Musical Data Mining for Electronic Music Distribution

François Pachet, Gert Westermann, Damien Laigre
Sony CSL-Paris, 6, rue Amyot, 75005 Paris, France
pachet@csl.sony.fr

Abstract

Music classification is a key ingredient for electronic music distribution. Because of the lack of standards in music classification – or the lack of enforcement of existing standards – there is a huge amount of unclassified titles of music in the world. In this paper we propose a method of classification based on musical data mining techniques that uses co-occurrence and correlation analysis for classification. This method allows for the rapid extraction of similarities between musical titles or artists. We investigate the usability of radio playlists and compilation CD databases for our data mining technique, and we compare the generated results with human similarity judgments. Based on a clustering technique, we show that interesting clusters can reveal specific music genres and allow classifying titles of music in a kind of objective manner.

1 Introduction

Electronic Music Distribution (EMD) concerns the digital transportation of music through networks. It has been gaining attention for a number of years, due to progress in data compression and network telecommunications. The number of musical titles, considering only Western music, ranges in the several millions. Besides issues related to copy protection and copyright management, the mere possibility of transporting these millions of music titles easily and efficiently raises the issue of *content management*: how to design efficient means of accessing, retrieving and exploring music titles?

One of the most successful approaches to this issue is similarity-based search. Similarity-based search allows users to find titles based on examples and counter-examples, without the drawbacks like the language mismatch problem [1] faced by explicit symbolic queries.

There are three main ways to extract musical similarities: signal-based approaches, collaborative filtering, and data mining. Signal-based approaches usually extract low-level descriptors such as tempo [2], fundamental frequency [3] or segmentation structure [4]. Current projects are devoted to extracting high-level descriptors (e.g. the Cuidado European IST project), with some preliminary results such as rhythm structure extraction [5]. These descriptors usually provide grounded, objective distance functions that can be used for similarity-based search. However, the descriptors are not yet sufficiently sophisticated to provide similarities at the music title level. Collaborative filtering techniques [6] are based on the comparison of user profiles, and they represent the main technique used today for music recommendation systems (Amazon, AllMusicGuide, etc.). The advantage of collaborative filtering is that the technique is relatively simple to implement. The main drawback is that it requires a huge number of actual users of a given system to be meaningful.

In this paper we study one particularly efficient means of extracting similarity for music titles, data mining. Although collaborative filtering may be considered as a particular form of data mining, it differs from the approach followed here in that it is based on subjective information (user's declared taste or rankings).

In this paper we consider techniques that use various sources of more objective information about music titles.

The rest of the paper is organized as follows: in section 2, we describe the background for musical data mining and identify corpora on which this technique can operate. Section 3 explains in detail the process of extracting titles and establishing similarity measurements between them. In section 4, we analyse the results of our different techniques and compare them with human judgments of similarity between music titles. In section 5, we discuss directions for future work.

2 Musical Data Mining

The notion of similarity is a complex one. For music, it is particularly complex because there are numerous dimensions of similarity: objective similarity based on musical features such as tempo, rhythm, timbre, but also less objective features such as musical genre, personal history, social context (e.g. music from the 60's), and a priori knowledge (e.g. the relation between The Beatles and Paul McCartney).

Since we are looking for purely automatic methods of similarity detection, it is difficult to make a priori distinctions regarding the nature of the similarity we extract. In a first step, our aim is 1) to determine whether data mining techniques can actually discover any kind of relevant similarity between music titles, and 2) to

characterize as much as possible the nature of these similarities, and if possible compare them with other sources of similarity discovery.

In this study, we choose to use a well-known technique used in statistical linguistics: co-occurrence analysis. Co-occurrence analysis is based on a simple idea: if two items appear in the same context, this is evidence that there is some kind of similarity between them. In linguistics, co-occurrence analysis based on large corpora of written and spoken text has been used to extract clusters of semantically related words [7]. Similarity measurements based on co-occurrence counts have been demonstrated to be cognitively plausible [8].

