Turn on the Base

Institute for Perception Research
PO Box 513 - 5600 MB Eindhoven

20.02.1996

Rapport no. 1094

Turn on the Base
(Project Evaluation)

Steffen Pauws

Voor akkoord: D.G. Bouwhuis

Steffen Pauws

Turn on the Base

(Project Evaluation)

Issue Date	:1996-20-02
Status	: Approved
Classification	: Company Restricted
Reviewers	: Berry Eggen
Issued by	: Institute for Perception Research

Abstract

This document is the fourth deliverable of the project 'Turn on the Base' as mentioned in the Project Contract **[Pauws1]**. It is the formal conclusion of the project evaluation phase.

The project 'Turn on the Base' is aiming at research and development of methods to access and select relevant information items in multi-media applications for home entertainment environments. We try to find tools that help to solve the selection problem which arises when the user is confronted with large amounts of information. This problem will be tackled with innovative techniques such as behaviour-based artificial intelligence and agent-based engineering (autonomous and situated agents). The CD jukebox demonstrator is chosen as a research vehicle for the first year of the project.

Contents

1 Introduction

This document represents the evaluation of a new functionality for a CD jukebox player. We have named this functionality PATS (Personalized Automatic Track Selection). It complements the random-play, program, and FTS (Favourite Track Selection) functionality already found on current products. By switching on PATS, the player will automatically select CD tracks. Which tracks will be selected is determined by the music preferences of the user. Actually, to support the user with the selection of favourite tracks, the system has to adapt automatically to these preferences in order to make an interesting track programme (personalized listening session). The user intervention for expressing his or her music preferences is kept as minimal as possible. Situated and autonomous agents are used to enable PATS. We are convinced that the PATS feature is also relevant for future TV sets (favourite TV program selection) and their peripherals such as set-top boxes (favourite service selection).

Coherence and variation between music tracks are recognized key factors for the success of PATS. The prime objective of the evaluation is to elicit the added value of PATS and validate its adaptivity to the music preferences of a music listener. The added value should aim at establishing a contribution to different activities of users such as active and passive music listening, and personalizing a compilation for a special occasion (party, present).

In Chapter 2, we survey (or acquaint new readers with) the past design decisions and the actual PATS functionality. In Chapter 3, some statistics of the music collection attributes are drawn to obtain a deeper insight into the contents. In Chapter 4, the induction algorithm and its modified version are quantitatively as well as qualitatively evaluated. In Chapter 5, the first steps to an overall functionality evaluation are set. Finally, Chapter 6 is devoted to discussion, recommendations, and reconsiderations.

2 Brief retrospection

In an exploratory study **[Eggen]**, Eggen pointed out that people enjoy accommodations that influence, or even ease, their music selection decisions such as advice from others, thematic radio programmes, or well-known music. In contrast with that, current CD player functionalities such as 'Favourite Track Selection (FTS)' and 'Shuffle/Random Play' are not longer adequately addressing user intentions regarding listening to music. This observation is even emphasized considering the ever growing amount of CD material and storage capacity of future home systems. In this perspective, Eggen already came up with a new functionality named 'Personalised Automatic Track Selection (PATS)' for CD audio players with jukebox capacity. This innovative feature should automatically make compilations from a music collection at home addressing the music preference of the listener(s). Hence, PATS can be an appreciated supplement to the present-day repertoire of CD player functionalities.

Cornerstone of PATS is applying more *coherence* amongst the offered music tracks than can be obtained in a random selection, but on the other hand evoking more *variation* over time amongst the offered selections than FTS can do without any need for explicit programming by the user. Both coherence and variation are recognized key factors for addressing music preference **[Eggen]**.

We distinguish *musical taste* from *music preference* **[Pauws2]**. Musical taste refers to a long-term commitment to a particular broad music style. It is partly affected by factors such as musical training or majority consensus. As such, we assume that musical taste is addressed by the music collection at home. On the other hand, music preference is defined as a person's liking of certain music at a certain moment. Consequently, it is influenced by the current activities and situations the user is involved in.

Music preference turns out to be a rather subjective matter, is hard to express verbally, and can be dominated by a strong personalised referential meaning of the music, i.e. a piece of music can be associated with certain moods or experiences of the listener. People tend to hear music in terms of melody, rhythm, texture, timbre, and atmosphere. However, we are convinced that guaranteeing coherence can address, to some extent, an individual's preference. Coherence draws forth a degree of expectation on the offered PATS compilation. It is operationally based on common objective attributes such as musicians, instruments, and music style amongst music tracks. This implies that the salient attributes for expressing an individual's music preference should be learned by the system. On the other hand, variation is based on the principle that people do not want to listen to the same sequenced material over and over again. PATS should elicit surprising effects from the music collection at home such as rediscovering forgotten music.

At first, coherence and variation seem to be rather opposite prerequisites. Whereas coherence stands for a criterion to order things, variation might initiate processes that undo this order. In fact, the on-going combination of building up and breaking down temporal structures amongst music tracks can be exploited in a self-organising fashion. Random perturbations initiated by breaking down structures should excite the system over and over again. To make these ideas on self-organisation more tangible, we started off from the framework of multi-agent systems:

- A multi-agent system is composed of agents, small autonomous entities with local situated behaviours.
- The agent's behaviours are stochastic and interactive in nature.
- A large population of agents is required to obtain deterministic regularities *emerged* from these interacting behaviours.

Therefore, we metaphorically designated an individual agent for each music track. Beside the attributes describing the track contents, the agent is equipped with behaviours that makes it wandering through a virtual 2-dimensional space and observing other track agents. Autonomously and continuously, agents can start or stop to following (i.e. stay close to) other agents. This follow behaviour is inspired by the observed local similarity[1] between the track attributes. Interactions amongst these follow behaviours give rise to decentralized cluster phenomena and should reconcile the requirement for both coherence and variation.

A cluster that is apt to some coherence, but additionally is ever changing due to the perturbations of its members, is considered a good candidate to be presented as a listening programme preferred by the user. Of course, the user has the last say in what track is good or bad for accommodating his/her intentions. At present, positive and negative user feedback is established by explicit judgement of the tracks. This leaves the system with a list of judged tracks that is subjected to an inductive machine learning algorithm wrapped into a user agent. This induction mechanism reveals the most salient attribute-values that distinguish the good from the bad tracks. Consequently, we have means to express the user music preference. This knowledge is used to focus more on these preference-related attribute-values while calculating similarities between music tracks. This cycle of automatic music compilation and explicit user judgement is iterated enabling the PATS functionality to adapt to (more than one) music preferences. In concrete terms, clusters should arise each addressing a particular music preference based on common attribute-values (common musicians, tempo, style etc.). Currently, user interaction is basically one of 'cognitive economy'; the listener has only to select from a list what music track is the best exemplar to meet his/her preference. The system subsequently presents the cluster that contains this 'prototypical' track as a listening programme.

The process of implementing PATS is fully documented **[Pauws1,Pauws2,Pauws3]**. This report describes the evaluation of PATS. Especially for the purpose of evaluation (and demon-

[1] For this application, we have defined similarity as a weighted sum of common attribute-values in which the weights express the salience of the attribute-values with respect to the user music preference.

stration), we gathered a music collection from numerous CDs. Each music track is attributed by data from CD-booklets, discographies, or acquired by careful listening.

3 Music collection statistics

The music collection comprises 300 one-minute excerpts from 100 jazz albums. Starting-point was to make up a single collection that reflects a personal music collection and to fulfil our need for an application carrier. Ideally, each future subject should be welcomed with his/her own collection at the user tests. For obvious reasons, the provision of this ideal situation can not be achieved in one go.

Although we have put a lot of effort in acquiring the contents, we have only partly succeeded in assembling a collection of reasonable size with attribute overlap. If more overlap or a more uniform distribution of attribute-values is required, the following statistical diagrams are helpful in acquiring new material.

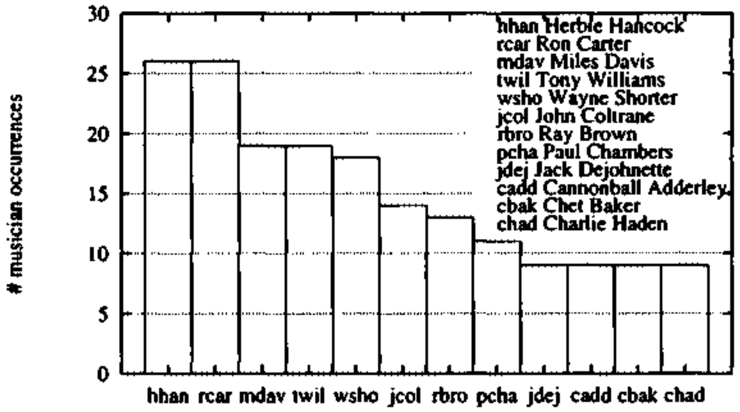


Figure 3.1 Twelve most prominent musicians in the music collection.

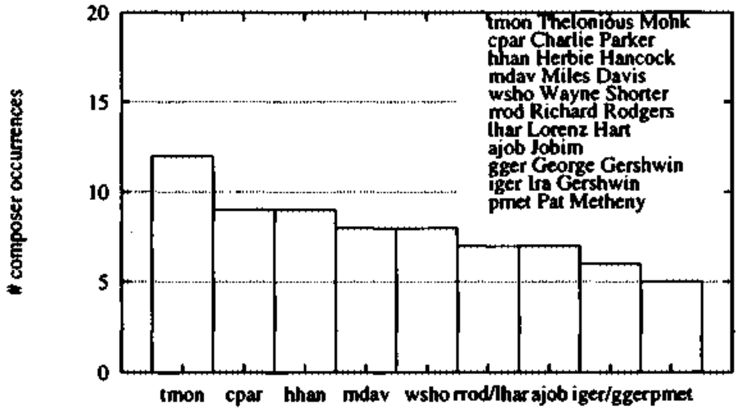


Figure 3.2 Nine most prominent composer (duos) in the music collection.

In Figure 3.1, the twelve most prominent musicians on the recordings are shown. Many persons in this list are considered accompanying musicians and not leading soloists. It should be noted that there is a small overlap in notable musicians.

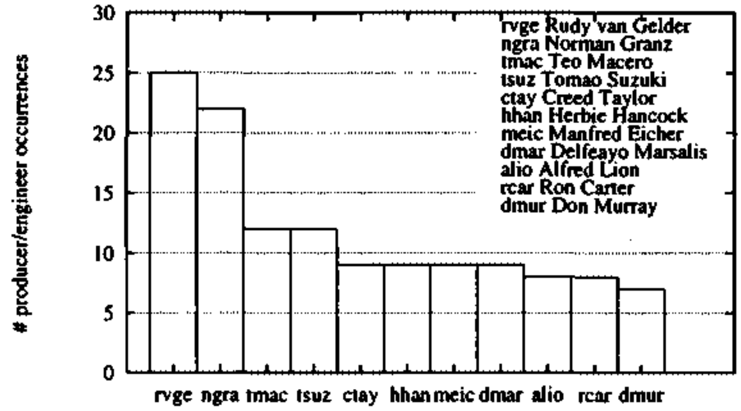


Figure 3.3 Eleven most prominent producers/engineers in the music collection.

The same observation can be drawn by looking at the list of most prominent composers (see Figure 3.2) and producers/engineers (see Figure 3.3). It is apparent that Herbie Hancock has (accidentally) played a central role in the music acquisition.

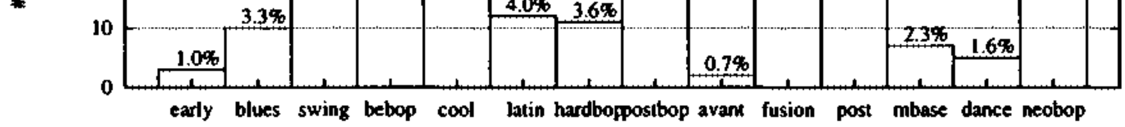


Figure 3.4 The jazz styles in the music collection.

As shown in Figure 3.4, some jazz styles are more represented than others. The collection shows some bias towards the music taste of a particular jazz lover. It should be advantageous to have a more uniform distribution of styles for adequately covering tastes of more subjects. As already mentioned, this can be sorted out in the future by taking notice of the statistics.

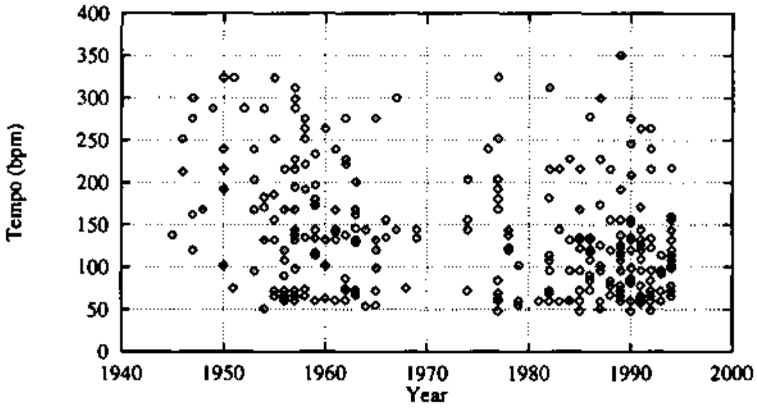


Figure 3.5 Scatterplot of recording year and tempo of the music.

