcal{Z}_{VA,D}$	Set of emotional-related words W^w specified in query semantic description
ψ^w	Semantic qualifier assigned to the word W^w
$Q_{EQ_d,D}(n_d)$	Query model for semantic non emotional-related description
$Q_{EQ_d,S}(n_d)$	Query model for non emotional-related song similarity
$Q_{VA,D}(\mathbf{n_{VA}})$	Query model for semantic emotional-related description
$Q_{VA,S}(\mathbf{n_{VA}})$	Query model for emotional-related song similarity
$\mathcal{M}$	List of songs candidates for query by semantic example that do not perfectly match titles in database
$\mathcal{C}$	List of songs candidates for query by semantic example that perfectly match titles in database

Table 4.7: Notation used for query models

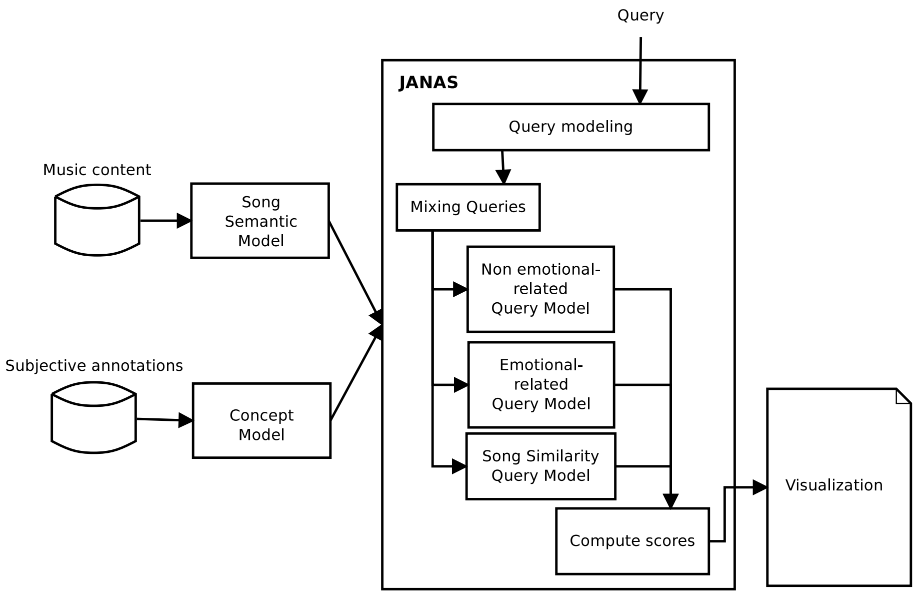


Figure 4.12: System block diagram with detail view for Scores Computing module.

The queries are combined so that each dimension appears just once in the final query:

$$\begin{aligned} Q_{VA,S} \in Q_S \iff Q_{VA,D} \notin Q_D; \\ Q_{EQ_d,S} \in Q_{EQ,S} \iff Q_{EQ_d,D} \notin Q_{EQ,D}, \forall d \in \mathcal{D}. \end{aligned} \tag{4.28}$$

The songs' similarity is computed for those factors that have not been specified in the semantic description. For example, for the query *I'd like a song like "Orinoco Flow", but happy and groovy*, the songs' similarity is not computed in the VA plane and it is computed for all the dimensions of EQ except for the grooviness; the emotional-related semantic similarity is computed for *happy* and the non- emotional-related semantic similarity for *groovy*.

All the query distributions are normalized such that their maximum value (i.e., maximum probability) is equal to one. For each song S^i and for each dimension, the system assigns a score ξ^i that is a direct computation of the posterior probability times a unanimity factor for the song.

For semantic example emotional-related description of we have:

$$\xi^i_{VA,D} = Q_{VA,D}(\mu^{\mathbf{i}}_{\mathbf{VA}})P(S^i_{VA}), \tag{4.29}$$

where $P(S^i_{VA})$ is the a-priori probability of S^i and it is directly proportional with $S^i_{VA}(\mu^{\mathbf{i}}_{\mathbf{VA}})$. This scaling factor takes into consideration the annotation unanimity around its mean. We scaled the a-priori probability so that

$P(S^i_{VA}) \in [0.8, 1]$ in order for the unanimity factor to be distinctive (to distinguish among equal posterior probabilities) but not discriminant (the a-priori probability sidely affects the score). In the same manner we obtain:

$$\xi^i_{VA,S} = Q_{VA,S}(\mu^i_{VA})P(S^i_{VA}); \tag{4.30}$$

$$\xi^i_{EQ_d,D} = Q_{EQ_d,D}(\mu^i_{EQ_d})P(S^i_{EQ_d}); \tag{4.31}$$

$$\xi^i_{EQ_d,S} = Q_{EQ_d,S}(\mu^i_{EQ_d})P(S^i_{EQ_d}); \tag{4.32}$$

$\forall d \in \mathcal{D} \setminus \{BPM\}$. For $d = BPM$ we chose $P(S^i_{EQ_{BPM}}) = 1$.

We combine these scores in order to obtain the partial scores:

$$\xi^i_{EQ,D} = (\prod_{d \in Q_{EQ,D}} \xi^i_{EQ_d,D})^{\frac{1}{|Q_{EQ,D}|}}, \tag{4.33}$$

$$\xi^i_{EQ,S} = (\prod_{d \in Q_{EQ,S}} \xi^i_{EQ_d,S})^{\frac{1}{|Q_{EQ,S}|}}, \tag{4.34}$$

where $|Q_{EQ,D}|$ is the amount of descriptors for the non emotional-related semantic description and $|Q_{EQ,S}|$ is the amount of descriptors for the non emotional-related song similarity, and

$$\xi^i_S = (\prod_{space \in Q_S} \xi^i_{space,S})^{\frac{1}{|Q_S|}}, \tag{4.35}$$

where $|Q_S|$ is the amount of types of descriptions for the song similarity. The total score is finally computed as:

$$\xi^i = (\xi^i_S \xi^i_{EQ,D} \xi^i_{VA,D})^{\frac{1}{|Q|}}, \tag{4.36}$$

where $|Q|$ is the amount of types of descriptions present in the query.

We chose to multiply scores in order to obtain the effect of a logical *AND*. This ensures higher scores for full matching songs and lower scores for partial matching. Moreover, since each subscore is less or equal to one, we use $n-th$ roots to avoid low scores for detailed queries. For example, if we have:

$$\begin{array}{llll} \xi^A_{VA} = 0.6; & \xi^A_{EQ} & = 0.7; \\ \xi^B_{VA} = 0.4; & \xi^B_{EQ} & = 0.9. \end{array} \tag{4.37}$$

the song A matches better both semantic descriptions and in fact we obtain:

$$\xi^A \quad = 0.64; \quad \xi^B \quad = 0.60. \tag{4.38}$$

The notation for scores is shown in table 4.8.

Symbol		Description
$Q_{EQ,D}$	$\{Q_{EQ_d,D}\}$	Sets of query models for non emotional-related semantic description
$Q_{EQ,S}$	$\{Q_{EQ_d,S}\}$	Sets of query models for non emotional-related song similarity
Q_D	$\{Q_{EQ,D}, Q_{VA,D}\}$	Sets of query models for semantic description
Q_S	$\{Q_{EQ,S}, Q_{VA,S}\}$	Sets of query models for song similarity
Q	$\{Q_S, Q_D\}$	Sets of query models
ξ^i		Total score for song S^i
ξ^i_S		Song similarity score for song S^i
$\xi^i_{EQ_d,D}$		Non emotional-related semantic description score for song S^i in the $d-th$ EQ dimension
$\xi^i_{EQ_d,S}$		Non emotional-related song similarity score for song S^i in the $d-th$ EQ dimension
$\xi^i_{EQ,D}$		Non emotional-related semantic description score for song S^i
$\xi^i_{EQ,S}$		Non emotional-related song similarity score for song S^i
$\xi^i_{VA,D}$		Emotional-related semantic description score for song S^i
$\xi^i_{VA,S}$		Emotional-related song similarity score for song S^i

Table 4.8: Notation used for scores

4.6 The Visualization module

Once the total scores are computed, they are sorted in a reversed-order list and presented to the user. The technology used for client-server communication in this prototype is CGI[32] based on Python code. The webpage presentation is managed via HTML and CSS. We implement two visualization: ranking list and playlist. In figure 4.13 the homepage of the system is presented. The block diagram detail is shown in figure 4.14

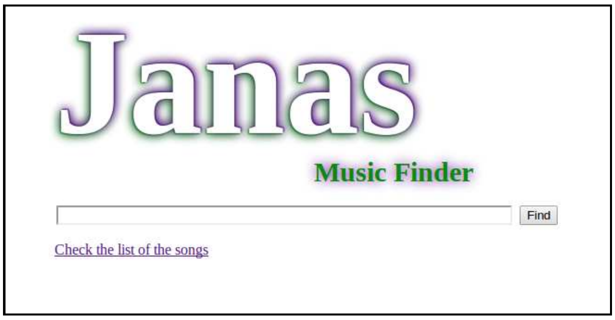


Figure 4.13: The homepage for research.