Here we do not seek to model the cognitive processes underlying similarity judgments in music, but we wish to simply compute similarity in various corpora, and evaluate our results for the purpose of EMD applications.

The first task in applying co-occurrence techniques is to identify relevant corpora. We have investigated two possible such sources: radio programs, and databases of compilation CDs.

2.1 Radio programs

The rationale behind analysing radio programs is that usually, at least for certain radio stations, the choice of the titles played and the choice of their sequence is not arbitrary. The radio programmer has, in general, a vast knowledge of the music he or she plays on air, and this knowledge is precisely what gives the program its characteristic touch. For instance, some radio stations specialize in back catalogues of the sixties (in France e.g. Radio Nostalgie and Europe 2), others in non-contemporary classical music (Radio Classique), and yet others have more diverse catalogues (such as FIP/Radio France). In all cases, however, the titles and their sequencing are carefully selected in order to avoid breaking the identity of the program.

It is this very knowledge (choice of titles and choice of sequencing) that we wish to utilize by data mining. Here, the co-occurrence analysis consists in testing how titles are actually chained together.

Several thousands radio stations exist in the occidental world, and many of them make their programs available on the web, or through various central organizations, such as Broadcast Data Systems. For our experiments we have chosen a French radio station that has the advantage of not being specialized in a particular music genre: Fip (Radio France).

2.2 Track Listing Databases

Another important source of information is actual CD albums, and in particular, samplers (compilations). Compilations, either official ones produced by labels, or those made by individuals, often carry some overall consistency. For instance, titles on compilations such as "Best of Italian Love Songs", "French Baroque Music", or "Hits of 1984" have explicit similarities of various sorts (here, social impact, genre, and period). Our main hypothesis is that if two titles co-occur in different compilations, this reinforces the evidence of some form of similarity between them.

3 Extracting Similarities

The automated extraction of similarities based on co-occurrence analysis requires three main steps: 1) information gathering, 2) music title and artist identification, and 3) co-occurrence analysis per se.

3.1 Information gathering

Web robots were implemented to automatically query the web servers containing appropriate information. The output of this phase is a collection of text files, each file representing either an album, a radio program, or any document containing at least two music titles, said to be co-occurring, as illustrated in Figure 1, 2 and 3.

The technique consists in first identifying a sample database. We have conducted experiments with databases of various sizes as described below. For every pair of titles in the database we perform a query in each of the data sources, to look for documents containing both items. In the case of radio programs, this query is slightly modified to ensure that the two titles are actually neighbours in the play list. We assume that co-occurrence is a symmetrical function so there is a total of `n(n-1)/2` queries to perform for a database of size n.

```
Tracks on this CD
Tears For Fears - Everybody Wants To
Rule The World
Split Enz - Message To My Girl
Suzanne Vega - Marlene On The Wall
The Bluebells - Young At Heart
James Brown - I Got You (I Feel Good)
The Christians - Harvest Of The World
Big Country Fields Of Fire
Roger Daltrey - Giving It All Away
The Moody Blues - Nights In White Satin
The Mission - Butterfly On A Wheel
Curiosity Killed The Cat - Down To
Earth
Was Not Was - Papa Was A Rolling Stone
D.N.A. Featuring Suzanne Vega - Tom's
Diner
Army Of Lovers - Give My Life
Yello - The Race
```

