CONTRIBUTING TO NEW MUSICOLOGICAL THEORIES WITH COMPUTATIONAL METHODS: THE CASE OF CENTONIZATION IN ARAB-ANDALUSIAN MUSIC

Thomas Nuttall **Miguel García Casado** **Víctor Núñez Tarifa**
Rafael Caro Repetto **Xavier Serra**
Music Technology Group, Universitat Pompeu Fabra, Spain
thomas.nuttall01@estudiant.upf.edu, miguel.garciac@upf.edu

ABSTRACT

Arab-Andalusian music was formed in the medieval Islamic territories of the Iberian Peninsula, drawing on local traditions and assuming Arabic influences. The expert performer and researcher of the Moroccan tradition of this music, Amin Chaachoo, is developing a theory whose last formulation was recently published in *La Musique Hispano-Arabe, al-Ala* (2016), which argues that centonization, a melodic composition technique used in Gregorian chant, was also utilized for the creation of this repertoire.

In this paper we aim to contribute to Chaachoo's theory by means of tf-idf analysis. A high-order n-gram model is applied to a corpus of 149 prescriptive transcriptions of heterophonic recordings, representing each as an unordered multiset of patterns. Computing the tf-idf statistic of each pattern in this corpus provides a means by which we can rank and compare motivic content across *nawabāt*, distinct musical forms of the tradition. For each *nawba*, an empirical comparison is made between patterns identified as significant via our approach and those proposed by Chaachoo. Ultimately we observe considerable agreement between the two pattern sets and go further in proposing new, unique and as yet undocumented patterns that occur at least as frequently and with at least as much importance as those in Chaachoo's proposals.

1. INTRODUCTION

The rich culture developed in the medieval Islamic territories of the Iberian Peninsula known as Al-Andalus gave birth to a refined musical and literary tradition that combines local musical practices with Middle Eastern Arabic poetry and sensibilities. The core of this tradition is the singing of *ṣanā‘i‘* (plural of *ṣan‘a*) or poems either by a choir accompanied by an instrumental ensemble or by a soloist. These *ṣanā‘i‘* are performed in suites known as *nawabāt* (plural of *nawba*), which also include orchestral pieces and both instrumental and vocal solo improvisations [6, 13]. The migration of Andalusian population to North Africa brought this tradition to this region, were it survived to this date after the disappearance of Al-Andalus in the 15th century. Nowadays, it is considered the classical musical repertoire in countries such as Morocco, Algeria, and Tunisia, in each of which it developed local characteristics, and is commonly known (among other names [22]) as Arab-Andalusian music. In this paper, we focus on the Moroccan repertoire of this tradition, which is known as *al-Āla* [6].

1.1 Music Theory and Centonization

Nawba is the essential form of Arab-Andalusian music. All the *ṣanā‘i‘* and other pieces in one *nawba* are composed in one single mode, known in this tradition as *ṭāb‘* (plural *ṭūbu‘*). In the specific case of the Moroccan *al-Āla* repertoire, pieces from certain *nawabāt* were lost during the process of oral transmission, so that the surviving ones where attached to other *nawabāt* according to modal similarity. In the 18th century, the scholar al-Haiek fixed the number of *nawabāt* in the *al-Āla* tradition to eleven (Table 1) [6].

The scholar and expert performer of *al-Āla* Amin Chaachoo is researching and developing a theoretical framework for this tradition. Chaachoo argues that regarding its musical aspect, Arab-Andalusian music heavily draws on local Iberian practices, and especially on plainchant in terms of compositional principles. A main argument for this proposal is the nature of Arab-Andalusian *ṭūbu‘*, which lack the microtonalities and nuances of Arabic *maqam*. Chaachoo, in his publication *La música andalusí al-Ála* [5] characterizes each *ṭāb‘* with a particular ascending and descending scale, a fundamental degree similar to the *finalis* of Gregorian modes, and one or two dominant degrees. In his more recent *La Musique Hispano-Arabe, al-Ala* [6], he also proposes the concept of a "persistent degree," inspired in the reciting tone from plainchant.