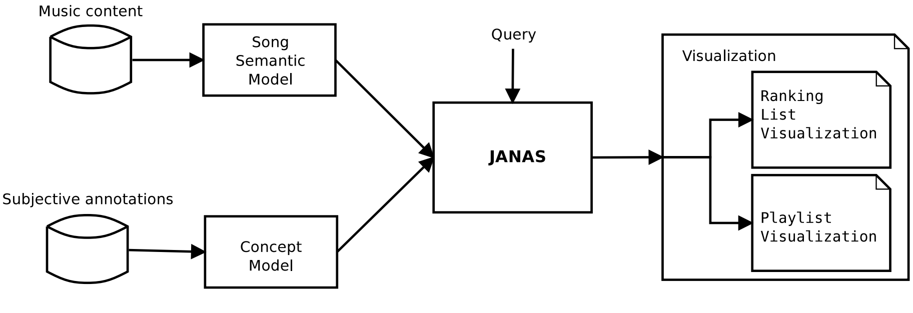
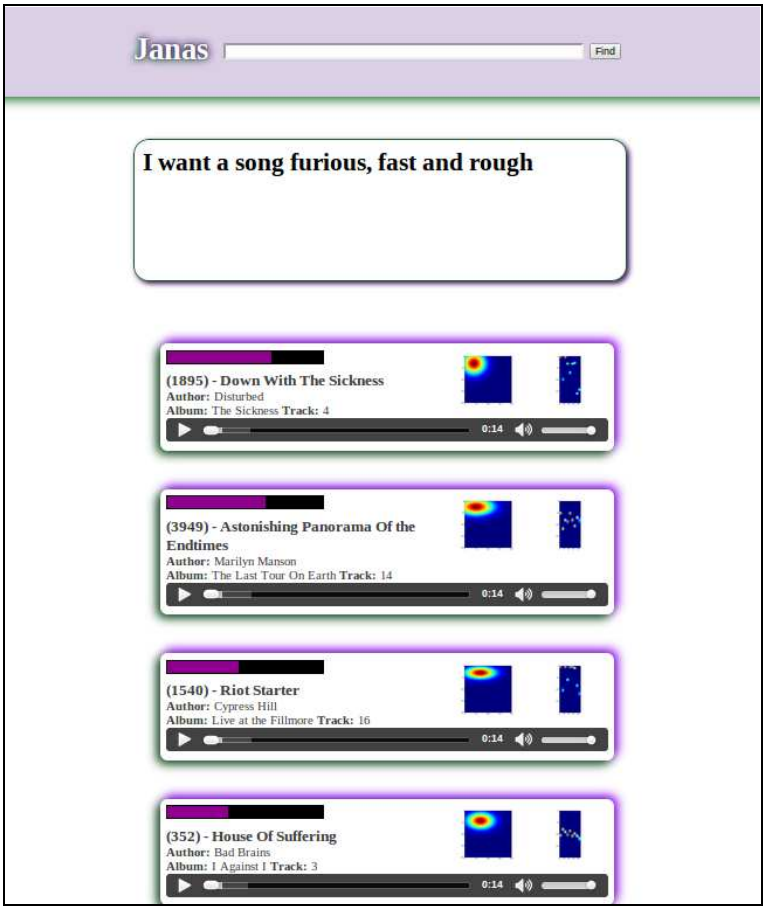


Figure 4.14: Block diagram with detailed view of the visualization module.

4.6.1 Ranking List Visualization

The songs are presented as the results of a web search. A threshold is imposed on scores to be presented: we chose 0.3 for queries without song similarity description and 0 otherwise. Each record of the ranking list displays the title, the artist, the album and the number of the track in the album. A bar is also displayed, filled proportionally to the score of the song. In addition, song's VA and EQ are shown. Ranking List Visualization is shown in figure 4.15



I want a song furious, fast and rough

(1895) - Down With The Sickness
Author: Disturbed
Album: The Sickness **Track:** 4

(3949) - Astonishing Panorama Of the Endtimes
Author: Marilyn Manson
Album: The Last Tour On Earth **Track:** 14

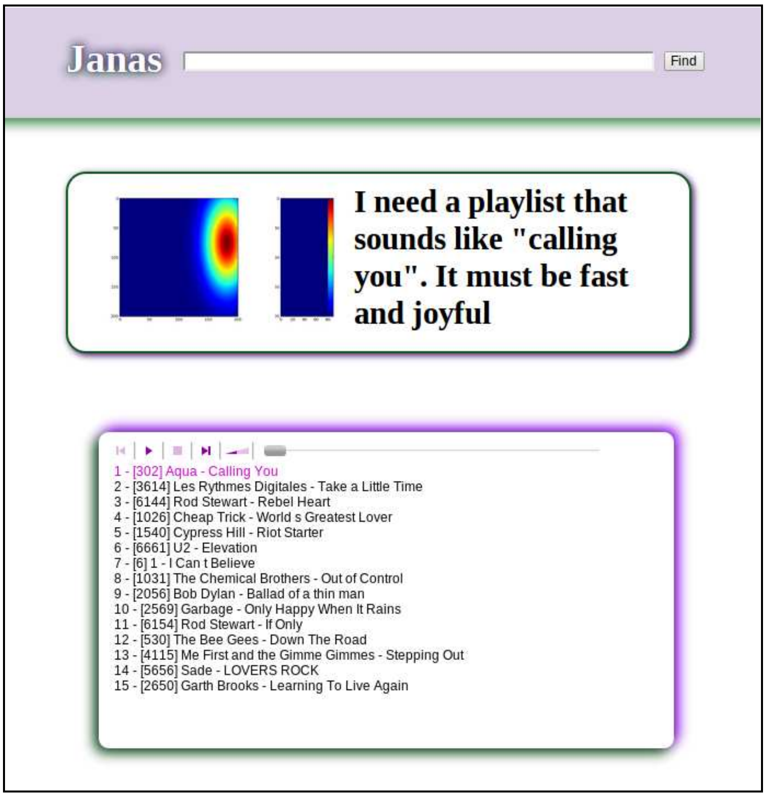
(1540) - Riot Starter
Author: Cypress Hill
Album: Live at the Fillmore **Track:** 16

(352) - House Of Suffering
Author: Bad Brains
Album: I Against I **Track:** 3

Figure 4.15: An example of ranking list visualization.

4.6.2 Playlist Visualization

If the text-based query contains the word *playlist*, the system presents the playlist visualization. Only information about title and artist are displayed. The fifteen songs with the highest scores are loaded in the playlist, regardless of the thresholds. The songs are played sequentially; the user can skip among tracks. Playlist Visualization is shown in figure 4.16



I need a playlist that sounds like "calling you". It must be fast and joyful

1 - [302] Aqua - Calling You
2 - [3614] Les Rythmes Digitales - Take a Little Time
3 - [6144] Rod Stewart - Rebel Heart
4 - [1026] Cheap Trick - World s Greatest Lover
5 - [1540] Cypress Hill - Riot Starter
6 - [6661] U2 - Elevation
7 - [6] 1 - I Can t Believe
8 - [1031] The Chemical Brothers - Out of Control
9 - [2056] Bob Dylan - Ballad of a thin man
10 - [2569] Garbage - Only Happy When It Rains
11 - [6154] Rod Stewart - If Only
12 - [530] The Bee Gees - Down The Road
13 - [4115] Me First and the Gimme Gimmes - Stepping Out
14 - [5656] Sade - LOVERS ROCK
15 - [2650] Garth Brooks - Learning To Live Again

Figure 4.16: An example of playlist visualization.

Chapter 5

Experimental results

In this chapter we will analyze the performance results of the system. We collected results through a questionnaire on paper. The questionnaire aimed at collecting the subjective opinion of testers about a prototype of the system. We will discuss the procedure we follow for collecting and processing evaluation rates for building the data set. We will shortly describe our tester sample. We will then analyze the evaluation results of the different parts of the test. We will finally review some impressions left by the subjects.

5.1 The data set

In the first studies on MIR, entire songs were annotated by testers or given as input to learning machines. In [1, chapter 2.1.3] the authors discussed the importance of taking into consideration the time-varying relationships between music and emotion. In [33] the authors focused on relationship between music and time-varying non emotional-related macro-descriptors. Nowadays the use of excerpts, either selected automatically [34] or manually [2], is widely diffused. There are several data set available for MIR purposes, like [35][36]. We selected the data set proposed in [24]. In [24] the authors created an online game called Mood Swing. In this game, people challenge friends or random users to tag songs in the Valence-Arousal plane. The tagging is made during the execution of the song and second-by-second, to annotate the the mood evolution variations. With this game, the authors obtained 50,000 Valence-Arousal dynamic annotation for songs. The music was obtained from a database of 8000 popular music tracks. A subset of this data set composed by Valence-Arousal annotations for each second of 15-seconds excerpts for 240 songs has been shared online. This subset is called MsLite. We expanded it by adding annotations for eight high-level non-emotional descriptors and one mid-level descriptor. The complete list of descriptors and their range of values can be found in table 4.2.

We designed and implemented an online survey called *Music Features Ranking Survey* in order to collect annotations for the semantic NEDs. The survey was both in Italian and English. We used the most faithful translation of the descriptors for the Italian version. We also added a short explanation of descriptors to fill the translation gap. The technologies used were HTML, CSS and PHP. Five excerpts were randomly chosen among the 240 songs in MsLite data set. For each song, the users were asked to rate each of the eight descriptor in a 9-point likert scale. The middle value 5 outlined no prevalence in the bipolar descriptors. We also collected personal data such as age, geographic area (among Africa, Asia, Europe, Middle East, North America, Oceania, South America), mother tongue (English, Italian or Other), the frequency of listening to music and some skills related to music, such as the capability of playing an instrument. It was possible not to insert personal data.