Figure 1 Example of a CDDB track listing. All titles in this compilation are said to be co-occurring. CDDB is a large CD database that is available on the web.

```
2:50 CUNNIE WILLIAMS       MY FATHER S
WORDS  COMIN FROM THE HEART OF THE
GHETTO (1994 YO MAMA)
2:54 GIL SCOTT HERON       NEW YORK
CITY   GLORY...THE GIL SCOTT HERON
COLLECTION (1977 ARISTA)
2:59 STEELY DAN     GASLIGHTING ABBIE
  TWO AGAINST NATURE (2000 BMG)
3:05 EDDY MITCHELL  HIP HUG HER   J
AI DES GOUTS SIMPLES
3:10 CHRIS JOSS     THE MAN WITH A
SUITCASE     MUSIC FROM THE MAN WITH
A SUITCASE (1999 PULP FLA)
3:14 TEARS FOR FEARS       EVERYBODY
WANTS TO RULE THE WORLD    S FRIENDS
(1992) (1985 EPIC)
3:18 NEY MATOGROSSO POEMA  OLHOS DE
FAROL (1999 EMARCY)
3:22 SMADJ   GLOGG  EQUILIBRISTE
(1999 MELT 200)
3:27 PORTISHEAD     ONLY YOU
  SINGLE 2 TITRES (1998 GO BEAT)
3:30 BOBBY WOMACK   SUMMERTIME   RED
HOT (1998 ANTILLES)
3:36 DIRECTION MICHEL PLASSON
  BERCEUSE EN RE MAJEUR   FAURE: L
OEUVRE D ORCHESTRE VOL II/PLASSON
(1979 EMI)
```

Figure 2 An excerpt of a Fip Radio program. Each title co-occurs with its direct neighbours.

http://www.amazon.com/…
…
1962-1966: The Red Album by **the Beatles**
…
Disc: 2- 12. **Eleanor Rigby**
…
Comment from a customer:
"Still, with just about every song here an absolute classic (the remainder are simply "great"), this essential album is as important to pop as Beethoven's symphonies and **Mozart's Requiem** are to classical music."

Figure 3 An excerpt of a web page (on Amazon) containing both occurrences of "Eleanor Rigby", "The Beatles", "Mozart", and "Requiem", thereby incrementing by one the number of co-occurrences of these two titles.

3.2 Title Identification

An important next step is then to actually identify the music titles and artists, based on the textual information provided by the various sources. This is a difficult problem indeed, because most of the time the title information in input by hand, by various kinds of people (in CDDB, it can be any individual), and without any general syntactic rule. Although the music industry has defined a standard music title reference (the ISRC code), it is usually not used for referencing music titles in existing information database such as the ones used here. For instance, a title such as Eleanor Rigby by The Beatles, could appear under a variety of formats such as:

- The Beatles – Eleanor Rigby,
- Eleanor Rigby / Beatles, The
- ELEANOR RIGBY; Beatles; Revolver – Track 2,

Etc.

We have designed a system that infers the most probable syntax from a given collection of track names, and is able to eventually identify the artist name (e.g. "THE BEATLES"), in a non-ambiguous fashion, and the title name ("ELEANOR RIGBY") with a high degree of success. Additionally, an ad hoc indexing procedure allows matching artist and title names independently of special characters, separators, and non-digit or letter characters. Special rules have also been introduced to handle frequent cases such as artist with or without "The" (e.g. Beatles appear also as THE BEATLES, or as BEATLES, The).

3.3 Co-Occurrence Analysis

Co-occurrence analysis consists in building a matrix with all titles in row and in column. The value at (*i*, *j*) corresponds to the number of times that titles *i* and *j* appeared together, either on the same sampler, on the same web page, or as neighbours in a given radio program. To define an actual distance function, we need to take into account several important factors.

First, two titles may never co-occur directly, but they may each co-occur with a third title. The distance function should take such indirect co-occurrence into account. Second, because we want to assess both the soundness (all found similarities are 'good') and completeness (all 'good' similarities are found) of the extracted similarities, we need to restrict the validation to a close corpus of titles that can then be used for comparisons with human similarity judgments.