The scatterplot in Figure 3.5 does not suggest a correlation between the year of recording and the tempo of the music. However, the acquisition has been focused on rather modish material (around 1990) with a relatively slow tempo and the heyday around 1960. Both the periods around 1950 and 1970 are not well represented.

4 Induction algorithm evaluation

The applicability of the ID3 algorithm is investigated in this real-world domain along two lines: quantitative and qualitative examination. The quantitative examination provides us with 'down-to-earth' estimation of overall performance, whereas the qualitative approach give us more insight in the induced general constructs on music preference.

4.1 Modification to ID3

The ID3 (Induction for Decision Trees) is one of the many decision-tree generation algorithms that are around. In our application, it is used to reveal the set of attributes that most accurately distinguish preferred and non-preferred music tracks. By starting off with a listening programme (training set) in which each track is explicitly labelled as preferred or non-preferred, the idea is to recursively partition this set into disjoint subsets along the track attributes. This means that at each partition an attribute has to be chosen that brings some clarity on what attributes a person's music preference is based. Eventually, each subset exclusively contains preferred or non-preferred tracks. As a result, a decision tree is build in which paths from root to leaf are annotated with the salient attribute-values for music preference.

The ID3 algorithm can be characterised as a greedy, heuristic, hill-climbing search method without any back-tracking facilities; it takes the best opportunity at each node without even pondering whether it should be more valuable at the end to choose for another possibility. A challenge is to find remedies for problems that result from this greediness.

One of the problems encountered was ID3's generation of rather trivial trees that made no sensible contribution to our application **[Pauws3]**. While splitting a track set along the values of a 'best' attribute, some values can be considered highly relevant for expressing someone's music preference, whereas other may not. For instance, if the attribute 'composer' has been chosen for partitioning a track set, it is not unthinkable that only some notable composers matter. As a consequence, we feel that it is more appropriate to partition tracks in one or two sets that have a certain common ground (e.g. by considering only the prominent jazz styles or musicians) and a remaining set that lack this common ground instead of dispersing all tracks in multiple sets (e.g. by considering *all* possible jazz styles or musicians).

Another problem is the existence of unknown attribute-values **[Pauws3]**. We tried to remedy this by replacing the unknown attribute-value by the most common value as present in the track set currently subjected to ID3. Unfortunately, but not surprisingly, this ad hoc solution brought highly suspicious information into the trees.

In general, it was observed that the decision trees as generated by the original ID3 algorithm were too specific to the listening programme at hand (i.e. the training material). To address the problems with a minimal re-design effort, we have applied some small changes to the ID3 algorithm. In Figure 4.1, an explanation by example is shown. Basically, the algorithm is similar except not every attribute value is used to branch on. Only values that appear to be relevant for expressing someone's music preference are considered candidates to branch on whereas for other (irrelevant) attribute-values a default branch is designated. We have defined *relevance* of an attribute-value simply as a threshold expressing the minimal number of tracks in a given set that must hold that value. Considering the objective of our ID3 application and the typical size of a listening session (10-15), we have fixed this threshold on two. Thus, minimal two tracks must hold the attribute-value before this value is a candidate to branch on. Values that do not exceed this threshold are treated as a so-called 'otherwise' value for which a default branch is devised. Actually in these cases, the choice of a best attribute is postponed to a lower tree level. As a result, the average branching factor is highly diminished resulting into better, more generic trees. In addition, the provision of a default branch gives us a good treatment for unknown attribute values; just treat them as 'otherwise' values.

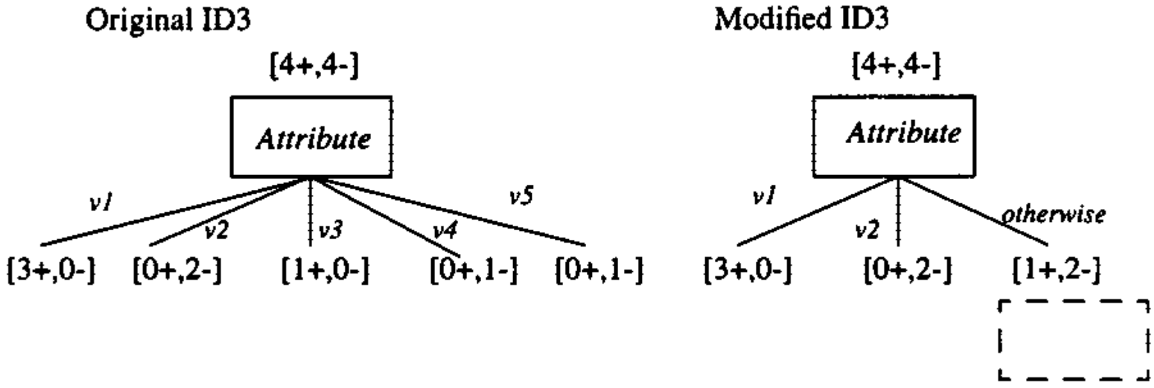


Figure 4.1 The difference in building a decision tree between the original ID3 and the modified version. Both versions start off with a listening programme consisting of 4 preferred and 4 non-preferred tracks. The original ID3 comes with a most decisive attribute with five offsprings whereas the modified version has devised an 'otherwise'- branch for those attribute-values (v3,v4,v5) that are not considered relevant. As a consequence, the resulting subset has to be further partitioned along a (possibly) other attribute as indicated by the dashed box.

4.2 Quantitative evaluation

A quantitative assessment of the overall performance of the induction algorithm was required to inform us about

- the improvement established by the modified version of ID3 with respect to the original one.
- the prediction on music preference using the decision trees. In other words, what a single listening session could say about the remaining part of the music collection in quantitative terms.
- the quality of the decision trees in terms of simplicity.

The prime objective of this evaluation is to find out the overall improvement of the modified ID3 version with respect to the original one. It should be stressed that the application of the algorithm in this evaluation (and as it was originally intended by Quinlan) totally differs from its use in our PATS setting. Within the PATS implementation, it is not used as a strict classification scheme for large amount of data, but merely as a tool for picking out the salient feature that distinguish the good from the bad in a small amenable data set. When a subject is confronted with a considerable amount of music over a long period of time, as is the case in this evaluation, it is unavoidable that the subject uses various music preference (or even musical taste) criteria intermingled. This might result in inconsistent, or even contradictory preference decisions when it comes to our objective attributes. This stands in contrast with the starting-point of PATS, in which we want to present a small compilation that the subject might find appropriate for accommodating his/her current preference criterion. It is therefore prudent to interpret the evaluation results in this perspective.

We have invited 4 subjects to listen to the total music collection (300 one minute excerpts = 5 hours of music) in 20 random sessions of 15 tracks. While listening, they were encouraged to judge the tracks.

Subsequently, we have compiled random sets varying in relative size (ranging from 8 (2.5%) to 60 (20%) tracks) from each subject's judgements. These sets were used as training set for building up a decision tree. This tree was used to classify (label) the remaining ‘unseen’ tracks, the so-called test set. To rule out the statistical variability, 100 trials were conducted for each training set size.

Five performance measures are calculated in each trial: accuracy, number of interior nodes, number of leaves, depth, and support. Prediction on music preference is expressed as the *classification accuracy* of a decision tree. It is defined as the percentage of test tracks that are correctly identified as preferred or non-preferred by the tree. Except for some tracks with unknown attribute-values, we assume that 100% correctness is guaranteed for tracks that are used for building up the tree (self-test accuracy). Trials with different relative training set sizes (2.5%, 5%, 10%, 20%) are conducted for finding appropriate answers on how many tracks are required for establishing a sensible decision tree. In addition, we have scored the attributes to find out what attributes are typically chosen for building up a tree.

The tree quality is measured by quantities such as the *number of interior nodes*, *number of leaves*, *depth*, and *support* of the trees. In general, the total number of nodes should be minimized considering the lurking irrelevant details (and branching) introduced by attributes with broad domains. The number of interior nodes characterizes the number of decision points. The number of leaves corresponds with the number of decision rules (paths from root to leaf) that are generated. The average depth expresses the average number of pre-conditions per rule. A low average depth means that highly discriminating attributes are found. Average support per rule is expressed by the average number of (unseen) tracks on which a particular rule was applicable. This measure should be maximized considering the targeted generic property of a decision tree.

4.2.1 Accuracy

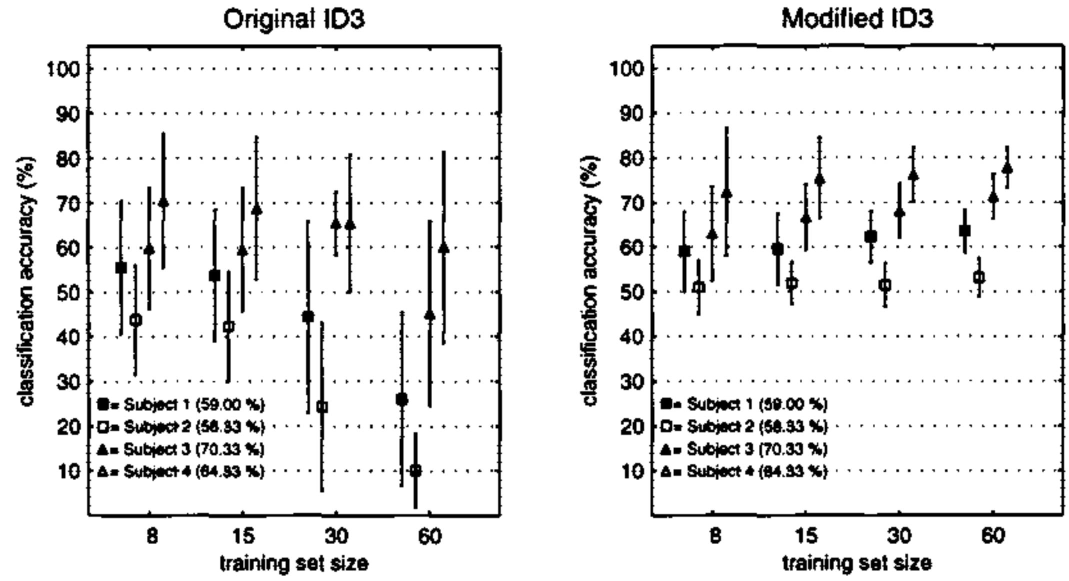


Figure 4.2 Classification accuracy of the original and modified ID3 algorithms.

The percentage of liked material for each subject is mentioned in Figure 4.2 and ranges from 58.33% for subject 2 to 84.33% for subject 4. This implies that by using a simple constant categoriser, which returns the label 'preferred' regardless of the track, outperforms in most cases this ID3 approach. It must be concluded that by looking at the accuracy, and as a consequence the prediction, the ID3 algorithm is impractical as a classification scheme for preference decisions for the entire music collection (see also the evaluation objectives as mentioned in Section 4.2). In particular, the original ID3 algorithm shows a dramatic drop in accuracy when the training set is enlarged. This observation has led to the modification of the ID3 algorithm by making the tree less specific to its training material. The modified ID3 algorithm is not afflicted with a performance deterioration at larger training set sizes.

One can conclude that the present ID3 algorithm and the attribute set are not sufficient to do predictions in this simple two-class problem. However, our intention is not to use the ID3 algorithm as a scheme for classifying the whole music collection but to reveal only those attribute-values that might be important to state a person's preference in a particular listening programme. In addition, the training sets are randomly created, lacking any form of pre-defined coherence.