We made the survey available online from July 23rd to August 8th 2012. In this period of time, 166 people completed the test. Among the 166 people who finished the test, the younger tester was 14 years old, the eldest 64 years old. Almost one fifth of the annotator was 23 years old; 100 people, the 60% of the population, were between 21 and 25 years old. 154 people were in Europe; only the 0.6% of our population was not european. The 89% of the users chose Italian as their mother tongue, 3% chose English and the remaining 8% declared another mother tongue. About the frequency and attention of listening to music, the 63% declared to listen to music very often, paying attention to it, the 30% listens to music quite often and only the 6% of people declared they do not listen to music very often. The statistics about skills are shown in chart on figure 5.1.

We collected all the rates for each couple excerpt-descriptor. We defined $\mathbf{r}_{i,k}$ the vector containing the rating given for the excerpt i and for the descriptor k and $|\mathbf{r}_i|$ the number of annotation received for each excerpt[1]. Since we used a nine-point scale, we had $\mathbf{r}_{i,k} \in [1, 9] \subset \mathbb{N}$. We were interested in extracting the means and the stardard deviations of our sample data $\mathbf{r}_{i,k}$. We first tried to identify outliers. The outliers detention was a substantial problem. Most of the outlier detention algorithms are based on the evaluation of the deviation of samples from a certain value, that is usually the mean, the median or the mode [37]. The evaluation of the deviation is thresholded to classify the sample as outliers or not. We chose the Modified Z-score [38]

[1] The number of annotation does not depend on the descriptor.

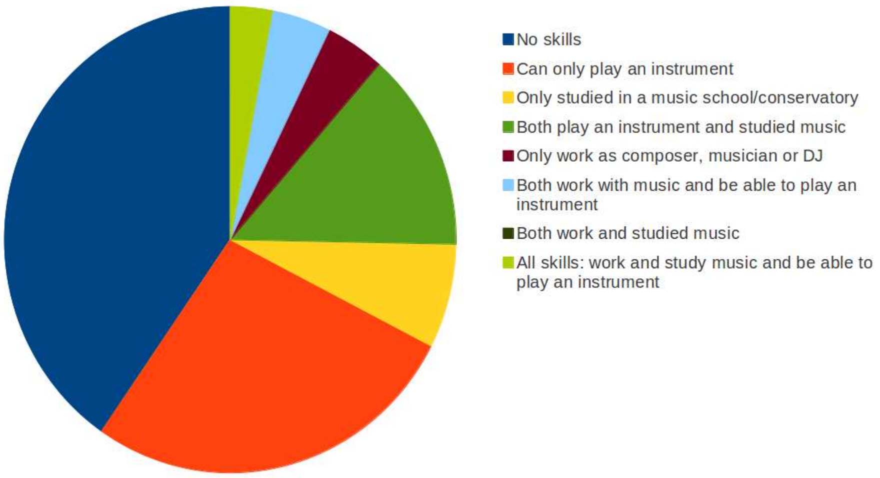


Figure 5.1: Skills owned by our survey population

algorithm (MZscore) for outliers detention, adding another modification in order for to make it more tailored to the problem. The Modified Z-score algorithm detects the outliers computing the evaluation of deviation as:

$$M_i = \frac{0.6475\dot{(}x_i - \tilde{x})}{MAD} \tag{5.1}$$

where $\tilde{x}$ is the sample median and $MAD = median\{|x_i - \tilde{x}|\}$ is the Median of Absolute Deviations. In [38] the authors suggested a threshold of 3.5. We observed that the performance of this algorithm was good, except for some specific cases. In the presence of a strong sample mode, any other value was classified as an outlier, independently on its value. If we had, for example, the rates $[4, 4, 3, 4, 4, 7]$ both the rate 3 and 7 were classified as outliers, whereas the former is a good rate and a natural consequence of human variance in tastes and judgment and the latter is completely different from the mode. We defined a range value $R_{i,k} = \max(\mathbf{r}_{i,k}) - \min(\mathbf{r}_{i,k})$ and we decided not to apply the MZscore algorithm for sample data such that $R_{i,k} \leq 2$. After we applied the MZ score algorithm, we recovered false positive samples with the following procedure. We defined $\tilde{\mathbf{r}}_{i,k}$ the sample data $\mathbf{r}_{i,k}$ after the outliers detention and removal. We focused on:

$$i, k : |\tilde{\mathbf{r}}_{i,k}| < |\mathbf{r}_i| \quad \wedge \quad \sigma^2_{\tilde{\mathbf{r}}_{i,k}} = 0 \tag{5.2}$$

i.e., on the annotations where MZscore algorithm detected and removes outliers and after removal all the sample data were equal. We defined a neighborhood for each value in the nine point scale and we chose to re-insert in the sample data those outliers that were inside the neighborhood. Neighborhoods have been arbitrary chosen. The neighborhood are referred in table 5.1.

$1 : [1, 3]$	$2 : [1, 4]$	$3 : [1, 5]$
$4 : [2, 5]$	$5 : [4, 6]$	$6 : [5, 8]$
$7 : [5, 9]$	$8 : [6, 9]$	$9 : [7, 9]$

Table 5.1: Rate neighborhood for false-positive outliers recover

Since the value 5 is the middle value that is able to split the bipolar descriptors from one meaning to the other one, no neighborhood crosses the 5 except for the neighborhood of 5 itself.

To eliminate poorly annotated data, we discarded all the excerpts

$$s_i : \exists k \in [1, ..., K] \text{ such that } |\tilde{\mathbf{r}}_{i,k}| < 3, \qquad (5.3)$$

i.e., that had not been rated at least three times for each descriptor. We discarded annotations for 110 excerpts. We finally computed the means $\mu^i_{EQ_k}$ and the standard deviations $\sigma^i_{EQ_k}$ of $\tilde{\mathbf{r}}_{i,k}$ in a similar manner than 4.1.

We used the 130 annotated excerpts to train a linear regressor and a robust linear regressor and annotate the discarded 110 excerpts. The linear regressor exhibited the best performance, hence we used it for the annotation. We consider the root mean-square errors as the standard deviation of the annotation. Performances are shown in table 5.2. The list of the songs in MsLite data set is indicated in appendix A.

High-level descriptors	LR RMSE	ROB RMSE
soft - hard	1.2	1.24
dull - clear	1.34	1.45
harmonic - rough	1.35	1.43
void - compact	0.883	0.93
static - dynamic	1.1	1.17
flowing - stuttering	1.1	1.16
easy - difficult	1.28	1.35
grooviness	1.54	1.54

Table 5.2: Linear Regressor and Robust Linear Regressor Root Mean-Square Errors for each non-emotional high-level descriptor.

Songs in MsLite data set are listed in appendix A.

5.2 Test procedure

The questionnaire introduced the instructions about how to build a semantic query and how to specify songs for query by sample example. We collected information about how frequently the testers listen to music. Five predefined queries were proposed to the subject. The subject was asked to rate the obtained results in a 9-point likert scale from 1 (very bad) to 9 (very good). Hence, the testers were asked to use the system and evaluate the general performances of the system in a similar 9-point scale. The complete text of the questionnaire can be read in appendix B.

During the test, subjects were left alone and no further explanation about test procedures and queries typologies was provided by us until they ended the evaluation. Each subject made one only test.

We collected 30 questionnaires. 54% of the subjects declared to listen to music less than three hours a day, 36% of subjects have been classified as expert since they listen to music more and 10% indicated they listen to music for reasons related to their job. The lates will be referred to as *professionist*. A pie chart of music listening habits distribution is shown in figure 5.2.

Beginner
Expert
Professionist

Figure 5.2: Music listening profiles in test population.

5.3 Predefined queries evaluation

Five predefined queries have been proposed to subjects. Subjects have been asked to evaluate each of them with a rate between 1 and 9. A summary of evaluations is listed in table 5.3.

Query	Mode	Mean	Std
I want a song very groovy and happy	7	7.1	1.2415
I want a song not happy at all, dull and flowing	7	6.9667	1.3257
I want a playlist that sounds angry, fast and rough	8	7.7667	1.04
I would like to listen to calm songs, like "Orinoco Flow", flowing and slow	8	7.7333	1.0148
I want a playlist not angry, and not stuttering and with a slow tempo	8	7.4333	1.3817

Table 5.3: Evaluation for the predefined queries

I want a song very groovy and happy

The histogram of evaluations for the query *I want a song very groovy and happy* is shown in figure 5.3. The mode of the rates is 7 and it corresponds to 50% of the total rates. 10% of the subjects rated the results of this query less than 5 and the 86.67% rated it with a grade higher or equal than 6. The mean of evaluations is 7.1, with a standard deviation of evaluation equal to 1.24. The mean of the rates assigned by professionists is 7.33, that is slightly better than the general mean, due to the first good impression on the system.

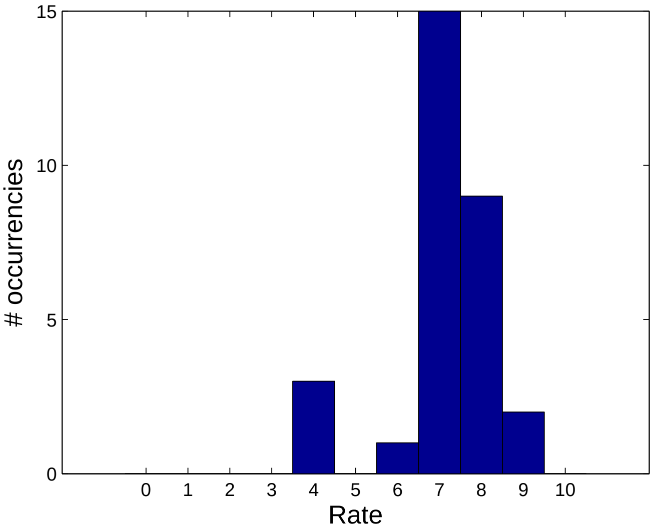


Figure 5.3: Histogram of evaluation rates for the query "I want a song very groovy and happy".