One of the most original proposals by Chaachoo is the use of centonization as the basic technique for melodic creation in Arab-Andalusian music. Centonization, from latin *cento* meaning patchwork, is defined by Paolo Ferretti as melodic composition by synthesis of pre-existing musical units known as *centos* [11], and is generally associated to melodic creation in Gregorian chant [1, 7]. Chaachoo argues that the melodic material in Arab-Andalusian music is also created by combination of *centos*, thus strengthening the connection between this tradition and Iberian local practices. The author associates these melodic units or *centos* to specific *ṭūbu‘* (See Fig.1), so that one particular *ṭāb‘* is characterized by a set of *centos*. To the best of our knowledge, Chaachoo's is the first attempt to explain melodic creation of Arab-Andalusian music through the concept of centonization.

Figure 1: Score example with the characteristic *centos* of the piece *Btayhi Rasd Dayl* of Nawba 4

1.2 Motivation

The use of centonization as composition technique in Arab-Andalusian music is still being developed by Chaachoo. Already published in two recent works [5, 6], the list of *centos* per *ṭāb‘* slightly varies from its first formulation to the second, as the theory is being consolidated. Chaachoo draws on his lifetime experience as performer and instructor of this music, as well as in his analytic work as a musicologist, for the definition of each *ṭāb‘* list of characteristic *centos* [1]. Therefore, our main goal is to contribute to the development of Chaachoo's theoretical work with findings from computational analysis of melodic patterns in a dataset of machine readable music scores.

Preliminary experiments with this dataset shows that, if we were to try and identify the *nawba* to which a specific score belongs by simply counting the occurrences of *centos* unique to the *ṭūbu‘* that constitute it as specified by Chaachoo, we would be unsuccessful. In fact, attempting to do so on the corpus used in this study results in a misclassification rate of 80%. This is due to certain *nawabāt* relying heavily on very general melodic sequences located around the tonic and the dominant of the modes, these general sequences can be found in many of the *centos* documented by Chaachoo as being specific to a certain *ṭāb‘* (and therefore *nawba*).

Our aim is to quantify the importance of particular melodic patterns with respect to *nawabāt* and provide an empirical ranking of melodic content for each. We must therefore rely on an approach that considers more than just frequency of occurrence, one that normalises for generality, putting forward new proposals, as to the melodic constitution of each *nawba* [6].

[1] Personal communication

2. RELATED WORK

There exists many studies into melodic pattern recognition, summaries of which have been made by Jansen et al. [15] and more recently Ren et al. [24]. Lack of agreement on the current state-of-the-art stems from the difficulty in evaluating approaches, with expertly annotated ground truth often required for performance measurement, more often than not on a study-by-study basis.

In no deviation from this trend we draw on the work of Chaachoo [6] for validation of our results and go further in suggesting/supplementing his studies with unique insights of our own. Accordingly, we aim to approach this investigation with interpretability in mind, seeking to build upon the string-based, frequency approaches found in [8, 14, 16], where patterns are represented by counting the number of instances of re-occurring sub-sections of notes in a musical sequence and their significance computed by comparing these counts, ignoring potential *interaction* between non-consecutive notes. The appeal of this method is that the theory is intuitive to a non-specialist and aligns with what a musician might consider important when characterising a musical piece melodically, an important consideration when wishing to contribute to and communicate with Chaachoo and his works.

The choice to consider only consecutive notes as belonging to the same pattern is supported by the nature of the music scores in our dataset (see section 3.1). These scores are manual transcriptions by Chaachoo himself from a collection of representative recordings. Due to the heterophonic nature of this music tradition, and the analytical purpose of the music scores, they contain only the common melodic line underlying the actual rendition of the choir and instruments in the orchestra. This rather prescriptive character of the transcriptions results in a representation of the music more theoretical than fine detailed, thus permitting the assumption that *centos* are literally repeated in the score, without the modifications of the ornamentation in actual performance.

As mentioned in the previous section however, simple frequency alone is not sufficiently powerful in characterising *nawba*. A more desirable attitude would be probabilistic, as found with Conklin in [9]. Conklin puts forth a novel method of computing pattern significance by comparing the probability of occurrence in a corpus with the probability of occurrence in an anti-corpus, patterns over-represented in the former are said to be distinct to that. Our tf-idf approach is very similar, the probability of occurrence of each pattern in a score is normalised by how likely it is to occur across all other scores.