4.2.2 Tree quality (simplicity)

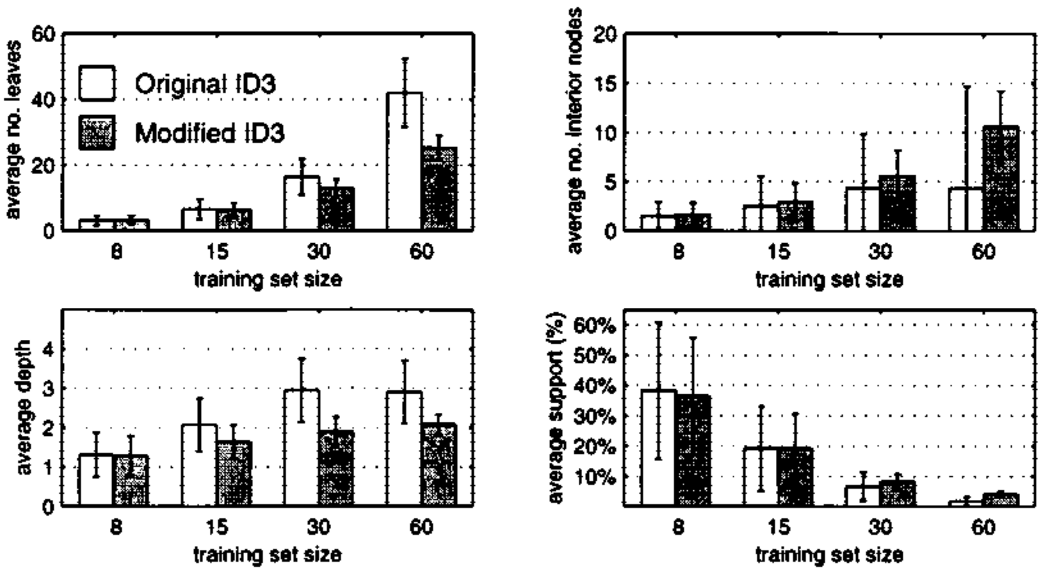


Figure 4.3 Tree quality of the original and modified ID3 algorithm.

In Figure 4.3, the tree quality (or simplicity) is expressed in the average number of nodes (interior as well as terminal), average depth of a path from root to leaf, and average support of the tree for both algorithms. Averages were calculated over all four subjects and trials. Prime motivation to modify the ID3 algorithm was to make it less specific to the training set (i.e. a listening programme). Trees build by the original ID3 tend to be too broad as expressed by the average number of leaves. This is remedied to some extent by the modified version at the expense of more interior nodes (decision points). However, the average depth of the tree is decreased meaning that there are less decision points per path from root to leaf (pre-conditions per rule). Thus, less attributes need to be inspected before a preference classification can be given. In addition, the average support shows some increase meaning that more tracks are applicable to a particular rule: the tree is more generic.

4.2.3 Attribute score

In Figure 4.4 and Figure 4.5, the attribute scores of both algorithms are shown. They simply express how many times a particular attribute occurs at one of the decision points within a decision tree. In Figure 4.4, it is apparent that the cause of the high specific property at large training sets is principally caused by the attribute *Title*. It needs no further explanation that the title of a track is too specific as classifier and does not say anything about other tracks. This phenomenon is highly suppressed by the modified version (see Figure 4.5). Another observation is that the quantitative attributes such as *NumMscns*, *Tempo*, and *Year* are predominant. We feel that this is rather an artefact of our approach than an attribute that is highly important. In general, these attributes are favoured by their minimal branching factor of only two and, hence, need to be divided into categories. In the case of *NumMscns*, we can think of categories such as duos, quartets, and big orchestras. In the case of *Tempo*, we can define categories such as very slow, slow, fast, and very fast. In the case of *Year*, we can hold on to jazz eras. However, the modified ID3 has already remedied the favouring of attributes *Tempo* and *Year* but did not find any improvement for *NumMscns*.

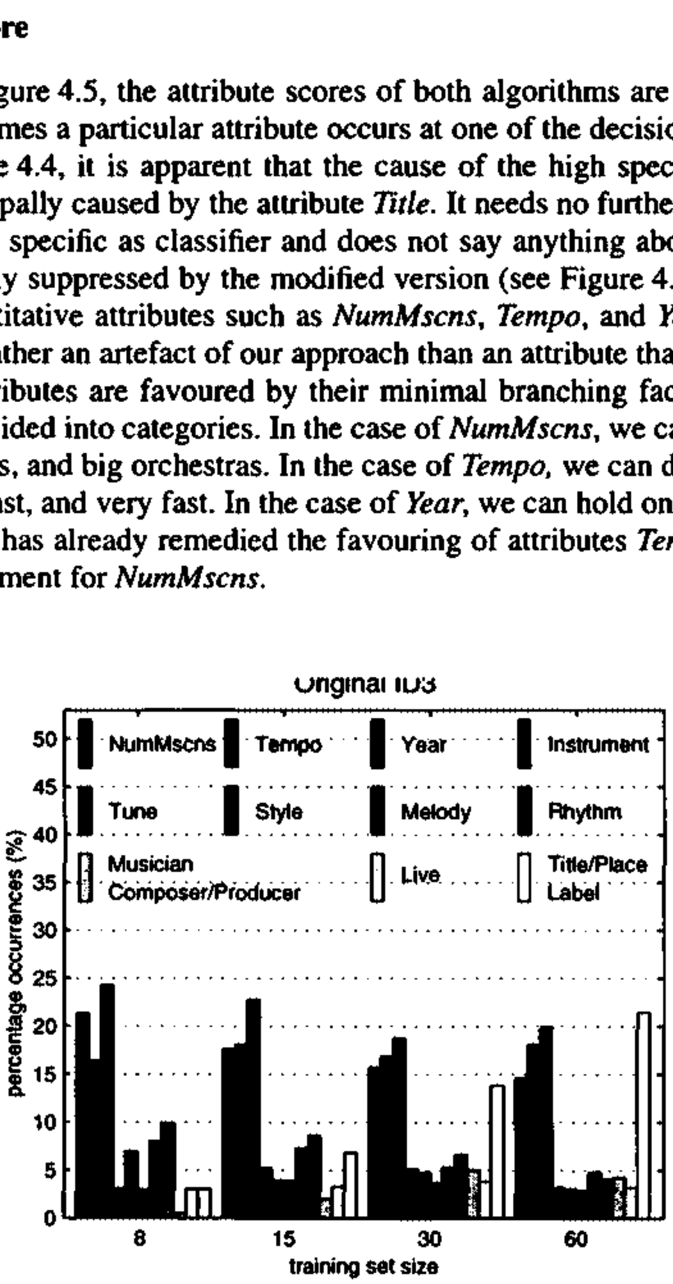


Figure 4.4 Attribute score of original ID3 algorithm.

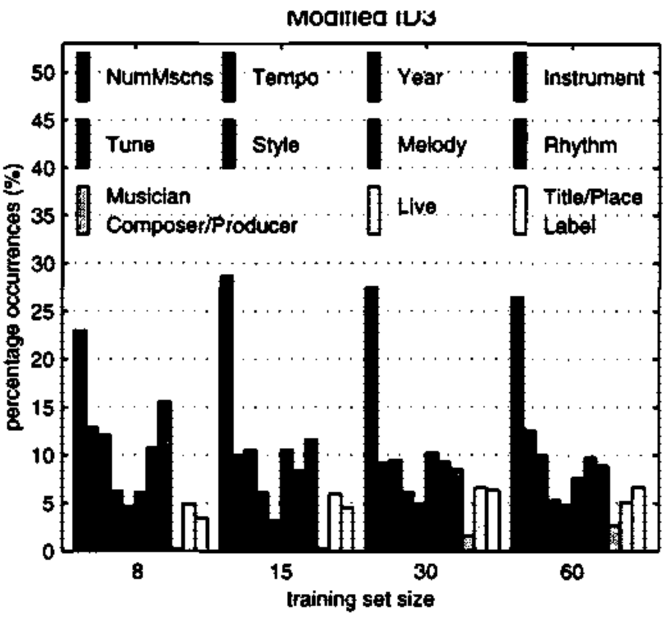


Figure 4.5 Attribute score of modified ID3 algorithm.

As a last remark, we mention that until now no hierarchical concepts are defined for the instruments. It might be profitable to recognize soprano, alto, baritone, and tenor saxophones simply as saxophones. In turn, saxophones can be categorized as a type of (reed) wind-instrument. Also, it is not explicitly annotated whether a particular instrument is acoustically or electrically amplified. In other words, specific domain knowledge is lacking.

4.2.4 Findings

We use the ID3 algorithm for other reasons than it was originally designed for. This fact becomes clear in this evaluation; the classification accuracy as shown in Section 4.2.1 is insufficient. As already emphasized, this is not particularly due to the algorithm itself, but more to what distinct, but intermingled preference decision criteria subjects use when they are confronted with 5 hours of music chunked into randomly created subsets. In addition, decision trees were constructed from small training set that were randomly chosen from this 5 hour of judged material. Subsequently, these decision trees were used to classify the remaining judged material. Obviously, there was no guarantee for a single preference criterion whatsoever in the set up of this evaluation. In contrast with that, our prerequisite for the ID3 applicability in the PATS context is that small sets are judged by one preference criterion that can (afterwards) be formulated in objective attributes. The validation of this requirement is tackled in Section 4.3.

Nevertheless, the modified ID3 algorithm shows a more generic attitude towards the judged material than the original one. This is reflected in an overall quality improvement and a more uniform distributed attribute score profile of the generated trees.

It is also notified that the constructive type of attributes are important for the algorithm performance. Beside the on-going effort in finding other objective attributes that correspond with subjective music experiences, the types of some current attributes are questionable (see Section 4.2.3): quantitative attributes require a coarser ordinal categorisation and a hierarchical order of instruments might be profitable.

4.3 Qualitative evaluation

The correspondence of the induced decision trees and an individual's formulated music preference are determined by means of an experiment. A subject who is familiar with the music contents compiled 14 listening sessions by hand. While compiling, he took care of establishing coherence amongst the preferred tracks by, for instance, addressing groups of songs to a notable jazz musician or a particular jazz style. In contrast, non-preferred tracks were characterized by not fulfilling this coherence. As a consequence, each compilation contained both preferred and non-preferred tracks, while the preference of the tracks was based on common attribute-values. It was the task of ID3 to find out these attribute-values. Although the subject was aware of the track attributes, he followed his own feeling of what attributes and criterion is important for making up a compilation. Beforehand, the preferred compilation criterion was precisely stated. They were written down (paraphrased) before the compilations were subjected to the ID3 algorithm. The resulting decision trees were discussed with the subject.

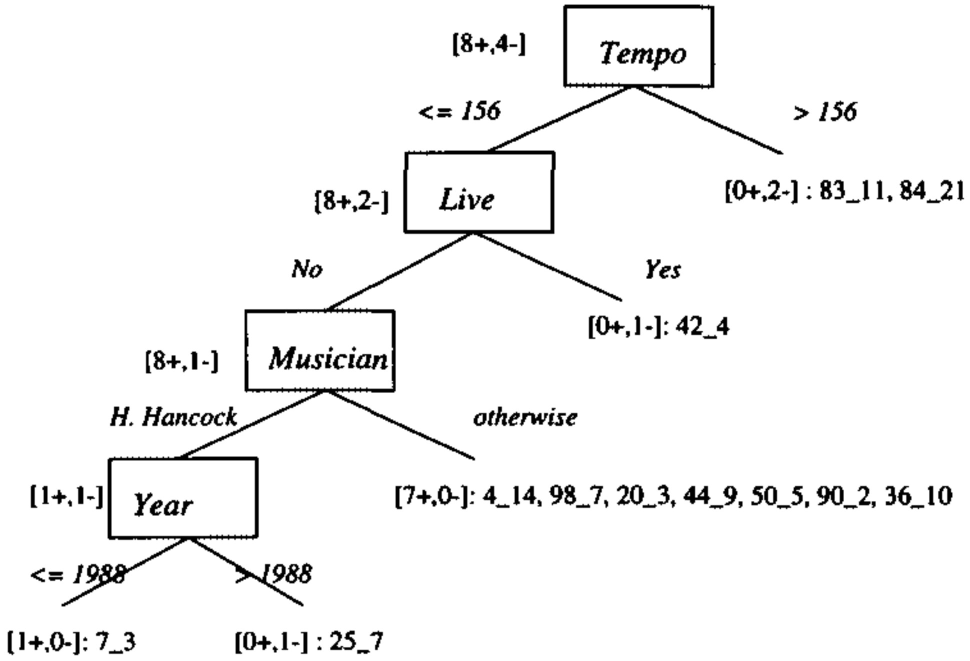
The compilation size is 12. Fourteen (14) compilation sets were composed.

4.3.1 Compilation 1

Compilation criterion: calm background music

Good tracks: (+)		
4_14	'You don't know what love is'	Chet Baker
98_7	'Ill wind'	Ben Webster
20_3	'Blue in green'	Miles Davis
7_3	'Scriabin'	Michael Brecker
44_9	'Ev'ry time we say goodbye'	Charlie Haden
50_5	'Peacocks'	Bill Evans
90_2	'Where are you'	Sonny Rollins
36_10	'Birks works'	Dizzy Gillespie
Bad tracks: (-)		
42_4	'Maiden voyage'	Chick Corea Herbie Hancock
83_11	'Five brothers'	Gerry Mulligan

84_21	'The skunk'	Fats Navarro
25_7	'Parallel realities'	Jack DeJohnette



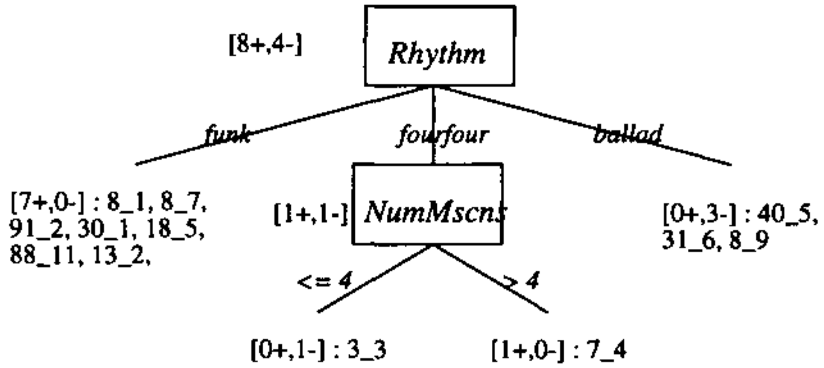
Calm, quiet music for background is conceived as songs with a moderate tempo and preferably not live recorded. The preference for a particular musician or year can not be interpreted.