I want a song not happy at all, dull and flowing

The mean of the evaluations for the query *I want a song not happy at all, dull and flowing* is 6.9667, with a standard deviation of evaluation equal to 1.32. 33% of the testers agree with the mode of 7, the 13.3% considered query results to be poor (below or equal 5) whereas the 70% considered them good (not lower than 7). The mean of rates assigned by professionist is 7.33, that again is slightly better than the general mean. The histogram of evaluations is shown in figure 5.4.

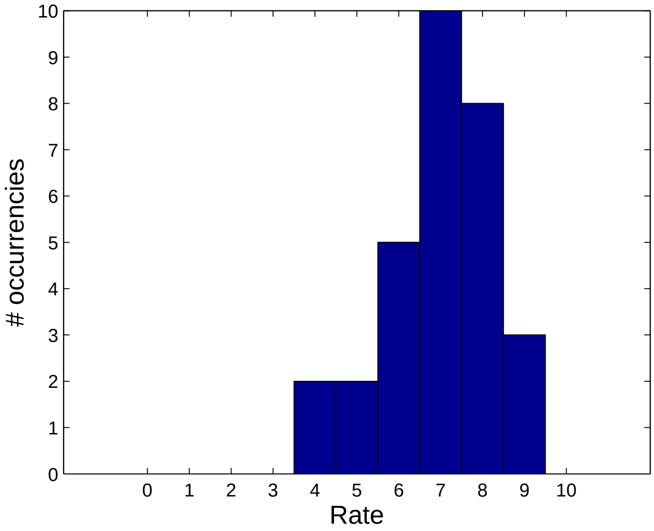


Figure 5.4: Histogram of evaluation rates for the query "I want a song not happy at all, dull and flowing".

I want a playlist that sounds angry, fast and rough

The mode of the rates for the query *I want a playlist that sounds angry, fast and rough* is 8, according to 46% of the testers. Results for this query did not receive any rate below the middle rate 5 and the 90% of subjects assigned a rate higher or equal to 7. The mean of the evaluations is 7.7667, with a standard deviation of evaluation equal to 1.04. The mean of rates assigned by professionist is 7, that is worser than the general mean. They probably expected something more characteristic. The histogram of evaluations is shown in figure 5.5.

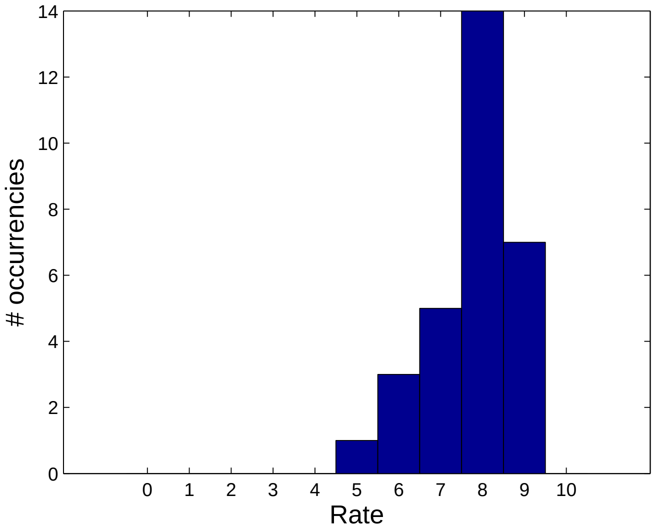


Figure 5.5: Histogram of evaluation rates for the query "I want a playlist that sounds angry, fast and rough".

I would like to listen to calm songs, like "Orinoco Flow", flowing and slow

The histogram of evaluations for the query *I would like to listen to calm songs, like "Orinoco Flow", flowing and slow* is shown in figure 5.6. The mean of the evaluations is 7.73, with a standard deviation of evaluation equal to 1.0148. 40% of testers created the mode 8. The mean of rates assigned by professionist is 7.67, that is slightly worser than the general mean, but their mode is still 8. Also this query did not receive any negative rate (lower than 5).

I want a playlist not angry, and not stuttering and with a slow tempo

The mode of the rates of results for the query *I want a playlist not angry, and not stuttering and with a slow tempo* is again 8 (43%), the mean of the evaluations is 7.43, with a standard deviation of evaluation equal to 1.3817. This query exhibits 6.67% of negative rates, also pretty bad, as 3. The mean of rates assigned by professionist is 6.33, that is considerably worser than the general mean. We consider that the negative qualifiers must to be better modeled. The histogram of evaluations is shown in figure 5.7.

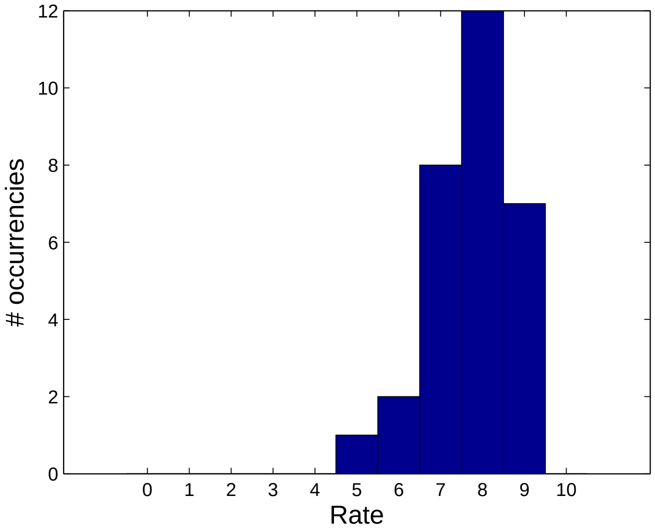


Figure 5.6: Histogram of evaluation rates for the query "I would like to listen to calm songs, like "Orinoco Flow", flowing and slow".

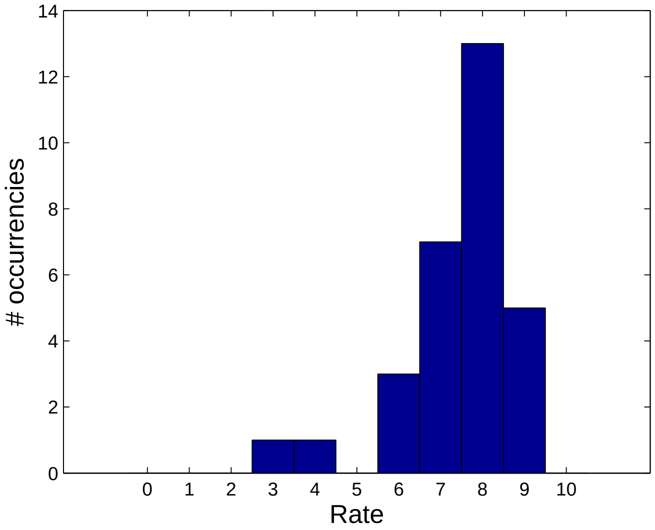


Figure 5.7: Histogram of evaluation rates for the query "I want a playlist not angry, and not stuttering and with a slow tempo".

5.4 General evaluation

The general evaluation has been rated after a free use of the system. The queries tried by testers had not been recorded. An overview of rates is shown in table 5.4.

Question	Mode	Mean	Std
Please indicate the general evaluation on the results obtained when using free queries	7	6.2667	1.5742
Do you think this system is useful?	9	7.4667	1.4794
Would you ever use this kind of system?	9	7.1000	2.0401
How do you evaluate the system in general?	7	7.2667	1.0807

Table 5.4: Evaluation for the system's general aspects.

Please indicate the general evaluation on the results obtained when using free queries

The first question was related to the free-text search. The subject were asked to evaluate the general quality of obtained results intended as the correspondence between queries and results. The mode of the rating is 7, agreed by 46.67% of testers. The means of rates is 6.2667 with a standard deviation of 1.5742. The professional listeners, who better know the applications currently available, assign an average rate slightly higher: 6.33. The 20 of testers evaluate the free-text experiment as insufficient, both for their high expectations and for the limitations of this prototype. Histogram of the rates are shown in figure 5.8.

Do you think this system is useful?

The idea of a music search engine based on semantic text-based queries has been widely appreciated. 33% of testers consider the system as completely useful, rating it 9 in the 9-point scale. The average general rate is 7.4667 with standard deviation of 1.4794. Professional listeners considered the system on average less useful, rating it 7, but the single rates 5, 7 and 9 are too different to suppose a general reason. A histogram of the evaluations is shown in figure 5.9

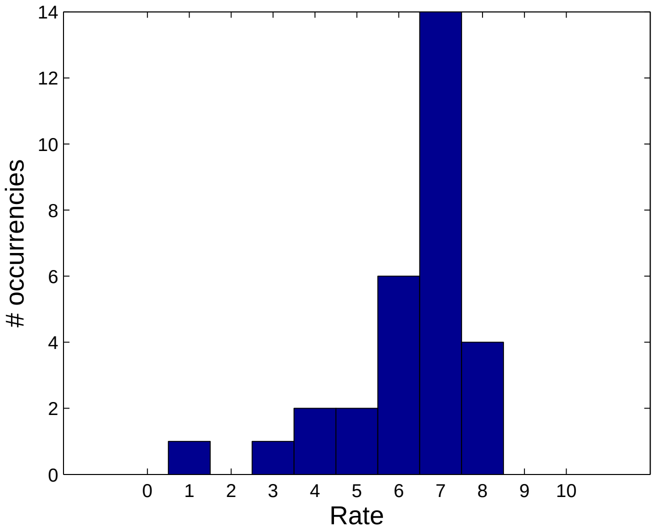


Figure 5.8: Histogram of evaluation rates for results of free-text queries.