Nawba number	*Nawba* transliterated name	Number of Scores
1	*raml al-māya*	19
2	*al-isbahān*	13
3	*al-māya*	13
4	*rasd al-ḏāyl*	18
5	*al-istihlāl*	24
6	*al-rasd*	10
7	*garībat al-ḥusayn*	13
8	*al-ḥiŷāz al-kabīr*	10
9	*al-ḥiŷāz al-māšriqī*	15
10	*'irāq al-'aŷam*	7
11	*al-'uššā*q	7

Table 1: Distribution of Scores across *nawabāt*

3. METHODOLOGY

Our analysis is implemented in Python, the code for which is available on Github [2] should the reader wish to reproduce this work. We use the music21 library [10] for processing scores and adopt the same convention for accidental notes in our reporting as in this toolkit, that is '#' for a sharp and '-' for a flat.

It is worth noting that we do not take into account any note/rest duration in this analysis, this decision is based on Chaachoo's theory which proposes the set of centos with no variation in their durations. Furthermore, we omit all octave information from our data, observing that melodic lines in our corpus very rarely jump between octaves and that same omission is made by Chaachoo in his work.

3.1 Dataset

Our dataset is a subset of 149 scores across all 11 Moroccan *nawabāt* from the CompMusic Arab-Andalusian corpus [25] of Dunya [23]. It has been selected such that each score is accompanied with the relevant *ṭāb'* metadata (and hence allows us to identify the *nawba* to which they belong). Each score is represented as an ordered list of consecutive notes (all scores are monophonic) with rests included. Table 1 shows the number of scores for each *nawba*.

3.2 Pre-processing

In analogy to a bag-of-words representation of a document, we represent each score as a bag-of-patterns. That is, we extract from each every possible n-gram up to a specified length, N. Any n-gram (or pattern as they will from here on be referred to as) that contains a rest, R, is discarded. For example, the bag-of-patterns for a score of [G, E, F, F, R, E, G, E] is:

$N = 2$: [GE, EF, FF, EG, GE];
$N = 3$: [GE, GEF, EF, EFF, FF, EG, EGE, GE];
$N \geq 4$: [GE, GEF, GEFF, EF, EFF, FF, EG, EGE, GE]

This method of motivic representation - where every possible sub-pattern is included alongside its parent patterns down to minimum length - is said to satisfy the *Submotif Existence Axiom* (SEA), first proposed in [18] and realised by Buteau in [2–4, 12].

3.3 TF-IDF Algorithm

Pattern importance is computed using the tf-idf statistic. Tf-idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [26]. With each of our 149 scores in bag-of-patterns form, we compute the tf-idf for each pattern on a score by score basis, using sub-linear term frequency so as to down-weight very popular or very rare terms.

Averaging our tf-idf values for each pattern for each *nawba* provides a metric for pattern importance per *nawba*.

3.4 Post-processing

A common problem in melodic pattern recognition by computational means is in dealing with the large quantity of results. As is typical of such methods, a set of selection rules is required. After applying the tf-idf algorithm to identify significant patterns, they are filtered according to the following:

1. We consider a pattern having occurred less than 50 times per score per nawba as being insufficiently frequent given our data size, this choice is informed by inspecting the frequency of patterns identified by Chaachoo as characteristic.
2. We do not consider strings that contain less than three notes or more than ten notes a pattern, these are discarded. This decision is informed by referring to existing literature on melodic composition of Arab-Andalusian music [6, 20, 25] and is somewhat justifiable intuitively, the same lower bound can be found in [9, 21] for example.
3. Any pattern that is a substring of another pattern in our selection of significant patterns (subject to previous selection rules) is discarded in favour of the longer pattern. A similar approach is adopted in [9, 14, 17, 19].

4. RESULTS AND EVALUATION

For each *nawba* we rank its pattern content by our averaged tf-idf, cross-referencing our findings with the patterns identified as characteristic of the *nawba* from Chaachoo.