4.3.2 Compilation 2

Compilation criterion: 'swinging'/solid jazz rock with electric instruments

Good tracks: (+)		
8_1	'Slang'	The Brecker Brother
7_4	'Suspone'	Michael Brecker
80_7	'Northern comfort'	Mezzoforte
91_2	'Better believe it'	David Sanborn
30_1	'Upbeat 90's'	Tom Scott
18_5	'Cockpit'	D-Code
88_11	'Headin' home'	Joshua Redman
13_2	'Down upbeat'	Casiopea
Bad tracks: (-)		
40_5	'Brigitte'	Kirk Lightsey Freddy Hubbard
31_6	'Someone to watch over me'	Ben Webster
3_3	'Sad walk'	Chet Baker

8_9	'And then she wept'The Brecker Brother

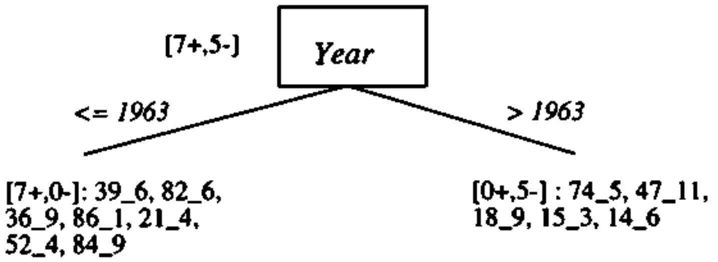


The criterion solid jazz rock is translated in a preference for funk oriented rhythms. Electric instruments are not annotated as such in the music collection.

4.3.3 Compilation 3

Compilation criterion: hard/bebop stylish, somewhat fast but not too fast

Good tracks: (+)		
39_6	'Ornithology'	Charlie Parker
82_6	'Bemsha swing'	Thelonious Monk
36_9	'A foggy day'	Red Garland
86_1	'Confirmation'	Charlie Parker
21_4	'Miles'	Miles Davis
52_4	'Cherokee'	Stan Getz
84_9	'Goin' to Minton's'	Fats Navarro
Bad tracks: (-)		
74_5	'Funky tamborim'	Tania Maria
47_11	'Big boy'	Hans Dulfer
18_9	'Charlie and Martino	Boehlee'D-Code
15_3	'Slow body poppin''	Billy Cobham
14_6	'Fridge blues'	Philip Catherine

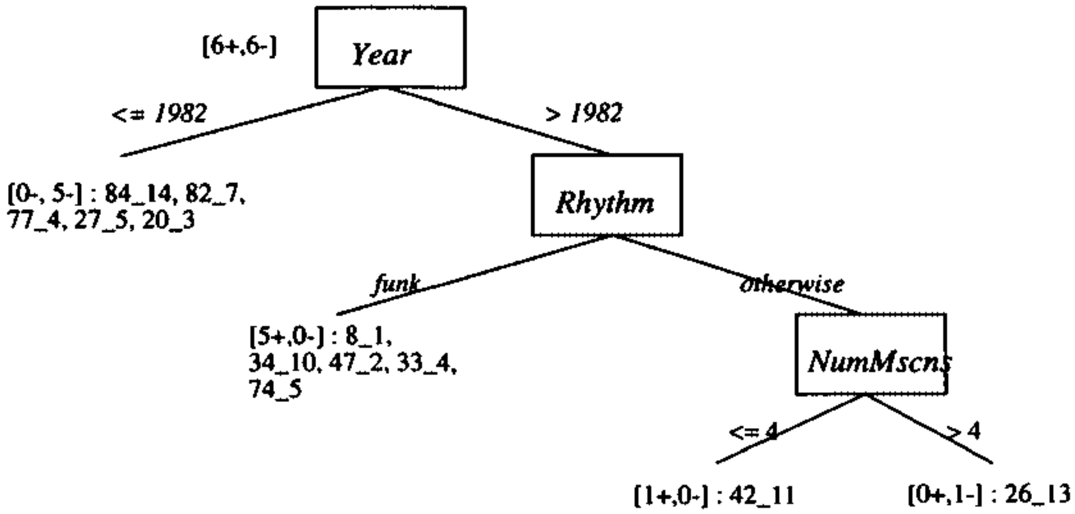


The heyday of hardbop and bebop is considered to be somewhere before the year 1963. The criterion of preferred tempo is not revealed.

4.3.4 Compilation 4

Compilation criterion: modish and dance

Good tracks: (+)		
8_1	'Slang'	The Brecker Brother
34_10	'Who dares'	Steve Williamson
42_11	'Why not'	Pee Wee Ellis
47_2	'Streetbeats'	Hans Dulfer
33_4	'Redneck'	The James Taylor Quartet
74_5	'Funky tamborim'	Tania Maria
Bad tracks: (-)		
84_14	'Wail'	Fats Navarro
82_7	'Epistrophy'	Thelonious Monk
77_4	'Hesitation'	Wynton Marsalis
27_5	'Doxy'	Dexter Gordon
20_3	'Blue in green'	Miles Davis
26_13	'On my own'	Al DiMeola

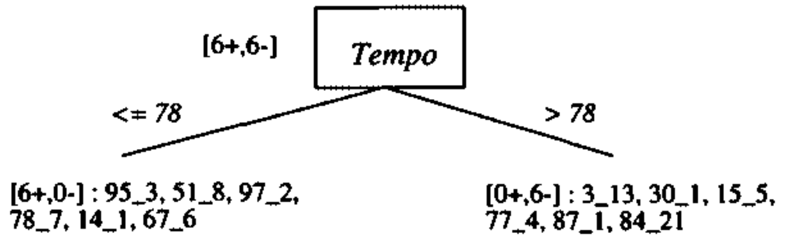


The modish character of the tunes is expressed by the requirement of being recorded after 1982. The preference for dance music is expressed by a funky rhythm. The preference for a particular setting size can not be interpreted.

4.3.5 Compilation 5

Compilation criterion: new age, meditative

Good tracks: (+)		
95_3	'Celeste'	Ralph Towner
51_8	'Etude'	Bill Frissel
97_2	'A remark you made'	Weather Report
78_7	'New born'	Lyle Mays
14_1	'Cote Jardin'	Philip Catherine
67_6	'Silence'	Keith Jarret
Bad tracks: (-)		
3_13	'In memory of Dick'	Chet Baker
30_1	'Upbeat 90's'	Tom Scott
15_5	'The dancer'	Billy Cobham
77_4	'Hesitation'	Wynton Marsalis
87_1	'Bird of paradise'	Charlie Parker
84_21	'The skunk'	Fats Navarro

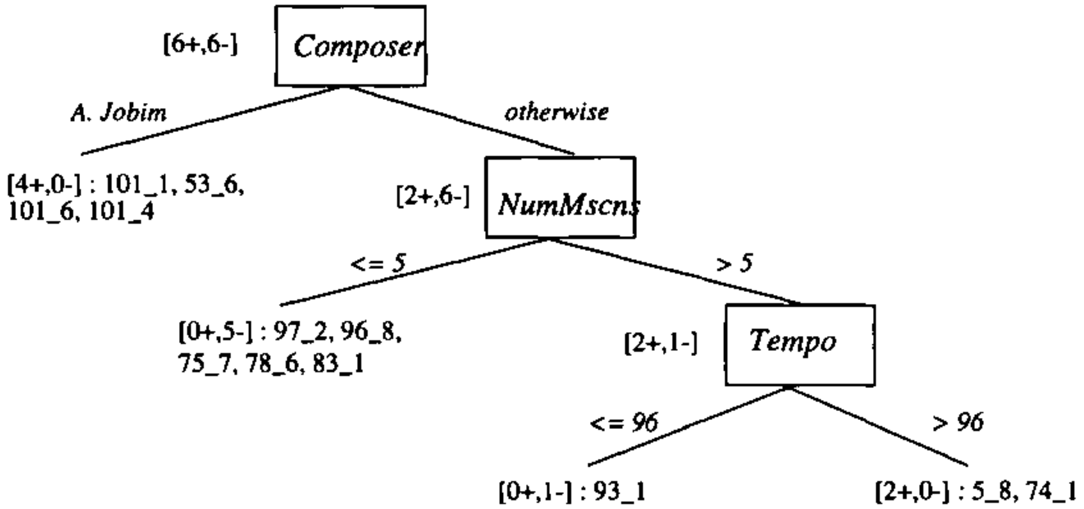


A new age style of music is not disclosed by the tree. For instance, 'A remark you made' is not known as a typical new age tune. The meditative character is expressed by a preference for a rather slow tempo.

4.3.6 Compilation 6

Compilation criterion: south-american / latin

Good tracks: (+)		
5_8	'Falsa baiana'	Batida
74_1	'Yatra-ta'	Tania Maria
101_1	'Garota de Ipanema - The Girl from Ipanema'	Stan Getz Joao Gilberto
53_6	'Corcovado'	Stan Getz
101_6	'S'o Danco Samba - Jazz Samba'	Antonio Carlos Jobim
101_4	'Morro Nao Tem Vez - Favela'	Stan Getz Luiz Bonf'a
Bad tracks: (-)		
93_1	'Sidewalk maneuvres'	Steps ahead
97_2	'A remark you made'	Weather Report
96_8	'Byrdlike'	Herbie Hancock
75_7	'Yesterdays'	Branford Marsalis
78_6	'Before you go'	Lyle Mays
83_1	'Line for lyons'	Gerry Mulligan

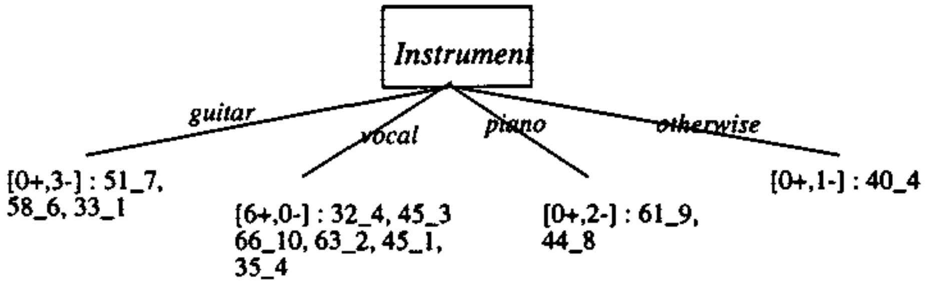


The brazilian Jobim is the most famous and notable composer of jazz latin songs. Unfortunately, the prerequisite for a particular setting size and tempo can not be conceived as relevant for latin grooves.

4.3.7 Compilation 7

Compilation criterion: vocals, ladies

Good tracks: (+)		
32_4	'Don't explain'	Gabrielle Goodman
45_3	'Stella by starlight'	Ella Fitzgerald
66_10	'The very thought of you'	Etta James
63_2	'Sophisticated Lady'	Billie Holiday
45_1	'You don't know what love is'	Dina Washington
35_4	'I can't help it'	Betty Carter
Bad tracks: (-)		
51_7	'Conception vessel'	Bill Frissel
61_9	'Maiden voyage'	Herbie Hancock
58_6	'Cats of Rio'	Dave Gruisin Lee Ritenour
44_8	'No more misunderstandings'	Stephen Scott
40_4	'Amsterdam after dark'	George Coleman
33_1	'Road song'	Jimmy Smith

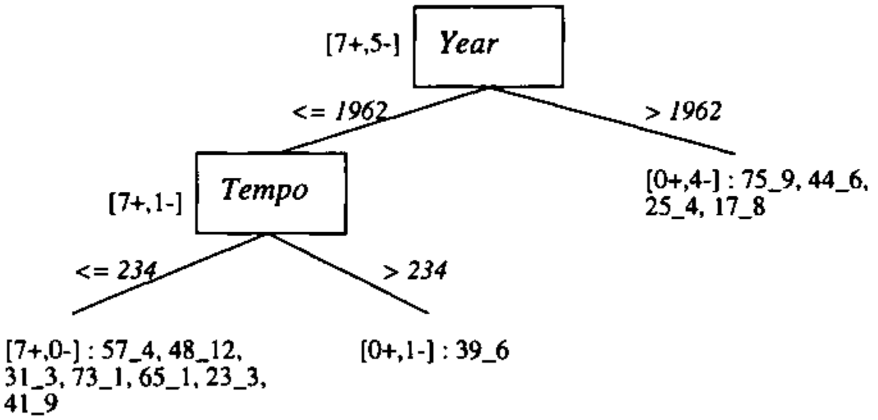


A preference for vocals is predominant is this compilation. However, the gender of musicians is not annotated in the music collection.