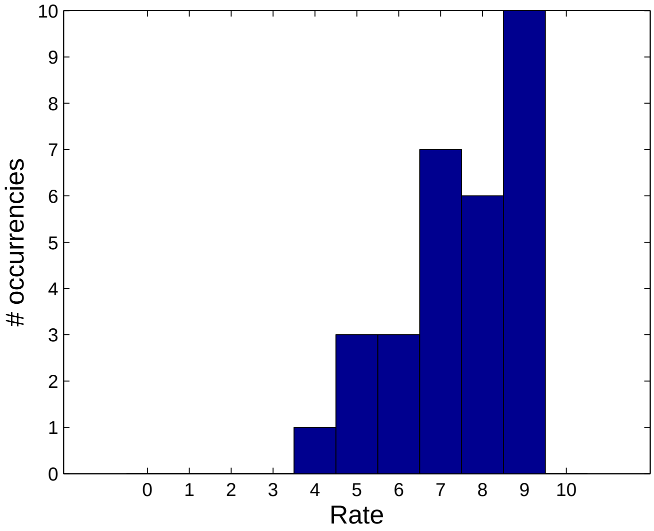


Figure 5.9: Histogram of evaluation rates for usefulness of the system.

Would you ever use this kind of system?

We asked subjects if they would ever use this kind of system. Most of the answers are positive: 76% of testers assigns a rate over 7 and in particular the 30% of them gives the maximum score. On the other side, 10% claim they likely would not use it. The means of rate is 7.1 with standard deviation 2.04, the highest of the questionnaire, because many people usually prefer to use applications they know rather than learn new tools. Even if all professional listeners would use this system, their average rate 7 is lower than the total mean. The histogram is referred to in figure 5.10

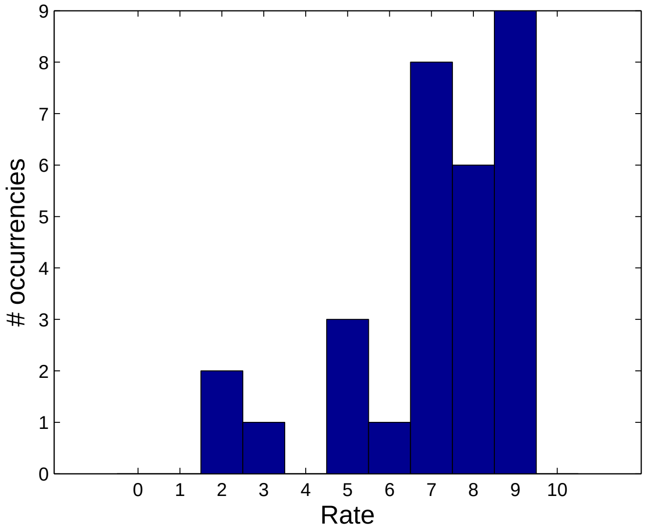


Figure 5.10: Histogram of evaluation rates for the question about personal potential use of the system.

How do you evaluate the system in general?

Finally we asked to rate the general concept of the system, taking into account all the elements: the results, the idea, the implementation and functionalities we propose, the usefulness and potentials. 10% of subjects seems doubtful, and assigns a rate between 4 and 5; hence 90% evaluate positively this work and its potentials. 7 is the mode for 46.67% of testers, the mean is 7.2667 and the standard deviation is 1.0807. Testers classified as professionist give a even better evaluation of 8. Since this was the first system of this kind that testers have ever tried, they probably appreciate this attempt. The histogram is shown in figure 5.11

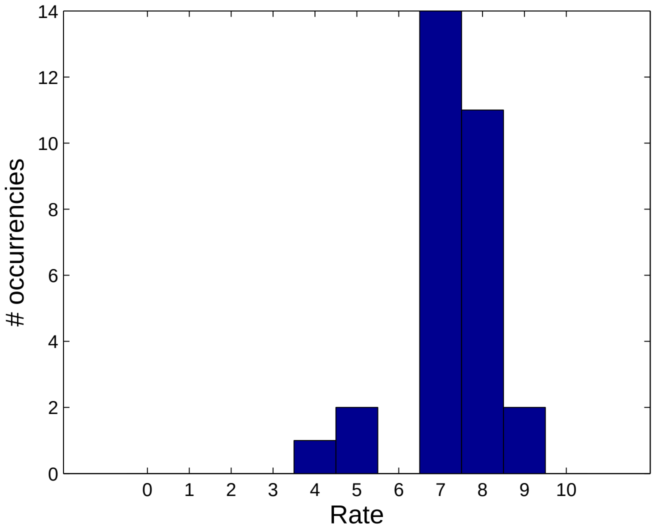


Figure 5.11: Histogram of evaluation rates about the general concept of the syste.

5.5 Notes and discussion on results

The questionnaire allowed testers to add optional notes. Some testers referred problems with the emotional-related description. We suppose that this is due to the gap between semantic mapping (from ANEW[18] data set) and song mapping (from MsLite) in the Valence-Arousal plane. We also received comments on issues on the role of qualifiers, both about their working principle and their actual effect. We assume that a set of perceptive test in order to better tune the model would increase system's performances. Moreover, the goodness of specification by qualifiers is strictly dependent on the sentence parsing precision. If the parser do not recognize qualifiers as belonging to a certain adjective, the qualifier is not recognized by the system and is not modeled. This is an important issue when using negative qualifiers such as *not* or *not at all*. Finally, the data set is composed by 240 songs. Some subjects found this amount too small.

Nevertheless, the questionnaire exhibits satisfying performances for our systems. All the sections received mainly positives rates and the modes of rate are everywhere higher or at least equal to 7. In particular, testers appreciate the playlist visualization, that makes the system suitable for music browsing. The current prototype have several limitations, as discussed above, but it represents a promising starting point for future developments.

Chapter 6

Conclusions and future developments

In this chapter we will review the work presented in this thesis and introduce future perspectives and applications for this study.

6.1 Conclusions

In this thesis we proposed a music search engine based on semantic text-based queries. The semantic text-based queries deal with: emotional, non-emotional and sample example description. The emotional-related description is based on affective words, the non emotional-related relies on a set of high-level NEDs. A set of semantic descriptors for the mid-level feature tempo is also included. The sample example allows to specify song similarity by naming the title or the author of the piece.

Our work is defined in the Music Information Retrieval research field. It is meant to address the problem of music search, retrieving the songs that match a semantic description. In order to create a relationship between semantic description and music results, we mapped music and semantic words on a common representation. We chose the Valence-Arousal plane to map the emotional description and a set of 1-dimensional spaces for non-emotional description, that we defined *semantic equalizer*. We defined the similarity among song as similarity in these two spaces, hence we built a model to uniform the results. As for our knowledge in the MIR literature, this is the first work in which natural language processing is applied to queries in order to exploit the significance nuances of the grammar. We used the natural language processing to tune the music search on the user's request. Through NLP, the system is able to accept complex queries that specify the desired intensity of features. We relied on probabilistic Bayesian decision theory in order to generate a ranking list of songs that match the query's request. We offered two kinds of visualization of the results. The first is based on a list of the retrieved songs, like a traditional search engine. The second presents a playlist in order to sequentially listen to pieces.

We collected evaluations about this prototype and the general concept of a semantic music search engine through a questionnaire. Subjects appeared pleased by the results and attracted by potential usefulness of a system of this kind. We consider this as a promising starting point for further developments.

6.2 Perspectives and future developments

We will present some future applications that can derive from this work.

6.2.1 Refining the semantic equalizer

In this work we chose 17 semantic non-emotional descriptors, belonging to 9 semantic dimensions. We called this space *semantic equalizer*, shortened to EQ. We considered EQ as a set of 1-D lines and we mapped only the semantic bounds of these dimensions. We suppose that potentials of EQ description may be further exploited in several manners.

As we stated before, we only mapped bipolar adjectives for each dimension in EQ (except for the tempo dimension). Nevertheless, several NEDs may be mapped in each EQ dimension. For example, the following semantic descriptors may be mapped in the *harmonious/rough*-dimension assigning them a rate between 1 and 9 : *smoothed*, *rocky*, *knobby*, *scraggy*, *crude*, *melodious*, *tuneful*, *musical*, *sweet-sounding.* This approach will increase the amount of possible non-emotional descriptors and hence accept queries such as *I'd like to listen to snappy songs that sound melodious and scattering*, with *snappy* mapped in the *groovy*-dimension, *melodious* mapped in the *harmonious/rough*-dimension and *scattering* in the *flowing/stuttering*-dimension.

In this work we define EQ with nine dimensions. This amount is not sufficient to map all the possible words for non-emotional description. Nevertheless, a short number of dimensions allows low computational burden and more intuitive representation. A possible approach for future developments may be to map words in a multidimensional fashion. The music for *party*, for example, can be defined as music with high arousal and positive valence, positively groovy, easy to be listened and dynamic. Modeling the word *party* may allow to search music or a party with no need to add other dimensions.