4.1 Visualising Patterns Importance

Fig.2 and Fig.3 show the top ranked patterns for *nawba* 1 and 2 respectively. Each bar is coloured such that patterns identified as characteristic of the *nawba* by Chaachoo are black, patterns that themselves contain patterns identified

[2] `https://github.com/MTG/quantifyingcentones`

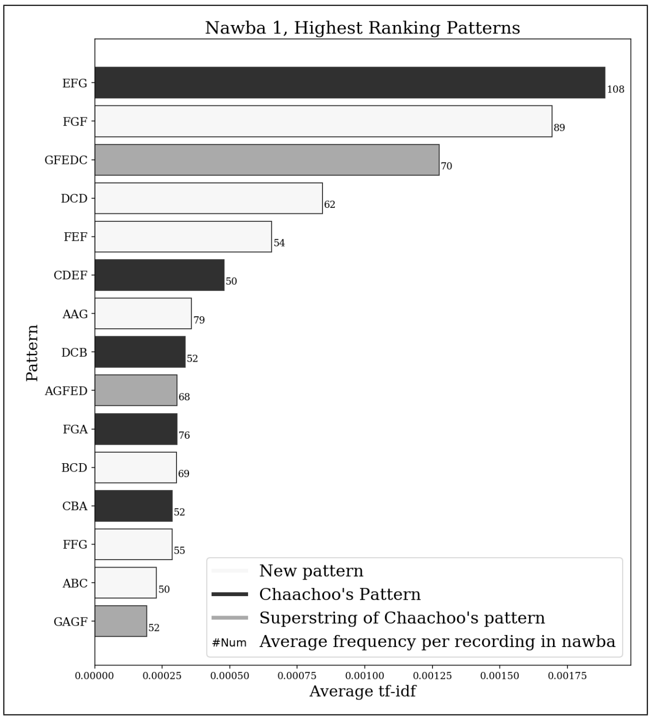

Figure 2: Ranked pattern importance for Nawba 1

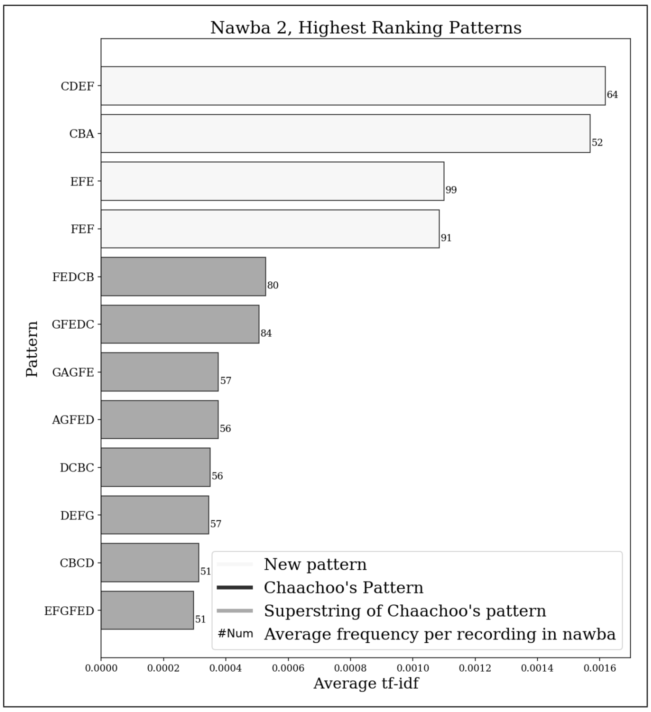

Figure 3: Ranked pattern importance for Nawba 2

by Chaachoo as characteristic of the *nawba* are grey and hitherto undocumented patterns white. Each bar is annotated with the average frequency of occurrence of that pattern in that *nawba*; that is the total number of occurrences in the corpus as a proportion of how many scores we have for that *nawba* (Table 1). These two examples are representative cases in which the most relevant patterns are a mix of Chachoo's patterns and new ones. Fig.4 displays the same score excerpt as in Fig.1 but with the new recognized relevant patterns for the *nawba*.

Table 3 presents the top 10 (where 10 exist) patterns determined by our analysis to be most characteristic of each *nawba*. The patterns identified by Chaachoo as characteristic of the *nawba* are bolded, those that themselves contain patterns identified by Chaachoo as characteristic of the *nawba* are italicized.

4.2 Evaluation

As a method of evaluating which patterns are most characteristic of each *nawba*, we compare the classification power of the patterns identified in this study to those documented by Chaachoo in [6] through a simple classifier.

4.2.1 Simple Classifier

An l2 regularised logistic regression model is applied to our corpus of 149 scores, we use a 60/40 train/test split and the frequency of occurrence of each pattern as features. In total, there are 184 patterns from Chaachoo's literature and 182 from our own.