4.3.8 Compilation 8

Compilation criterion: swing jazz

Good tracks: (+)		
57_4	'Cheese cake'	Dexter Gordon
48_12	'A foggy day'	Roy Eldridge
31_3	'Kings road blues'	Quincy Jones
73_1	'I am in love'	Shelly Manne
65_1	'They can't take that away from me'	Milt Jackson
23_3	'In your own sweet way'	Miles Davis
41_9	'9-20 special'	Oscar Peterson
Bad tracks: (-)		
75_9	'Crepuscule with Nellie'	Branford Marsalis
44_6	'The call'	Randy Weston
39_6	'Ornithology'	Charlie Parker
25_4	'Nine over reggae'	Jack DeJohnette
17_8	'King cockroach'	Chick Corea

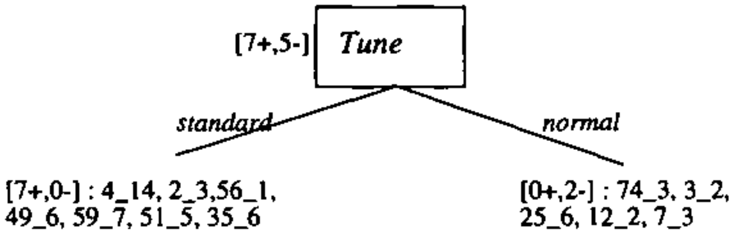


The years before 1963 characterize the jazz styles from the old days. However, a preference for a rather arbitrary tempo can not be interpreted.

4.3.9 Compilation 9

Compilation criterion: standards

Good tracks: (+)		
4_14	'You don't know what love is'	Chet Baker
2_3	'Autumn Leaves'	Gene Ammons Sonny Stitt
56_1	'Yesterdays'	Paul Gonsalves
49_6	'My romance'	Bill Evans
59_7	'Alone together'	Charlie Haden
61_5	''Round midnight'	Herbie Hancock
35_6	'In a sentimental mood'	John Coltrane
Bad tracks: (-)		
74_3	'Vem P'ra roda'	Tania Maria
25_6	'Indigo dreamscapes'	Jack DeJohnette
12_2	'Bottums up'	Ron Carter
7_3	'Scriabin'	Michael Brecker
3_2	'Mid-forte'	Chet Baker

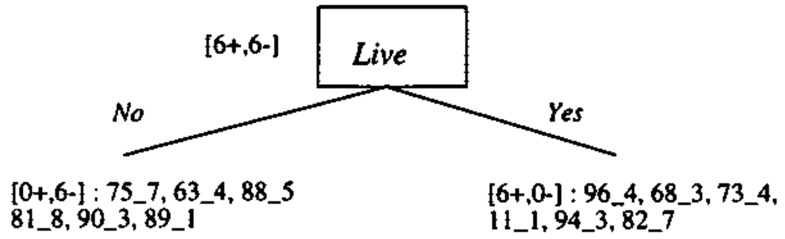


Tunes that are known as standards (in our definition, songs that are commonly played by jazz musicians) are filtered out.

4.3.10 Compilation 10

Compilation criterion: Recorded in front of a live audience

Good tracks: (+)		
96_4	'Lawra'	Herbie Hancock
68_3	'Falling in love with you'	Keith Jarret
73_4	'Whisper not'	Shelly Manne
11_1	'Summer in Central Park'	Ron Carter Richard Galliano
94_3	'Tenor madness'	Toots Thielemans
82_7	'Epistrophy'	Thelonious Monk
Bad tracks: (-)		
75_7	'Yesterdays'	Branford Marsalis
63_4	'I don't want to cry anymore'	Billie Holiday
88_5	'Alone in the morning'	Joshua Redman
81_8	'I mean you'	Thelonious Monk Gerry Mulligan
90_3	'John S.'	Sonny Rollins
89_1	'St. Thomas'	Sonny Rollins



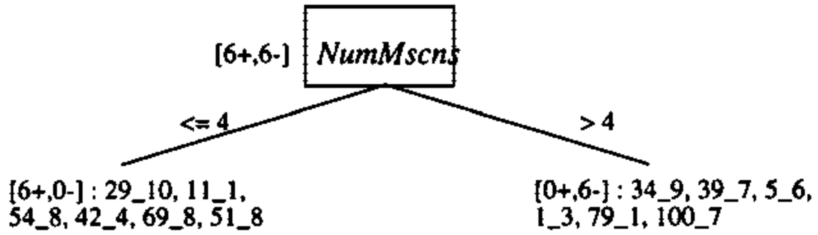
Tunes that are recorded in front of a live audience are easily traced back.

4.3.11 Compilation 11

Compilation criterion: Duo/solo

Good tracks: (+)		
29_10	'Cookin''	Kevin Eubanks
11_1	'Summer in Central Park'	Ron Carter Richard Galliano
54_8	'Conversation in G'	Hein van de Geyn
42_4	'Maiden voyage'	Chick Corea Herbie Hancock
69_8	'A nightingale sang in Berkeley square'	David Kikovski

51_8	'Etude'	Bill Frissel
Bad tracks: (-)		
34_9	'This won't work'	Quite sane
39_7	'Rocker'	Miles Davis
5_6	'Tudo bem'	Batida
1_3	'St. Louis blues'	Cannonball Adderley
79_1	'Have you heard'	Pat Metheny
100_7	'Little rootie tootie'	Joe Zawinul

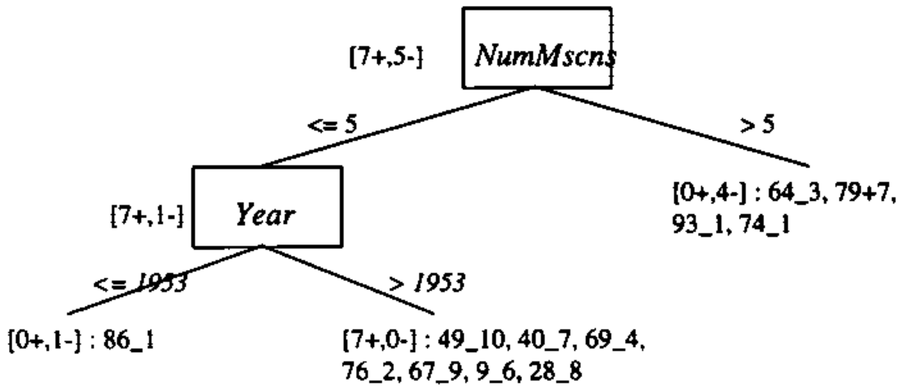


Although the number of musicians that participate in the recording is recognized as the main preferred factor, its exact number is not identified. This is due to a rather tricky definition of a setting size. Actually, we count the number of instruments played by the musicians. Some musicians play more than one instrument on a recording.

4.3.12 Compilation 12

Compilation criterion: Trio

Good tracks: (+)		
49_10	'Porgy - I love you Porgy'	Bill Evans
40_7	'Mean streets'	Tommy Flanagan
69_4	'Presage'	David Kikovski
76_2	'Emily'	Ellis Marsalis
67_9	'Blackberry winter'	Keith Jarret
9_6	'Little girl blue'	Ray Brown
28_8	'Angel eyes'	Dodo Moroni Ron Carter
Bad tracks: (-)		
86_1	'Confirmation'	Charlie Parker
64_3	'Gertrude's favourite'	The Houdini's
79_7	'Beat 70'	Pat Metheny
93_1	'Sidewalk maneuvres'	Steps ahead
74_1	'Yatra-ta'	Tania Maria

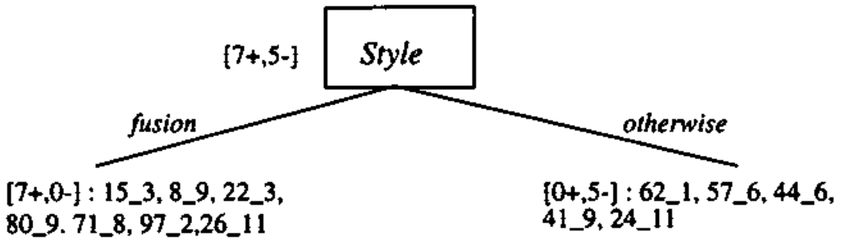


The same remark as in set 11 (see 4.3.11) applies for this compilation. In addition, the attribute 'year' can not be conceived as a compilation criterion.

4.3.13 Compilation 13

Compilation criterion: jazz rock, calm

Good tracks:(+)		
15_3	'Slow body poppin''	Billy Cobham
8_9	'And then she wept'	The Brecker Brother
22_3	'Portia'	Miles Davis
80_9	'Rising'	Mezzoforte
71_8	'Easy morning'	Koinonia
97_2	'A remark you made'	Weather Report
26_11	'Precious little you'	Al DiMeola
Bad tracks: (-)		
62_1	'Watermelon man'	Herbie Hancock
57_6	'Soy Califa'	Dexter Gordon
44_6	'The call'	Randy Weston
41_9	'9-20 special'	Oscar Peterson
24_11	'Gut bucket blues'	Joey DeFrancesco

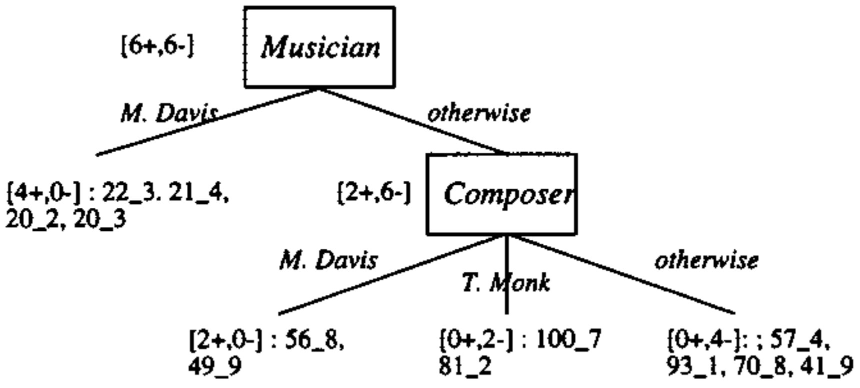


The music style ‘fusion/jazz rock’ is easily recognized as the main preference factor for this compilation. A calm character of the songs is not identified.

4.3.14 Compilation 14

Compilation criterion: Miles Davis

Good track: (+)		
56_8	‘Walkin"	Paul Gonsalves
49_9	‘Milestones"	Bill Evans
22_3	‘Portia’	Miles Davis
21_4	‘Miles’	Miles Davis
20_2	‘Freddie Freeloader’	Miles Davis
20_3	‘Blue in green’	Miles Davis
Bad tracks: (-)		
57_4	‘Cheese cake’	Dexter Gordon
100_7	‘Little rootie tootie’	Joe Zawinul
93_1	‘Sidewalk maneuvres’	Steps ahead
41_9	‘9-20 special’	Oscar Peterson
70_8	‘Ana Maria’	Kenny Kirkland



The criterion 'Miles Davis' is easily recognized as the main preference factor for this compilation.

4.3.15 Findings

We have demonstrated that the purpose of the ID3 application is well-served. A considerable number of compilation criteria could be traced back and interpreted. For some attributes that were chosen at lower tree levels of the trees, no direct interpretation regarding the compilation criterion could be made. Although some unwanted results can be attributed to the ID3 implementation, some are a result of the fact that not everyone has the same perception of a song tempo, and no exact borderline exists between categories of music styles, rhythms, and melodies. Categorisation clashes may exist between subjects and the person who completed the music collection database.

5 System evaluation

Experiments that evaluate the whole system can be conducted by real users as well as simulated users recognizing the fact that

- subjects are not available at any time,
- subjects cost a certain amount of resources,
- subjects show a great inter-individual variability,
- subjects express different interests to the contents of an unknown music collection,
- subjects behave in a way that can not always be anticipated from our assumptions on music preference, and
- subjects must be encouraged to participate in long-lasting tests.