6.2.2 User profiling

High-level features carry a great semantic significance, but they have the side-effect to be highly subjective. In this work we took into account the subjectiveness of the semantic and song description by modeling words and songs with the standard deviation of annotations together with their mean. Possible future developments include building semantic and songs' models tailored to users, in order to tune the results of a query to *what he meant* rather than *what he said.*

6.2.3 Expansion of data set

The current data set include 240 excerpts of 15-seconds each. An expansion of the amount of pieces may lead to better performances. Moreover, the algorithms for dealing with whole songs instead of excerpts are yet to be implemented. This aspect has several issues, like the correct segmentation to obtain excerpts to be annotated (content-based) and the availability of time-segment query, such as *I'd like a song that contatins at least 30 seconds of anger* or *Give me a playlist composed by songs that sounds harmonious only half of time and rough elsewhere.*

6.2.4 Query by speech

Future technologies are based on interaction between machine and human as more natural and intuitive as possible. Applications like Siri by Apple or the Google Glass project[1] are two popular examples for this new paradigm. Google has included text-by-speech functionalities in its mobile operative system Android[2]. Future developments include easier interaction with the system like via query by speech.

6.2.5 Music browsing and thumbnailing

The Valence-Arousal plane has just been used in several application for music browsing. Other kind of descriptions has marginally been employed. We consider that Valence-Arousal plane and semantic equalizer may be used in the future to navigate into music libraries, like is currently made for other descriptors such as genre or decades (e.g. *80s rock*).

[1] Google Glass, http://www.google.com/glass/start/

[2] http://www.google.com/mobile/voice-search/

Moreover, the VA plane gives an immediate mood representation of a song, whereas typical music players only provide meta-information (such as title, artist and genre). We expect new applications to show some content-based representation for music in order to make users intuitively and immediately understand how a song sounds.

Appendix A

List of songs

We list here the songs in MsLite data set. The songs that have been annotated via content-based machine learning are indicated in bold.

ID	Artist	Album	Title
1086	Chicago	Chicago X	Gently I ll Wake You
6	1	1	I Can t Believe
51	**1**	**1**	**Sweet**
57	**3 Doors Down**	**The Better Life**	**Be Like That**
55	**3 Doors Down**	**The Better Life**	**Duck And Run**
109	Abba	Arrival	Tiger
123	**Abba**	**Voulez-Vous**	**The King Has Lost His Crown**
124	**Abba**	**Voulez-Vous**	**Does Your Mother Know**
140	AC/DC	Back In Black	What Do You Do For Money Honey
138	AC/DC	Back In Black	Hells Bells
181	Ace of Base	The Sign	Happy Nation
207	Aerosmith	Nine Lives	Taste Of India
197	Aerosmith	Live Bootleg	Back In The Saddle
194	**Aerosmith**	**A Little South of Sanity - Disk 1**	**Same Old Song And Dance**
214	**Aerosmith**	**Nine Lives**	**Pink**
3151	Alan Jackson	Who I Am	Let s Get Back To Me And You
3148	**Alan Jackson**	**Who I Am**	**All American Country Boy**
4323	Alanis Morissette	MTV Unplugged	Ironic
1650	Alice DeeJay	Who Needs Guitars Anyway	Celebrate Our Love
257	All Saints	All Saints	Never Ever
264	All Saints	All Saints	Lady Marmalade
302	Aqua	Aquarium	Calling You
312	Aqua	Aquarius	Cuba Libre
293	**Aqua**	**Aquarium**	**My Oh My**
331	Backstreet Boys	Black Blue	Everyone
352	**Bad Brains**	**I Against I**	**House Of Suffering**
398	BBMak	Sooner Or Later	Love On The Outside
455	Beatles	A Hard Day s Night	If I Fell
502	Beatles	Magical Mystery Tour	All You Need Is Love
491	Beatles	Beatles For Sale	Everybody s Trying To Be My Baby
457	Beatles	A Hard Day s Night	And I Love Her
481	**Beatles**	**Beatles For Sale**	**Rock And Roll Music**
452	**Beatles**		**The Long And Winding Road**
558	**Ben Folds Five**	**Whatever And Ever Amen**	**Brick**
3304	**Billy Joel**	**Piano Man**	**Captain Jack**
3308	**Billy Joel**	**The Stranger**	**Scenes From an Italian Restaurant**

604	**Blind Melon**	**Blind Melon**	**Holyman**
639	**Blink 182**	**Enema Of The State**	**Dysentery Gary**
678	**Blood Sweat Tears**	**Blood Sweat Tears**	**Spinning Wheel**
689	**Bloodhound Gang**	**One Fierce Beer Coaster**	**Shut Up**
2056	Bob Dylan	Live at Budokan Disc 1	Ballad of a thin man
743	Bon Jovi	New Jersey	Living In Sin
754	**Bon Jovi**	**Slippery When Wet**	**Livin On a Prayer**
764	Boston	Boston	Foreplay Long Time
6090	**Bruce Springsteen**	**Live 1975-1985 disc 3**	**The Promised Land**
815	Bryan Adams	On A Day Like Today	Inside Out
819	Bryan Adams	On A Day Like Today	Where Angels Fear To Tread
838	**Bryan Adams**	**So Far So Good**	**Cuts Like A Knife**
878	Busta Rhymes	Extinction Level Event - The Final World Front	Just Give It To Me Raw
868	**Busta Rhymes**	**Anarchy**	**Here We Go Again**
900	Cake	Fashion Nugget	She ll Come Back To Me
1011	Cheap Trick	Silver	Day Tripper
1026	Cheap Trick	Silver - Disc 1	World s Greatest Lover
229	Christina Aguilera	Christina Aguilera	So Emotional
228	**Christina Aguilera**	**Christina Aguilera**	**I Turn To You**
1090	**Chumbawamba**	**Tubthumper**	**Amnesia**
1209	Collective Soul	Hints Allegations and Things Left Unsaid	Breathe
1194	Collective Soul	Collective Soul	Gel
1202	**Collective Soul**	**Hints Allegations and Things Left Unsaid**	**Wasting Time**
1370	Counting Crows	This Desert Life	Hanginaround
1351	**Counting Crows**	**Across A Wire - Live In NYC From The Ten Spot CD 2**	**Raining In Baltimore**
1364	**Counting Crows**	**August and Everything After**	**Perfect Blue Buildings**
1352	**Counting Crows**	**Across A Wire - Live In NYC From The Ten Spot CD 2**	**Round Here**
1643	Craig David	Born To Do It	Last Night
1419	Creedence Clearwater Revival	Pendulum	It s Just A Thought
1392	**Creedence Clearwater Revival**	**Cosmo s Factory**	**Before You Accuse Me**
1524	Cypress Hill	IV	Dead Men Tell No Tales
1540	Cypress Hill	Live at the Fillmore	Riot Starter
1580	**D'Angelo**	**Voodoo**	**Chicken Grease**
1611	Dave Matthews Band	Live at Red Rocks 8 15 95 Disc 1	Best Of What s Around
1620	**Dave Matthews Band**	**R.E.M.ember Two Things**	**The Song That Jane Likes**
1691	Def Leppard	Adrenalize	I Wanna Touch U
1724	Deftones	White Pony	Rx Queen
1797	**Depeche Mode**	**People Are People**	**People Are People**
1895	Disturbed	The Sickness	Down With The Sickness
1893	**Disturbed**	**The Sickness**	**Voices**
1909	**Dixie Chicks**	**Wide Open Spaces**	**Never Say Die**
1915	**Dixie Chicks**	**Wide Open Spaces**	**Give It Up Or Let Me Go**
1916	**DMX**	**Flesh Of My Flesh Blood Of My Blood**	**Bring Your Whole Crew**
4068	**Don McLean**	**Favorites And Rarities - Disc 1**	**American Pie**
2000	Dr. Dre	00	Forgot About Dre ft Eminem
2014	Duran Duran	Arena	Hungry Like The Wolf
5037	**Elvis Presley**	**Elvis Christmas Album**	**I Believe**
2138	Enya	Watermark	Orinoco Flow
2146	Erasure	Chorus	Joan
1121	Eric Clapton	Crossroads 2 Disc 4	Kind hearted woman
1105	Eric Clapton	Crossroads 2 Disc 2	Layla
1130	Eric Clapton	Unplugged	Tears in Heaven
1138	Eric Clapton	Unplugged	Old Love
2222	Everclear	So Much For The Afterglow	I Will Buy You A New Life
2239	**Everclear**	**Sparkle And Fade**	**Pale Green Stars**
2261	Everlast	Whitey Ford Sings the Blues	Hot To Death
2244	**Everlast**	**Eat At Whitey s**	**I Can t Move**