We bootstrap our accuracy score on test 100 times. The results are displayed in Table 2.

Pattern Set	μ	σ
Chaachoo	70.8	4.9
Ours (Table 3)	72.2	5.0

Table 2: Bootstrapped accuracy when classifying nawba using two pattern sets (n=100)

5. DISCUSSION

The patterns discovered by this analysis are found to be as predictive as those of Chaachoo for classifying *nawabāt* and 44% of our top 109 patterns match the *centos* outlined in his works. Furthermore, we have demonstrated that in Arab-Andalusian music of the Moroccan tradition there exists at least 61 unique melodic patterns that occur as frequently (and in many cases more frequently) as those identified and documented to date and with at least as much significance as measured via our normalised frequency approach.

From a musical perspective it is clear that our patterns share some similarities; they constitute very simple cells, based on the relevant degrees of the *ṭāb'* (tonic and dominant). Since Arab-Andalusian music has uniform melodic contour and intervals rarely larger than a third, some of these new patterns could probably be clustered into other more generic groups. This could mean that we are not discovering new important material rather than variations of the known *centos* proposed previously by Chaachoo.

6. CONCLUSIONS

Our methodology is not without its limitations. The variability of our results is conditioned by factors such as whether we include rests in our patterns (and if so to what

Nawba	Most Characteristic Patterns
1	**EFG**, FGF, *GFEDC*, DCD, FEF, **CDEF**, AAG, **DCB**, *AGFED*, **FGA**
2	CDEF, CBA, EFE, FEF, *FEDCB*, *GFEDC*, *GAGFE*, *AGFED*, *DCBC*, *DEFG*
3	**B-AG**, EEF, FGEF, FGA, DCB, *AGFEDC*, *GAGFE*, EFEDC, CEE, *DEFG*
4	**CDEC**, ECD, *DEFG*, GAGFE, EDCDE, CDEF, AGFEDC, FAG, CDC, CCD
5	**ABC**, DCBA, **FAG**, *AGFEDC*, *EDCB*, *GAGFE*, *DEFG*, BCD, EDE, CBAG
6	**EDC**, *AGED*, *ABAGE*, *EDECD*, *EF#GA*, *F#GAG*, GF#G, GAB, AGG, GGA
7	EDE, *GAGFE*, *DEFG*, **CDE**, FGA, *AGFED*, AAG, *GFEDC*, CCD, DCD
8	E-DC, GAGF#E-, **AGF#E-D**, DEF#, BAGF#, **EF#GAG**, **CBAG**, DCB, BCB
9	EFG, FEF, FGF, **FGA**, FFG, *AGFED*, *GFEDC*, GAG, DEF, EFE
10	BAGF#E, **GF#ED**, *F#GAG*, CBAG, *EF#GA*, EDC, GAB, GGA, DCBA, ABC
11	GAG, GFED, DED, **GAB**, **FEDC**, *ABC*, EDCB, GGA, EFE, **CBAG**

Table 3: Top 10 most characteristic patterns per *nawba*. Those bolded exactly match patterns identified by Chaachoo. Those italicized are superstrings of one of those identified by Chaachoo

Figure 4: Score example with the new patterns discovered of the piece *Btayhi Rasd Dayl* of Nawba 4

length), the minimum frequency of occurrence for a pattern to be considered significant and maximum/minimum pattern length. The alteration of these and others like them would likely have marked consequences for our results. The justification for our decision was driven by necessity and intuition but would indeed benefit from further collaboration with experts of the tradition.

A further limitation exists in our inability to compare the significances computed here with other methods given that the tf-idf statistic is proportional to corpus size and hence only comparable within our training corpus.

However, as one of the first computational analysis on Arab- Andalusian music, we hope to have contributed to the musicological theory around centonization and hope that our approach may serve as a first reference for pattern recognition in Arab-Andalusian music, establishing the principles and basis for future and helpful study for musicological theories that can contribute to a better understanding and preservation of the musical tradition.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge Mr. Amin Chaachoo, expert musicologist and experienced musician of this tradition, main curator of the Arab-Andalusian corpus and author of the music transcriptions and for his help with musicological explanations. This work was carried out with the support of the Musical Bridges project, funded by RecerCaixa.