5.1 Simulated users

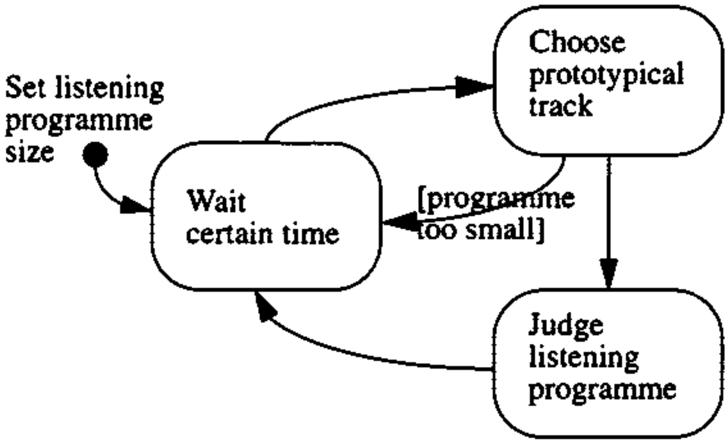

Set listening programme size
Choose prototypical track
Wait certain time
[programme too small]
Judge listening programme

Figure 5.1 State diagram of simulated user.

Setting a simulated user behind the terminal making human-like preference decisions requires a music preference model that complies with our assumptions on music preference **[Pauws2]**.

We have implemented a simulated user following a simple state diagram as shown in Figure 5.1. Iteratively, it takes three actions. Before a trial is actually started, the listening programme size is stated at 15. First, the simulated user waits a *certain* amount of time required for the system to settle itself. Second, it chooses a particular track that is prototypical for its music preference. As a response, the system offers a listening programme consisting of 15 tracks. If the system cannot offer 15 tracks for some reason, the user will simply retracts and 'will come back later'. Third, it judges all tracks in the offered listening programme. This cycle of waiting-choosing-judging is repeated 15 times, i.e. the simulated user judges 15 listening sessions. To smooth the statistical variability, each experiment consists of 10 trials.

Wait a certain amount of time - We have defined the moment in which a track agent stops or starts a following behaviour (i.e. a change in the cluster organisations) as a time notion. This notion is used as time step in which the unfolding of clusters over time is monitored. In some preliminary tests, we have investigated how much 'average' system configuration time is required between listening sessions to make up new listening programmes. First, the system needs some initialisation to let all track agents be participated in a particular follow behaviour. This initialisation time is fixed at 750 time steps. Between requests for listening programmes, the system requires time to adapt and reconfigure itself to the new induced preferences. This re-configuration time is fixed at 250 time steps.

Choose a prototypical track - Several strategies can be used how to determine a preferred 'prototypical' track. We have adopted two strategies:

strategy 1) Take each time one and the same prototypical track; this track is randomly chosen at the very beginning of the trial.

strategy 2) Take each time a prototypical track at random from the set of preferred tracks.

The first strategy is important to investigate the clustering around a single prototypical track but it is less useful if a user preference is formulated as a disjunction of aspects. Imagine one likes a track if musician *x or* musician *y* plays along. It may be difficult to find such a prototypical track in which both musicians are really present. Therefore, strategy 2 should address these disjunctive preference aspects by randomly choosing tracks in which at least one of the musicians play along.

Judge the listening programme - The judgement strategy determines whether a given track is suitable in a particular listening programme. It stresses the context-dependency of a judgement decision. The context is determined by the music preference of the user. This reflects real-life decisions: a music song can be highly preferred if it is collected in a particular volume, whereas it may be irrelevant in combination with others. A tree representation is imposed to the judgement process. Along each branch, an attribute and its value is associated that conclusively states the user's music preference[1]. For a given track, one starts at the root and selects the branch that has an attribute-value that matches with some of the tracks' attribute-values. Tracks that successfully pass these tests are indicated as preferred. Other elaborated representations are also possible by introducing some inherent

[1] One could also include attributes and values that the simulated user dislikes.

variability (or inconsistency) in the preferential judgement **[Pauws3]**, but this makes the task for the system more complicated.

The metrics precision, coverage, and variation are monitored. Former documents **[Pauws2]** have introduced some clumsy notation. Therefore, we introduce the following sets:

S_p : set of all preferred (in relation with the prototypical track) tracks in the music collection This set can be easily obtained by using the judgement strategy and preference tree model. However in the case of a real user, it is infeasible to come across such a set without letting the user judge each track.

$O_p(t)$: set of preferred tracks as offered in session at time *t*.

$O_{np}(t)$: set of non-preferred tracks as offered in session at time *t*.

The ordinary set operations are defined: $S_1 \cup S_2$ set-union, $S_1 \cap S_2$ set-intersection, $S_1 - S_2$ set-difference. In addition, we define the operator *#(S)* that returns the number of elements in a set *S*.

Now, *precision* is a metric varying from 0 to 1 that gauges the selection accuracy of preferred tracks,

(EQ 1)

$$\text{precision}(t) = \frac{\#(O_p(t))}{\#(O_p(t) \cup O_{np}(t))}$$

Coverage keeps track of how many distinctive preferred tracks are offered over all listening sessions,

(EQ 2)

$$\text{coverage}\ (t) = \frac{\#\left(\bigcup_{k=1}^{t} O_p(k)\right)}{\#(S_p)}$$

As already mentioned, the set S_p can be obtained from the preference tree model.

The coverage metric accurately gauges at what speed a considerable portion of the preferred material is offered but does not give us good insight in the variation among two adjacent listening programmes. Therefore, we have devised an extra metric named *variation* that measures the preferred tracks that are offered in session *t* and that were not already offered in session *t-1*,

(EQ 3)

$$\text{variation}(t) = \frac{\#(O_p(t) - O_p(t-1))}{\#(O_p(t) \cup O_{np}(t))} \qquad ,(t > 1)$$

It should be stressed that the variation metric only considers the preferred tracks. We have not defined a metric that works on non-preferred tracks. Consequently, variation can only reach its maximum of 1 if the precision is maximal. The variation metric can be a valuable indicator how much dynamics (or spirit) is still left in the system. In other words, if the variation tends to decrease, the system will reach some equilibrium state in which no more surprising effects must be expected.

The PATS functionality should outperform a random track selection strategy. An indication of minimum performance level is given by exploiting the hypergeometric distribution **[Pauws2]**. Let's consider $N = 300$ tracks in the collection from which $N_p (= \#(S_p))$ tracks are preferred. We take at random and without replacement n tracks from this collection. The stochastic variable X is the number of preferred tracks in this selection. X is hypergeometrically distributed. More precisely,

(EQ 4)

$$P_X(X = k) = \frac{\binom{N_p}{k}\binom{N - N_p}{n - k}}{\binom{N}{n}}$$

for $\max(0, n - N + N_p) \le k \le \min(N_p, n)$.

The expected number of preferred EX tracks and variance $\sigma^2(X)$ are expressed as

(EQ 5)

$$EX = \frac{nN_p}{N} \quad \text{and} \quad \sigma^2(X) = \frac{n(N-n)N_p(N-N_p)}{N^2(N-1)}$$

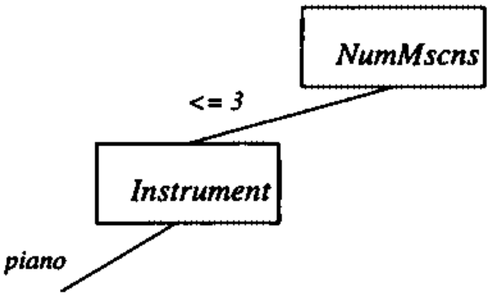
Let's consider in more detail the assessment of a random functionality in which listening programmes of 15 tracks are offered. The user is allowed to choose one track that match his/her preference. Subsequently, the system comes up with $n = 14$ randomly chosen tracks from the remaining $N = 299$ tracks. Thus, at least one track matches the preference tree model. An expected number of preferred tracks EX in a random session of 14 tracks can be calculated by substituting (EQ 5). Consequently, the expected precision for a random functionality equals

$(EX + 1) / (n + 1)$. Although we only evaluate the system by observing the trends of the performance indicators, this value is a precise lower limit for the metric precision.

We have evaluated at what rate the system adapts to one particular preference model. In addition, we have combined three preference models in an interleaving mechanism to assess the interaction between different preference models. We have experimented with 6 distinct preference tree models.

5.1.1 Preference 1: lover of piano jazz with a small accompaniment

$N_p = 25$ tracks out of 300 tracks (8.33%) are present in the music collection that satisfy this description. The 'expected' precision for a random approach equals 0.142. Prototypical tracks are recordings from 'Keith Jarret' or 'Bill Evans'.



Results

As shown in Figure 5.2 and Figure 5.3, the PATS functionality demonstrates a considerable higher precision than a simple random approach. Looking at the precision curve, a learning effect is observed and reaches 80%. It is questionable whether the adaptation time should be shortened. It may be effected by longer configuration time between the sessions. With respect to the variation of the offered sessions, we observe that there is considerable more variation when applying strategy 2. This also results into a steeper coverage curve meaning that more distinctive preferred tracks are offered in less time.

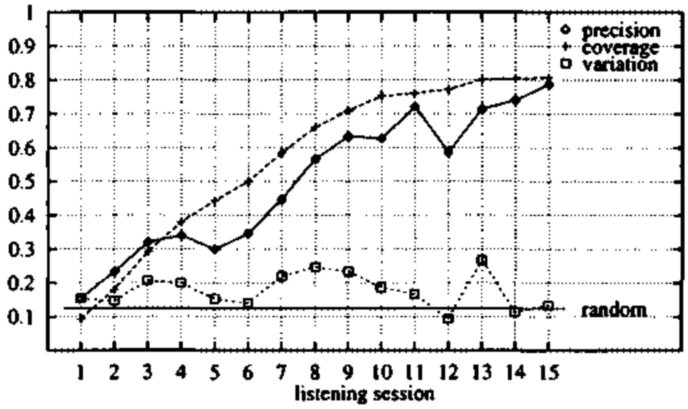


Figure 5.2 Performance statistics of simulated preference 1 (lover of piano jazz with a small accompaniment) involved in 15 listening sessions. During each trial, one and the same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

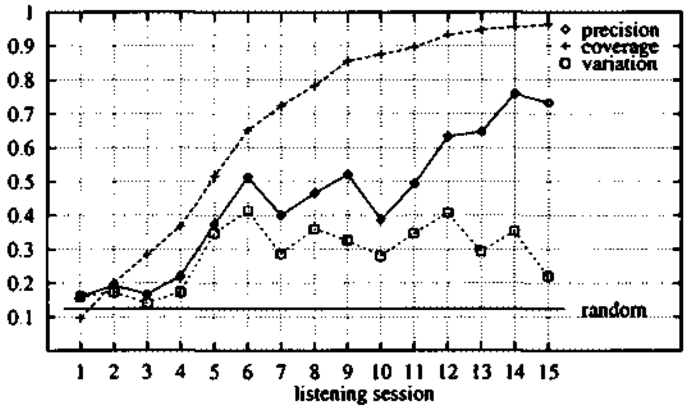
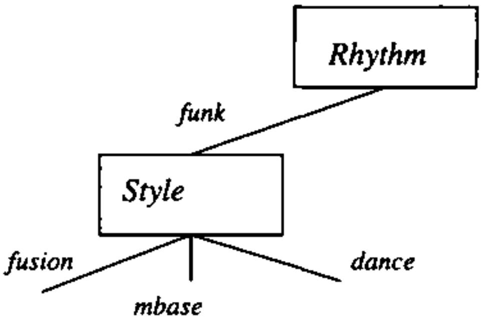


Figure 5.3 Performance statistics of simulated preference 1 (lover of piano jazz with a small accompaniment) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.2 Preference 2: lover of funk (dance-oriented) modern jazz

N_p = 35 tracks out of 300 tracks (11.67%) are present in the music collection that satisfy this description. The 'expected' precision of a random approach equals 0.173. Prototypical tracks are some recording from 'The Brecker Brothers' or 'Mezzoforte'.



Results

As shown in Figure 5.4 and Figure 5.5, it is remarkable that the PATS functionality has no difficulty in adapting to this preference profile. Whereas the precision is optimal, the variation in strategy 1 decreases over time due to some 'overlearning': the cluster with the prototypical track is too static. This phenomena is not observed if strategy 2 is applied. There is probably more than one cluster representing the preference profile. This is also reflected in a higher coverage.