2269	**Everlast**	**Whitey Ford Sings the Blues**	**Years**
2285	Everything but the Girl	Amplified Heart	Rollercoaster
2290	Everything but the Girl	Amplified Heart	Missing
2350	Fatboy Slim	You ve Come a Long Way Baby	Kalifornia
2377	**Finger Eleven**	**The Greyest Of Blue Skies**	**Suffocate**
2379	**Finger Eleven**	**The Greyest Of Blue Skies**	**Famous**
2416	Fleetwood Mac	The Dance	Dreams
2471	**Foreigner**	**Agent Provocateur**	**I Want To Know What Love Is**
2569	Garbage	Garbage	Only Happy When It Rains
2637	**Garth Brooks**	**Ropin The Wind The Limited Series**	**Which One Of Them**
2650	**Garth Brooks**	**The Chase**	**Learning To Live Again**
7050	**Gary Wright**	**The Dream Weaver**	**Made To Love You**
2678	Genesis	From Genesis To Revelation Disky version	In The Wilderness
2709	**Genesis**	**Live - The Way We Walk - Volume One - The Shorts**	**Jesus He Knows Me**
2811	Green Day	Dookie	Burnout
3626	**Huey Lewis and the News**	**Fore**	**I Never Walk Alone**
3122	**Ja Rule**	**Venni Vetti Vecci**	**World s Most Dangerous feat Nemesis**
3183	Janet Jackson	Rhythm Nation 1814	Someday Is Tonight
4805	Jennifer Paige	Jennifer Paige	Always You
4809	Jennifer Paige	Jennifer Paige	Between You and Me
287	Jessica Andrews	Who Am I	Who Am I
2896	**Jimi Hendrix Experience**	**Are You Experienced**	**The Wind Cries Mary**
1171	Joe Cocker	Joe Cocker Live	When The Night Comes
1734	John Denver	An Evening With John Denver - Disc 2	Take Me Home Country Roads
6372	**Keith Sweat**	**Keith Sweat**	**Chocolate Girl**
3764	**Kenny Loggins**	**Outside from the Redwoods**	**Now And Then**
3533	La Bouche	Sweet Dreams	Fallin In Love
2309	**Lara Fabian**	**Lara Fabian**	**I am Who I am**
2950	Lauryn Hill	The Miseducation of Lauryn Hill	Final Hour
3557	**Led Zeppelin**	**In Through The Out Door**	**Carouselambra**
3562	**Led Zeppelin**	**Led Zeppelin I**	**You Shook Me**
3614	Les Rythmes Digitales	Darkdancer	Take a Little Time
3618	Les Rythmes Digitales	Darkdancer	Sometimes
3644	**Lifehouse**	**No Name Face**	**Sick Cycle Carousel**
3652	**Lifehouse**	**No Name Face**	**Quasimodo**
3692	**Live**	**The Distance To Here**	**Run to the Water**
3712	**Live**	**Throwing Copper**	**Waitress**
3752	LL Cool J	mr smith	I Shot Ya
3716	**LL Cool J**	**G O A T**	**Imagine That**
3717	**LL Cool J**	**G O A T**	**Back Where I Belong**
555	Lou Bega	A Little Bit Of Mambo	Mambo Mambo
553	**Lou Bega**	**A Little Bit Of Mambo**	**The Trumpet Part II**
3814	Lynyrd Skynyrd	Lyve From Steel Town CD 1	Saturday Night Special
3823	Madison Avenue	Polyester Embassy	Who The Hell Are You Original Mix
3922	**Marilyn Manson**	**Holy Wood**	**Coma Black**
3949	**Marilyn Manson**	**The Last Tour On Earth**	**Astonishing Panorama Of the Endtimes**
2655	**Marvin Gaye**	**Let s Get It On**	**Let s Get It On**
4115	**Me First and the Gimme Gimmes**	**Are a Drag**	**Stepping Out**
3218	Michael Jackson	Off The Wall	Rock With You
3233	**Michael Jackson**	**Thriller**	**Human Nature**
3219	**Michael Jackson**	**Off The Wall**	**Working Day And Night**
3380	Montell Jordan	Get It On Tonight	let s cuddle up featuring LOCKDOWN
3395	**Montell Jordan**	**This Is How We Do It**	**Down On My Knees**
4347	Mudvayne	L d 50	Prod

4343	**Mudvayne**	**L d 50**	**Internal Primates Forever**
4369	MxPx	On The Cover	No Brain
4392	**Mystikal**	**Let s Get Ready**	**Mystikal Fever**
1835	Neil Diamond	Hot August Night - Disc 1	Sweet Caroline
1846	Neil Diamond	Hot August Night Disk 2	Canta Libre
1840	Neil Diamond	Hot August Night - Disc 1	Shilo
7108	Neil Young	Harvest	Words Between The Lines Of Age
4478	New Radicals	Maybe You ve Been Brainwashed Too	Technicolor Lover
4472	**New Radicals**	**Maybe You ve Been Brainwashed Too**	**I Don t Wanna Die Anymore**
4484	Next	Welcome II Nextasy	Cybersex
4563	**Nine Inch Nails**	**The Fragile Right**	**The Big Come Down**
3355	Olivia Newton-John	Olivia	Summer Nights Grease
4779	**Our Lady Peace**	**Happiness Is Not A Fish That You Can Catch**	**Blister**
4825	Papa Roach	Infest	Broken Home
130	**Paula Abdul**	**Forever Your Girl**	**Opposites Attract**
4838	Pennywise	Straight Ahead	Might Be a Dream
4840	Pennywise	Straight Ahead	Straight Ahead
1220	**Phil Collins**	**But Seriously**	**Heat On The Street**
1218	**Phil Collins**	**But Seriously**	**I Wish It Would Rain Down**
1243	**Phil Collins**	**Hello I Must Be Going**	**Thru These Walls**
4964	**Placebo**	**Black Market Music**	**Passive Aggressive**
5232	Queen	The Game	Save Me
5243	**Queen**	**The Works**	**I Go Crazy**
5191	**Queen**	**Live Magic**	**Is This The World We Created**
5242	**Queen**	**The Works**	**Is This The World We Created**
5436	R.E.M.	Dead Letter Office	Burning Hell
5444	**R.E.M.**	**Dead Letter Office**	**Femme Fatale**
5345	Radiohead	OK Computer	No Surprises
5370	Rage Against the Machine	Renegades	Microphone Fiend
5418	**Rancid**	**and out Come the Wolves**	**As Wicked**
3998	Richard Marx	Repeat Offender	Satisfied
6144	Rod Stewart	Vagabond Heart	Rebel Heart
6154	Rod Stewart	Vagabond Heart	If Only
6152	**Rod Stewart**	**Vagabond Heart**	**Have I Told You Lately**
5531	Rolling Stones	Tattoo You	Worried About You
5557	Roxette	Look Sharp	Dance Away
5545	**Roxette**	**Joyride**	**soul deep**
5611	Run-D.M.C.	Raising Hell	Hit It Run
5630	**Sade**	**Love Deluxe**	**Like A Tattoo**
5656	**Sade**	**Sade LOVERS ROCK**	**LOVERS ROCK**
5708	**Savage Garden**	**Affirmation**	**The Animal Song**
5738	**Scorpions**	**World Wide Live**	**Make It Real**
5797	Seven Mary Three	American Standard	Anything
6609	Shania Twain	Come On Over	Honey I m Home
6615	**Shania Twain**	**The Woman In Me**	**Home Ain t Where His Heart Is Anymore**
1460	**Sheryl Crow**	**Live from Central Park**	**There Goes The Neighborhood**
5857	Sisqo	Unleash The Dragon	Unleash The Dragon feat Beanie Sigel
5949	Soul Asylum	Grave Dancers Union	Somebody To Shove
6043	Spineshank	Strictly Diesel	Slipper
6047	Spineshank	Strictly Diesel	While My Guitar Gently Weeps
7014	Steve Winwood	Back in the High Life	Split Decision
7026	Stevie Wonder	Songs in the Key of Life Disc 2	Isn t She Lovely
7031	**Stevie Wonder**	**Songs in the Key of Life Disc 2**	**As**
7020	**Stevie Wonder**	**Songs In The Key Of Life Disc 1**	**Sir Duke**
6205	Stone Temple Pilots	Tiny Music Songs from the Vatican Gift Shop	Adhesive
6217	Stroke 9	Nasty Little Thoughts	One Time

6245	Styx	Return To Paradise Disc 2	Fooling Yourself The Angry Young Man
6246	Styx	Return To Paradise Disc 2	Show Me The Way
6258	Styx	The Grand Illusion	Come Sail Away
378	**The Bangles**	**Different Light**	**Following**
530	The Bee Gees	Here At Last Bee Gees Live Disc Two	Down The Road
904	The Cardigans	Gran Turismo	Starter
1031	**The Chemical Brothers**	**Surrender**	**Out of Control**
1284	The Corrs	In Blue	Somebody for someone
1386	The Cranberries	No Need To Argue	Ridiculous Thoughts
1388	**The Cranberries**	**No Need To Argue**	**Yeat s Grave**
2284	**The Everly Brothers**	**The Fabulous Style of**	**All I Have To Do Is Dream**
3006	The Human League	The Very Best of	Heart Like A Wheel
4997	The Police	Live Disc One - Orpheum WBCN Boston Broadcast	Hole In My Life
5004	The Police	Live Disc Two - Atlanta Synchronicity Concert	Walking In Your Footsteps
5014	The Police	Live Disc Two - Atlanta Synchronicity Concert	So Lonely
5024	The Presidents of the United States of America	unknown	Body
6881	**The Verve**	**Urban Hymns**	**Weeping Willow**
4022	Tim McGraw	A Place In The Sun	Somebody Must Be Prayin For Me
6500	**Tina Turner**	**Tina Live In Europe CD 1**	**What s Love Got To Do With It**
6564	TLC	FanMail	Don t Pull Out On Me Yet
3447	**Toby Keith**	**How Do You Like Me Now**	**Do I Know You**
778	Toni Braxton	Secrets	Come On Over Here
795	**Toni Braxton**	**Toni Braxton**	**I Belong to You**
6577	Tool	Aenima	Stinkfist
6582	**Tool**	**Aenima**	**Hooker with a Penis**
6661	**U2**	**All That You Can t Leave Behind**	**Elevation**
6750	Ugly Kid Joe	America s Least Wanted	Cats In The Cradle
6785	Van Halen	98	House of Pain
2884	Wade Hayes	Old Enough To Know Better	Kentucky Bluebird
6917	Westlife	Westlife	I Need You
6973	**White Zombie**	**Supersexy Swingin Sounds**	**Electric Head Pt Satan in High Heels Mix**
2996	Whitney Houston	Whitney Houston	Greatest Love Of All
7073	**Wu-Tang Clan**	**Wu-Tang Forever Disc 2**	**Dog Shit**
7066	**Wu-Tang Clan**	**Enter The Wu-Tang 36 Chambers**	**WuTang th Chamber Part II**
7079	**Wu-Tang Clan**	**Wu-Tang Forever Disc one**	**Reunited**
7094	Xzibit	Restless	Rimz Tirez feat Defari Goldie Loc Kokane
7091	**Xzibit**	**Restless**	**D N A DRUGSNALKAHOL feat Snoop Dogg**

Appendix B

Perceptive test

In the following we append the content of the questionnaire we use for the evaluation of the system.