8. REFERENCES

[1] W. Apel. *Gregorian Chant*. A Midland book. Indiana University Press, 1958.

[2] C. Buteau. Melodic clustering within topological spaces of schumann's träumerei. *International Computer Music Conference, ICMC 2006*, 01 2006.

[3] C. Buteau and G. Mazzola. Motivic analysis according to Rudolph Réti: formalization by a topological model. volume 2, pages 117–134. Taylor & Francis, 2008.

[4] C. Buteau and J. Vipperman. Melodic clustering within motivic spaces: Visualization in openmusic and application to schumann's träumerei. volume 37, pages 59–66, 01 2009.

[5] A. Chaachoo. *La música andalusí Al-Ála: Historia, conceptos y teoría musical*. Editorial Almuzara, Córdoba, 2011.

[6] A. Chaachoo. *La musique hispano-arabe, al-Ala*. Univers musical. Editions L'Harmattan, 2016.

[7] G. Chewand and J. W. McKinnon. Centonization. *Oxford Music Online*, 2001.

[8] D. Conklin. Discovery of distinctive patterns in music. In *Intell. Data Anal., Vol 14*, pages 547–554, 2010.

[9] D. Conklin and C. Anagnostopoulou. Comparative pattern analysis of cretan folk songs. In *Proceedings of 3rd International Workshop on Machine Learning and Music*, MML '10, pages 33–36, New York, NY, USA, 2010. ACM.

[10] M. Cuthbert, C. Scott, and C. Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 637–642. International Society for Music Information Retrieval, 2010.

[11] P. Ferretti and A. Agaësse. *Esthétique grégorienne ou Traité des formes musicales du chant grégorien. Volume I. Traduit de l'italien par Dom A. Agaësse*. Desclée, 1938.

[12] H. Fripertinger and L. Reich. Topological motive spaces, and mappings of scores' motivic evolution trees. *Grazer mathematische Berichte*, (347):35, 2005.

[13] M. Guettat. La musique arabo-andalouse, l'empreinte du maghreb. page 560, 2000.

[14] J. Hsu, C. Liu, and A. L. P. Chen. Discovering nontrivial repeating patterns in music data. In *IEEE Transactions Multimedia*, pages 311–325, 2001.

[15] B. Janssen, W. De Haas, A. Volk, and P. Van Kranenburg. Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, Marseille, France*, volume 20, page 74, 2013.

[16] I. Karydis, A. Nanopoulos, and Y. Manolopoulos. Finding maximum-length repeating patterns in music databases. In *Multimedia Tools and Applications*, volume 32, pages 49–71, 2006.

[17] I. Karydis, A. Nanopoulos, and Y. Manolopoulos. Closed patterns in folk music and other genres. In *6th International Workshop on Folk Music Analysis*, pages 15–17, 2016.

[18] G. Mazzola. *The Topos of Music: Geometric Logic of Concepts, Theory, and Performance*. Birkhäuser, 2011.

[19] D. Meredith, K. Lemström, and G. A. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

[20] M.Sordo, A. Chaachoo, and X.Serra. Creating corpora for computational research in arab-andalusian music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology (DLfM '14). ACM, New York, NY, USA, 1-3. DOI: https://doi.org/10.1145/2660168.2660182*, 2014.

[21] O. Nieto and M. M. Farbood. Perceptual evaluation of automatically extracted musical motives. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, pages 723–727, 2012.

[22] C. Poché. *La música arábigo-andaluza (con CD)*. Músicas del mundo. Ediciones Akal, 1997.

[23] A. Porter, M. Sordo, and X. Serra. Dunya: A system for browsing audio music collections exploiting cultural context. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 101–106, Curitiba, Brazil, 04/11/2013 2013.

[24] I. Y. Ren, H. V. Renand Koop, A. Volk, and W. Swierstra. In search of the consensus among musical pattern discovery algorithms. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 671–678, 2017.

[25] R. Caro Repetto, N. Pretto, A. Chaachoo, B. Bozkurt, and X. Serra. An open corpus for the computational research of arab-andalusian music. In *Proceedings of 5th International Conference on Digital Libraries for Musicology (DLfM 2018)*, pages 78–86, 2018.

[26] H. Wu, R. W. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3):13:1–13:37, June 2008.