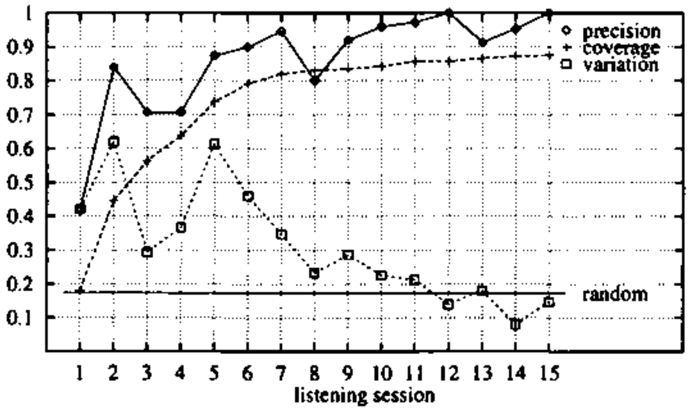


Figure 5.4 Performance statistics of simulated preference 2 (lover of funk dance-oriented modern jazz) involved in 15 listening sessions. During each trial, one and the

same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

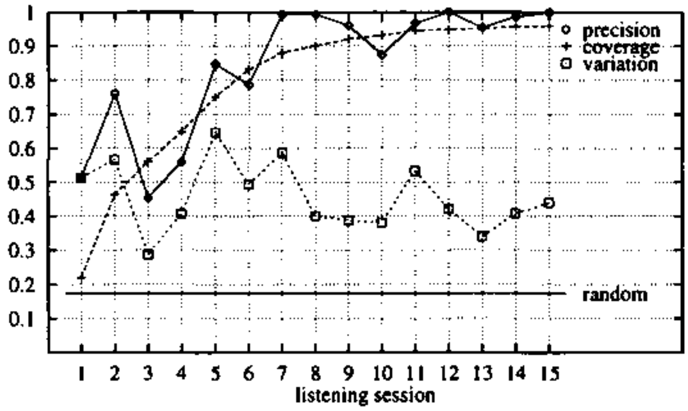
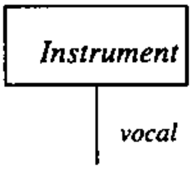


Figure 5.5 Performance statistics of simulated preference 2 (lover of funk dance-oriented modern jazz) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.3 Preference 3: lover of vocal jazz

N_p = 29 tracks out of 300 tracks (9.67%) are present in the music collection that satisfy this description. The 'expected' precision of random approach equals 0.154. Prototypical tracks are recordings from 'Billie Holiday' or 'Chet Baker'.



Results

Although we feel that this preference profile should be relatively easy to adapt to, it is apparent in Figure 5.6 and Figure 5.7 that the attribute-value 'vocals' is probably obscured by other attribute-values and consequently hard to induce. The variation in Figure 5.6 tends to decrease indicating that no considerable progress can be made although only 80% of all vocal jazz is offered. This is not true in Figure 5.7: almost all vocal jazz is offered at least once.

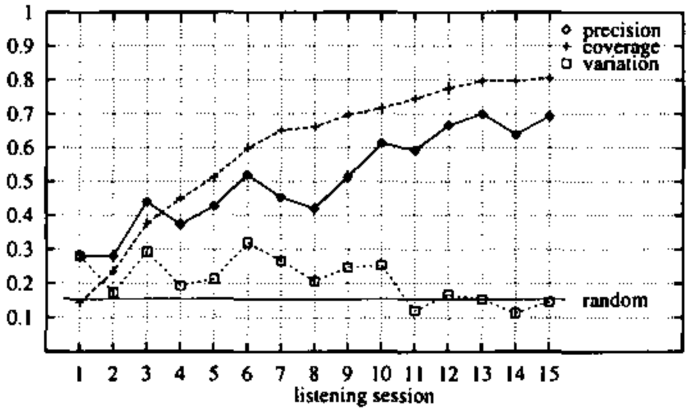


Figure 5.6 Performance statistics of simulated preference 3 (lover of vocal jazz) involved in 15 listening sessions. For each trial, one and the same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

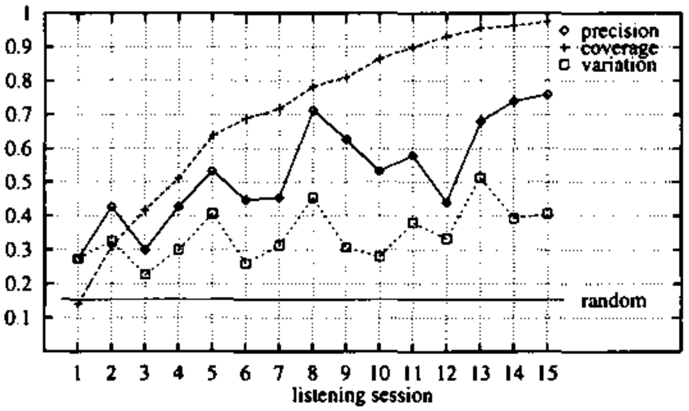
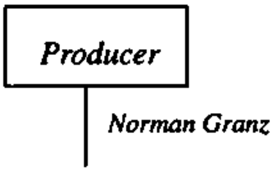


Figure 5.7 Performance statistics of simulated preference 3 (lover of vocal jazz) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.4 Preference 4: lover of Norman Granz jazz

N_p = 22 tracks out of 300 tracks (7.33%) are present in the music collection that satisfy this description. The 'expected' precision of a random approach equal 0.132. Prototypical tracks are recordings from the Verve label.



Results

As shown in Figure 5.8, the average precision is over 80% at the end of a trial, although the variation is negligible. On the other hand, the coverage is maximal and the precision is optimal while the variation between two sessions is about 30% for the last two sessions in Figure 5.9.

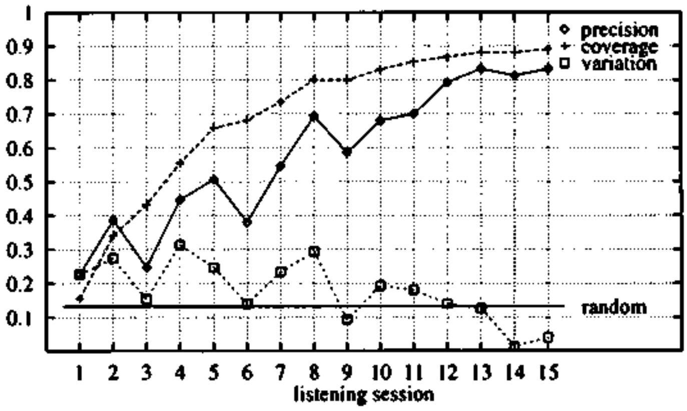


Figure 5.8 Performance statistics of simulated preference 4 (lover of Norman Granz jazz) involved in 15 listening sessions. For each trial, one and the same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

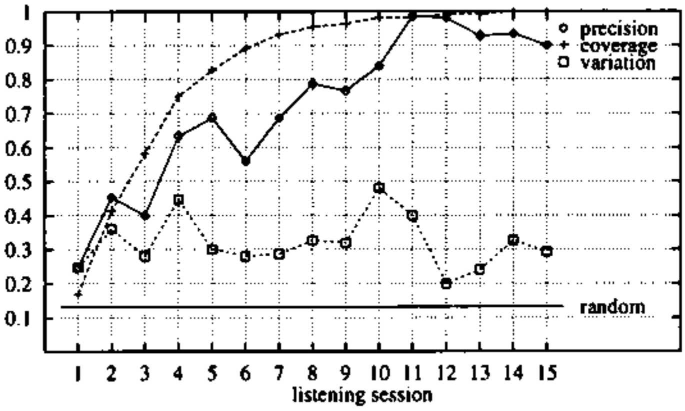
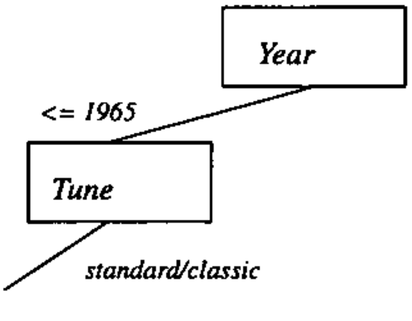


Figure 5.9 Performance statistics of simulated preference 4 (lover of Norman Granz jazz) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.5 Preference 5: lover of old-days classical jazz tunes

N_p = 45 tracks out of 300 tracks (15%) are present in the music collection that satisfy this description. The 'expected' precision of random approach equals 0.204. Prototypical tracks are recordings from songs of the composer duos 'Rodgers-Hart' or 'Gershwin-Gershwin'.



Results

Although the precision is near optimal when applying strategy 1 as shown in Figure 5.10, only half of the preferred material is offered: the other half is missing (for some obscure reason).

This is in contrast with strategy 2 (as shown in Figure 5.11), in which almost all preferred material is offered at least once with the same precision characteristics.

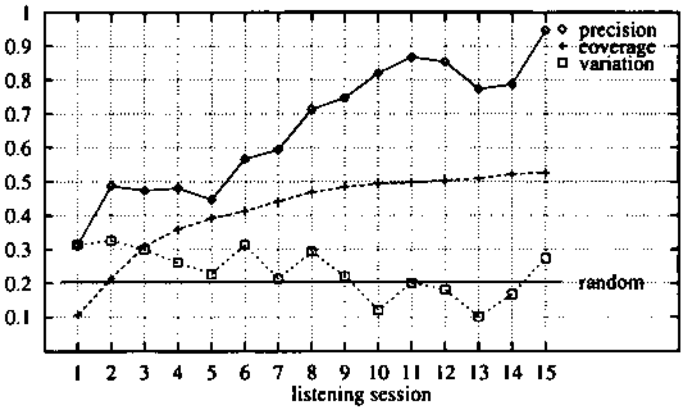


Figure 5.10Performance statistics of simulated preference 5 (lover of old-days classical jazz tunes) involved in 15 listening sessions. For each trial, one and the same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

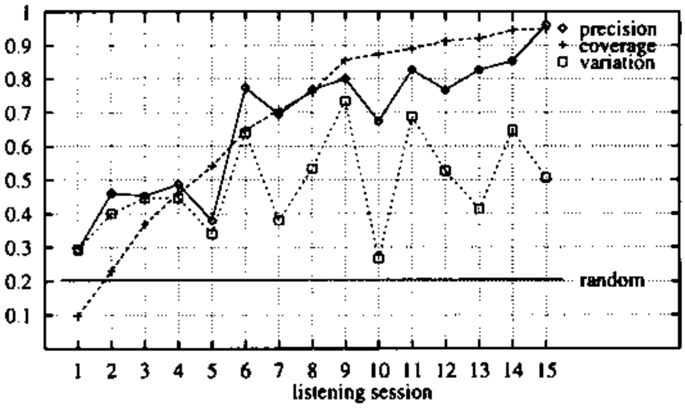
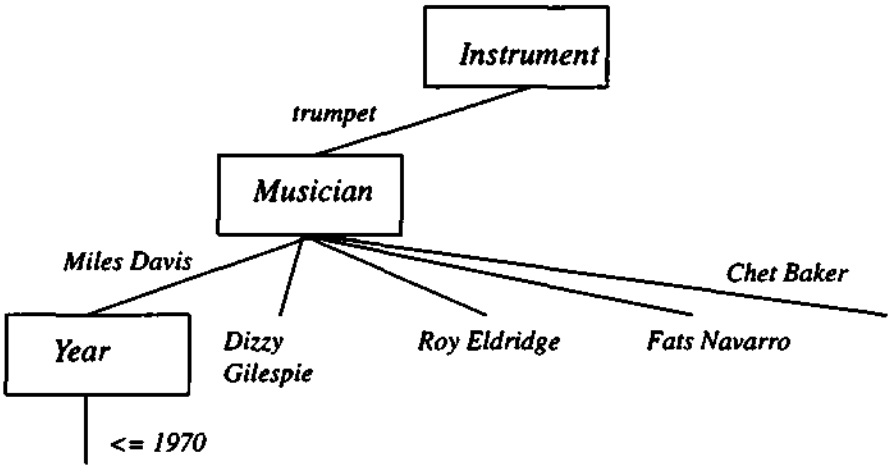


Figure 5.11 Performance statistics of simulated user 5 (lover of old-days classical jazz tunes) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.6 Preference 6: lover of notable jazz trumpet players

N_p = 33 tracks out of 300 tracks (11%) are present in the music collection that satisfy this description. The 'expected' precision of a random approach equals 0.167. Prototypical tracks are self-evident.



Results

With respect to the former preference profiles, the PATS functionality shows some difficulty in adapting to this preference. This can be clearly observed in Figure 5.12 and Figure 5.13 in which the average precision does not exceed 70%. This can be explained by recognizing the fact that the preference profile is a disjunction of distinct objects; the five trumpet players are distinct and cannot be clustered solely by name. Although there is a common concept between the musicians (they all play the trumpet), this is not always found out by the system primarily due to the lack of some kind of inference mechanism. As a matter of fact, the combination of the used track similarity measure and the adjustment of weight factors by interpreting the decision trees rather make things worse; these particular musicians get higher weights causing repulsion of their music tracks. As a result, the system does not succeed to join the musicians in one and the same listening programme, while though they all play the same instrument. An overly simple patch for this particular problem is the introduction of joint recordings in the music collection. A more concrete, long-standing solution is to refine the similarity measure or introduce some simple inference rules.