Janas Music Finder - test

Janas Music Finder is a music search engine based on text-based semantic queries.

Semantic Description → JANAS → songs

The kind of queries Janas can accept are:

- Emotional description: based on words that concern mood and feelings.
Example: I want a song happy and joyful

- Non-emotional description: base on 17 words: *hard-soft, clear-dull, rough-harmonic, dynamic-static, stuttering-flowing, difficult-easy, fast-slow, groovy*
Example: something hard and easy

- It is possible to specify qualifiers for both emotional and non emotional descriptors. Qualifiers are: *not at all, not, hardly, a little, slightly, partly, somewhat, average, in-between, medium, moderately, fairly, rather, quite, quite a bit, mainly, considerably, very, highly, very much, fully, extremely, completely*
Example: very hard and not happy at all

- Tempo marking through Italian words used in music sheets: *adagio, andante, moderato, allegro, presto* **Example: I want a song that plays allegro**

- Similarity with other songs. Both songs' titles and authors must be written in quotation marks (""). Authors must be indicated immediately after the title of the song they refer to and mist be preceded by the keyword **by**. A link to the list of songs in the database is in the homepage or on Janas logo at the top bar. The research of a song can be made by using just the titles, the titles and the artist, or just the artist. Should the research be performed only by author, the system will ask to specify which songs by the that particular artist are the user wishes to use for its research. This kind of research is based only on semantic-based song similarity and on query specification. It is not based on other parameters such as timbre similarity, mood suggested by lyrics, genre and so on.
  Example: Something like "isn't she lovely" by "stevie wonder"
  I want to listen to some music similar to songs by "the cranberries"

- Comparison with other songs via qualifiers more e less:
  Example: Something like "isn't she lovely" but faster

- Including the keyword **playlistv** in the query, the system will generate a playlist based on the semantic description instead of a list of songs.
  Example: I want a playlist that sounds angry, fast and rough

- A mixture from previous typologies:
  Examples: I'd like something that plays like "isn't she lovely", but in an adagio tempo, more groovy and partly sad

It is necessary for queries to be English, in the form of a sentence with actual significance in order for Janas to better recognize them.

The test is divided in two sections:

1. predefined queries: some predifined queries will be proposed

2. free-text queries: you will be free to try the music search engine

The test requires the use of the search engine for at least ten minutes.

What kind of listener are you? Please choose only one answer

Beginner (I listen to music less than three hours a day)	
Expert (I listen to music mor than three hours a day)	
Professional (I listen to music also for reasons related to my job)	

Predefined queries

You have to evaluate the quality of results with a mark in a 9 point-scale, where 1 means very bad and 9 is the optimum. Quality is intended as the correspondence of songs results with respect to the query content. 5 indicates a neutral mark.

Query	1	2	3	4	5	6	7	8	9
I want a song very groovy and happy									
I want a song not happy at all, dull and flowing									
I want a playlist that sounds angry, fast and rough									
I would like to listen to calm songs, like "Orinoco Flow", flowing and slow									
I want a playlist not angry, and not stuttering and with a slow tempo									

Free-text queries

Feel free to try some queries and do evaluate the performances.

	1	2	3	4	5	6	7	8	9
Please indicate the general evaluation on the results obtained when using free queries. The evaluation is intended on the correspondence between query and results.									
Do you think this system is useful? (1: not at all - 5 can't really say - 9 : very useful)									
Would you ever use this kind of system? (1: not at all. 5: I don't know. 9: Yes, very often)									
Taking into account the results, the idea of semantic research and the implementation, the functionalities, usefulness and potentials, how do you evaluate the system in general? (1: very bad. 5: neutral. 9: very good)									

Please indicate optional notes

Please, fold this sheet before giving it back, in order for the answers to be hidden.

1

Bibliography

[1] H. H. C. Yi-Hsuan Yang, *Music Emotion Recognition.* CRC Press, 2011.

[2] M. L. B. D. B. H. D. M. J. P. M. M. Lesaffre, L. De Voogdt, "How potential users of music search and retrieval systems describe the semantic quality of music," *Journal of the american society for information science and technology*, pp. 697–707, 2008.

[3] B. Rohrmann, "Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data," *Project Report. University of Melbourne, Australia*, 2003.

[4] T. Grill, A. Flexer, and S. Cunningham, "Identification of perceptual qualities in textural sounds using the repertory grid method," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM '11. New York, NY, USA: ACM, 2011, pp. 67–74. [Online]. Available: http://doi.acm.org/10.1145/2095667.2095677

[5] G. Prandi, A. Sarti, and S. Tubaro, "Music genre visualization and classification exploiting a small set of high-level semantic features."

[6] D. T. G. L. D. Turnbull, L. Barrington, "Towards musical query-by-semantic-description using the cal500 data set," *SIGIR 2007 Proceedings, Session 18: Music Retrieval*, July 2007.

[7] D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet, "Identifying words that are musically meaningful."

[8] C. Mckay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets," in *In Int. Conf. on Music Information Retrieval, ISMIR 2004*, 2004, pp. 525–530.

[9] Celma, "Foafing the music bridging the semantic gap in music recommendation," in *5th International Semantic Web Conference (ISWC)*, Athens, GA, USA, 2006.

[10] L. Chiarandini, M. Zanoni, and A. Sarti, "A system for dynamic playlist generation driven by multimodal control signals and descriptors," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, 2011, pp. 1–6.

[11] A. L. chun Wang and T. F. B. F, "An industrial-strength audio search algorithm," in *Proceedings of the 4 th International Conference on Music Information Retrieval*, 2003.

[12] M. Slaney, "Semantic-audio retrieval," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, May, pp. IV–4108–IV–4111.

[13] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval.* John Wiley & Sons, 2005.

[14] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, Jul.

[15] T. Pohle, E. Pampalk, and G. Widmer, "Evaluation of frequently used audio features for classification of music into perceptual categories," in *In Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05*, 2005.

[16] D. M. Randel, *The Harvard dictionary of music / edited by Don Michael Randel*, 4th ed. Belknap Press of Harvard University Press, Cambridge, MA :, 2003.

[17] J. F. T. Hastie, R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[18] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Tech. Rep.

[19] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* Wiley, 2001.

[20] H. L. Somers, R. Dale, and H. Moisl, *Handbook of natural language processing / edited by Robert Dale, Hermann Moisl, Harold Somers.* Marcel Dekker, New York :, 2000.

[21] D. Jurafsky, J. H. Martin, and A. Kehler, *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition.* MIT Press, 1999, vol. 2.

[22] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing.* Cambridge, MA, USA: MIT Press, 1999.

[23] L. Barrington, D. Turnbull, D. Torres, and G. Lanckriet, "Semantic similarity for music retrieval," in *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[24] L. E. Y. E. Kim, E. Schmidt, "Moodswings: a collaborative game for music mood label collection," *ISMIR 2008 - Session 2c - Knowledge Representation, Tags, Metadata*, 2008.

[25] L. M. Zbikowski, "Modelling the groove: Conceptual structure and popular music," *Journal of the Royal Musical Association*, vol. 129, no. 2, pp. 272–297, 2004.

[26] M. E. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1009–1020, 2007.

[27] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of LREC*, vol. 4, 2004, pp. 1083–1086.

[28] J. Cu, R. Cabredo, R. Legaspi, and M. Suarez, "On modelling emotional responses to rhythm features," *PRICAI 2012: Trends in Artificial Intelligence*, pp. 857–860, 2012.

[29] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[30] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 423–430. [Online]. Available: http://dx.doi.org/10.3115/1075096.1075150

[31] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions.* Association for Computational Linguistics, 2006, pp. 69–72.

[32] D. Robinson and K. Coar, "The common gateway interface (cgi) version 1.1," IETF, RFC 3875, October 2004, status: INFORMATIONAL.

[33] A. S. S. T. M. Zanoni, D. Ciminieri, "Searching for dominant high-level features for music information retrieval," *20th European Signal Processing Conference*, 2012.

[34] D. Totaro, "Example-based definition of high-level descriptors of musical excerpts," Master's thesis, Politecnico di Milano, 2010-2011.

[35] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[36] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," in *Computer Music Journal*, 2003.

[37] F. E. Grubbs, "Procedures for detecting outlying observation in samples," *American Statistical Association*, pp. 1–21, February 1969.

[38] D. H. B. Iglewicz, *How to detect and handle outliers.* ASQC Quality Press, 1993.