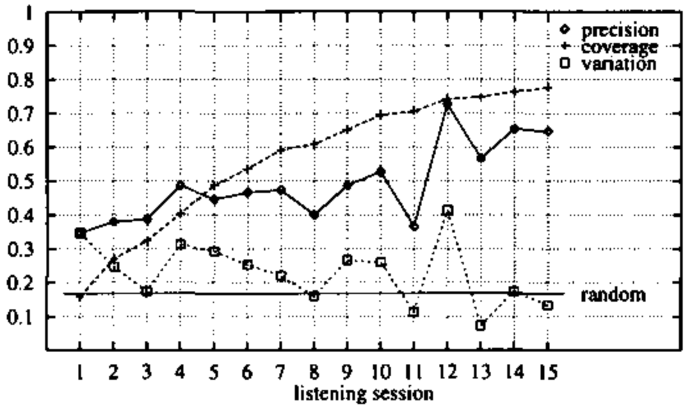


Figure 5.12Performance statistics of simulated preference 6 (lover of notable jazz trumpet players) involved in 15 listening sessions. For each trial, one and the same prototypical track is used (strategy 1). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

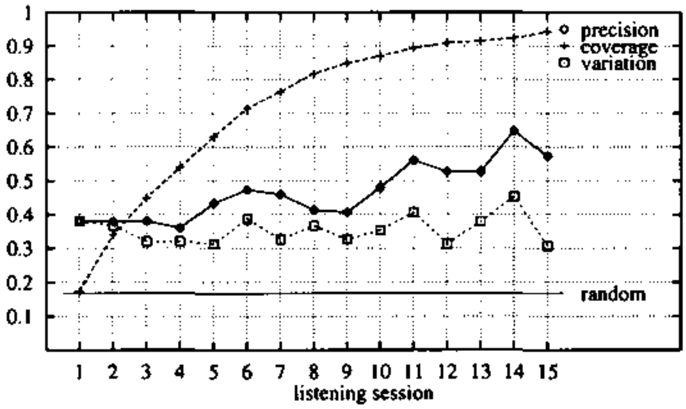


Figure 5.13Performance statistics of simulated preference 6 (lover of notable jazz trumpet players) involved in 15 listening sessions. For each session, a prototypical track is randomly chosen (strategy 2). Arithmetic means of 10 trials are shown. As a reference, the 'expected' precision of a random approach is shown.

5.1.7 Preferences 1, 2, and 4 combined

We are interested whether the system is capable in managing more than one preference at the same time. Therefore, we have combined the preferences 1, 2, and 4. One at a time, a preference is activated in a 'round-robin' fashion. Each preference is characterized by the same prototypical track as dictated by strategy 1 (thus resulting in three distinct prototypical tracks). In this particular preference combination, we have kept the interaction between the three preferences small: only the sets of preferred tracks associated with preference 1 and 4 have one track in common, all other intersections of preference sets are empty. The simulated user judges 30 listening sessions in each trial. Each experiment consists of 10 trials. Arithmetic means of 10 trials are shown.

Results

Only the precision of this combined preference profile is shown in Figure 5.14. It is evident that the system is, to some extent, capable in managing more than one preference if there is no or little interaction between these preferences.

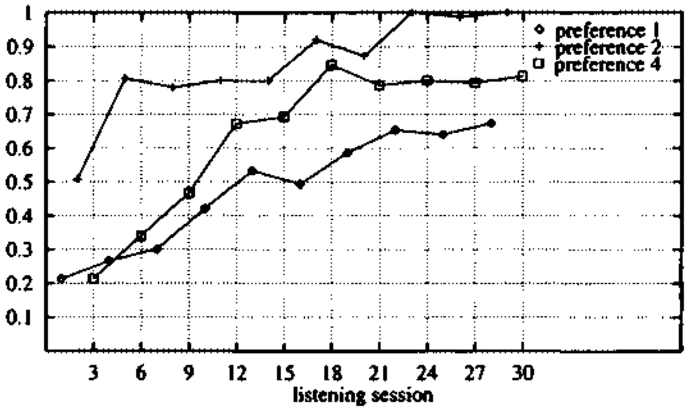


Figure 5.14. Precision statistics of a combination of three simulated preferences 1, 2, and 4 involved in 30 listening sessions. By means of an interleaving mechanism, each preference is activated. For each trial, each preference is characterised by one and the same prototypical track (strategy 1). Arithmetic means of 10 trials are shown.

5.1.8 Preferences 3, 4, and 5 combined

The last experiment comprises the combination of preference profiles 3, 4, and 5. The distinction with the former is that there is now a considerable interaction between the preferences. More specifically, denote the set of preferred tracks under preference i as $S_{\text{pref } i}$ then

$$\#(S_{\text{pref } 3} \cap S_{\text{pref } 4}) = 2,$$
$$\#(S_{\text{pref } 3} \cap S_{\text{pref } 5}) = 8,$$

$$\#(S_{\text{pref}\,4} \cap S_{\text{pref}\,5}) = 9,$$
$$\#(S_{\text{pref}\,3} \cap S_{\text{pref}\,4} \cap S_{\text{pref}\,5}) = 2.$$

Results

As shown in Figure 5.15, this experiment demonstrates that the current system is less adequate in preference adaptation if there is a considerable interaction between preferences. It may be concluded that similarity among tracks is not transitive within distinct preferences.

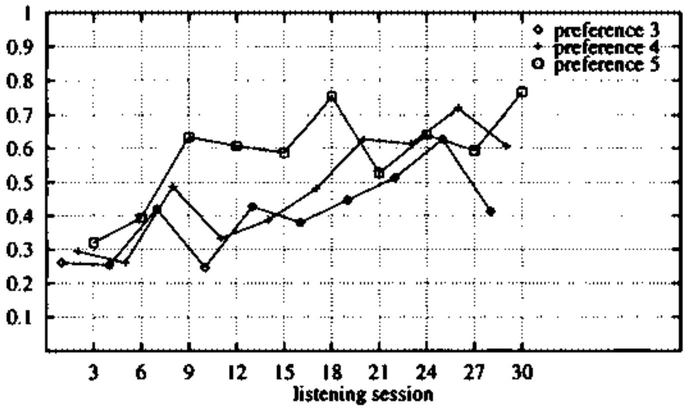


Figure 5.15. Precision statistics of a combination of three simulated preferences 3, 4, and 5 involved in 30 listening sessions. By means of an interleaving mechanism, each preference is activated. For each trial, each preference is characterised by one and the same prototypical track (strategy 1). Arithmetic means of 10 trials are shown.

5.1.9 Findings

Experiments with simulated users and preferences have shown that the system is capable to adapt to these precisely formulated preferences. Though, it finds more difficulties when the preference profile consists of a disjunction of attribute-values (see Section 5.1.6). If there is no common property between these disjunct aspects (or the system does not succeed in finding this), the weighted similarity measure (combined with the induction mechanism) between tracks does not associate distinctive attribute values but rather further discriminate between these distinctions.

The strategy how one chooses a prototypical track to indicate one's preference plays a dominant role in the degree of variation. Especially if one does not stick to the same prototypical track, more average variation and higher average coverage over time is obtained. This is plausible if one consider that one may extract another cluster if one selects different prototypical tracks. One could hence compare all present clusters to find the most appropriate one.

Managing more than one preference is only successful when there is no considerable interaction between the preferences. In particular, tracks that are preferred by more than one preference profile can fulfill a bridge between these preference profiles. In fact, tracks that enter a cluster bring along all its followers from another cluster expressing another preference. Due to the non-transitivity of the local weighted similarity measure between tracks, the global preference criterion within cluster may be disturbed. In other words, if we know that *A* is quite similar to *B* within some preference, and *B* is quite similar to *C* within some other preference, we cannot imply that *A* is not very dissimilar to *C* within one of the former preferences.

5.2 Real user tests

Real user tests should demonstrate the real profit of the PATS feature. If the system adapts to the simulated preferences, we like to know whether the system adapts to real users and whether the users are aware of this fact. Besides, it is questionable how much system adaptation is desired by the user. It is commonly known that real user tests take a considerable amount of time. We cannot always estimate in advance its throughput or required effort accurately and, hence, it caused a considerable backlog in this project. Some causes of this 'being-behind-schedule' are identified:

- Although some performance aspects are quantifiable, it is mainly a qualitative matter in which users should set forth why a particular listening programme might be interesting by means of a semi-structured interview. In addition, we have the opinion that this first encounter between system and user is a rather diagnostic one than a comparative one (with respect to a control group or other system). Laying our hands on possible wrong assumptions and false hopes from our side can be better tackled by a qualitative than a quantitative approach. However, measurement criteria, conditions, and assumptions should be known at the very beginning of the first pilot experiment. At this lab, little experience is currently present in setting up these kind of experiments.
- The music collection is highly personalized implying users have to acquaint themselves with the contents or should have a considerable knowledge on jazz recordings.
- The experiments are long-lasting because it takes time for the system to adapt to a user's music preference.

It was felt that an accurate description of the problem statement, methods, and analyses were requirement before a reliable experiment can be conducted. Therefore, this work is re-defined and boarded out as an apprenticeship and graduation **[Ober]**.

6 Discussion

In this report, we have shown that the PATS functionality adapts to a certain kind of music preference as we have defined in previous documents. However, this is only validated in simulations and the final evaluation should be conducted with real users. Humans behave quite different and make unpredictable preference decisions. In fact, we have the invalidated assumption that humans make preference decisions based on relative importance of attribute-values. It is however also true, that humans prefer tracks by gradually eliminating less attractive alternatives by sequential evaluation of the attribute-values **[Tversky1]**. If some attributes do not meet some minimal criterion, their corresponding tracks are eliminated. In other words, preference decisions have more than one angle. Compare in this respect the two statements "I like this song because it has the right tempo", and "I dislike this song because it is played too fast". Some fundamental insight with respect to preference decisions is desirable. In addition, if the music listener is aware of some sort of adaptation phenomena in an appliance such as this one, it is questionable how much adaptation is desired. An added value with respect to present-day CD player functionalities should be made clear. For instance, some may even say that it remains rather thrilling to browse through your racks and make the music selection yourself.

The simulations made clear that the system finds troubles in grouping tracks that are not (or hard to make) compatible based on their common attributes (see Section 5.1.6) and in managing more than two interacting preferences (see Section 5.1.7 and 5.1.8). This is due to our actual implementation of the similarity measure between tracks and the way we use the output results of the induction algorithm. The similarity is defined as some linear combination of common attribute-values. Hence, it is not capable in associating tracks that have nothing much in common, but are felt to belong together. For example, it is not immediately evident how the system can group tracks from 'Dizzy Gillespie' and 'Chet Baker' although they both play the trumpet but never played together. Only resorting to attribute-values that have tracks of both musicians in common, such as 'trumpet' in this case, is too brittle. Even so much so, the similarity measure and the way induced results are incorporated rather separates them. The phenomenon of poor listening programmes with conflicting preferences is due to the non-transitivity of the local similarity measure within the context of different preferences. This is in accordance with the violation of the triangle inequality in mathematics: if A is quite similar to B, and B is quite similar to C, the A can not be very dissimilar to C. However, it appears that this non-transitivity is constraining in our approach **[Tversky2]**. It might be worthy to investigate another family of similarity measures such as the *contrast model* and the *ratio model* that both incorporate contributions of common attribute-values as well as distinctive attribute-values. Another possibility is to incorporate simple inference rules and some domain knowledge.

By definition, our similarity measure is an asymmetric relation: if A is quite similar to B, it is not necessarily so that B is quite similar to A. It seems to be a sensible decision to have this sort of directional property in which there is an explicit subject and referent (or prototype) in the similarity calculation. In daily life too, we say things like "this kid plays like Miles Davis" whereas the converse statement might not be true.

Immediate small implementation actions that might contribute to some improvement were already discussed in Section 4.2.4. The quantitative attributes require a coarser ordinal categorisation and a hierarchical order for instruments (or some domain knowledge in general) might be profitable.

If we could eliminate time boundaries and limited capacities and resources, we also could immediate validate the generic power of the PATS functionality on another carrier (e.g. another music domain or even multi-media domain). Another interesting research issue is to see for design concepts that resolve possible conflicts of PATS with other functionalities such as 'random/shuffle play' and 'favourite track selection'.

References

[Eggen 1995]

Turn on the Base, Accessing large amounts of information in multi-media applications for home entertainment environments,
Eggen J.H., PRL Redhill Technical Note, 1995.

[Ober 1996]

Evaluatie van de PATS functionaliteit,
Ober D., forth-coming (in dutch), 1996.

[Pauws1 1995]

Turn on the Base, Project Contract,
Pauws S.C., IPO-report no. 1032, 1995.

[Pauws2 1995]

Turn on the Base, Project Analysis and Design,
Pauws S.C., IPO-report no. 1051, 1995.

[Pauws3 1995]

Turn on the Base, Project Realisation,
Pauws S.C., IPO-report no. 1074, 1995.

[Tversky1 1972]

Elimination by Aspects: A Theory of Choice,
Tversky A., Psychological Review, VOL.79 No.4, 1972, pp 281-299.

[Tversky2 1977]

Features of Similarity,
Tversky A., Psychological Review, VOL.84 No.4, 1977, pp 327-352.