Given a corpus of titles $S = (T_1, \ldots, T_N)$, we compute the co-occurrence between all pairs of titles T_i and T_j. The co-occurrence of T_i with itself is simply the number of occurrences of T_i in the considered corpus. Each title is thus represented as a vector, with the components of the vector being the co-occurrence counts with the other titles. To eliminate frequency effects of the titles, components of each vector are normalized according to:

$$Cooc_{norm}\left(T^1, T^2\right) = \left(\frac{Cooc\left(T^1, T^2\right)}{Cooc\left(T^1, T^1\right)} + \frac{Cooc\left(T^2, T^1\right)}{Cooc\left(T^2, T^2\right)} \right) \Big/ 2$$

The normalized co-occurrence values can directly be used to define a distance between titles; this distance will be expressed as:

$$Dist_1(T^1, T^2) = 1 - Cooc_{norm}(T^1, T^2)$$

This first distance will be used to give what we will call *direct* similarity between titles, because it is based on the co-occurrence itself and does not reveal indirect links that a title can have with other titles. Example: if "Eleanor Rigby/The Beatles" co-occurs with "Good Vibration /The Beach Boys" and "Good Vibration/The Beach Boys" co-occurs with "God only knows/The Beach Boys", the co-occurrence measure will not show similarity between "Eleanor Rigby" and "God only knows".

A measure of similarity that takes such indirect links into account is the correlation between the vectors representing two songs. If both songs are equal and their vectors point in the same direction, the correlation is 1. If they do not share any components and are orthogonal, the correlation is -1. Given that the vectors are normalized, we can compute the correlation between two titles T^1 and T^2 as :

$$Sim(T^1, T^2) = \frac{Cov_{1,2}}{\sqrt{Cov_{1,1} \times Cov_{2,2}}}$$

where $Cov_{1,2}$ is the covariance between T^1 and T^2 and:

$$Cov(T^1, T^2) = E((T^1 - \mu_1) \times (T^2 - \mu_2))$$

E is the mathematical expectation and $\mu_i = E(T^i)$.

We then define the *distance* between T^1 and T^2 as:

$$Dist_2(T^1, T^2) = 1 - (1 + Sim(T^1, T^2))/2$$

The correlation analysis can also be performed with artists themselves instead of titles. Two samplers can contain the same artists with different titles, and a radio can broadcast different titles consecutively from two artists only. Assuming that an artist generally has a characteristic style and all songs of one artist are similar to each other, we can apply this artist-based analysis especially when the database is small.
In the analyses described below we will also discuss artist based analyses.

4 Assessing the extracted similarities

There are several validation experiments that can be performed to assess the quality of the similarities resulting from the data mining approaches. The most general situation is illustrated in Figure 4. A general evaluation would consist in assessing the respective size of the intersections between 3 sets: 1) human similarity judgments, 2) similarity extracted by radio co-occurrence, 3) similarity extracted by sampler co-occurrence.

A complete evaluation would therefore consist in answering the following questions:
- Consistency: are the human judgements consistent? Are there consensual similarities? Are the similarities extracted from radio programs and samplers the same?
- Soundness: do the extracted similarity all correspond to human judgement?
- Completeness: are all possible human similarities extracted?

It is of course difficult to prove any of these assertions, as this would imply a user evaluating all possible titles (several millions). What we can do, however, is test them on small subsets.

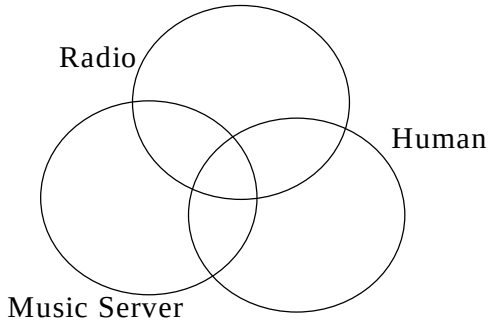

Figure 4 Assessing similarities: how do the various similarities actually match?

For the evaluation, we have defined three such subsets of the databases. One very small (12 titles), one consisting of 80 frequent titles played on the selected radio station (Fip), and one made up of 100 artists only selected by hand so as to represent different genres. Similarity analysis of artists instead of titles gives better results on a small set of items.
The clustering technique performed is the Ascendant Hierarchical Classification [9]. We call *co-occurrence clustering* the clustering applied to co-occurrences values expressed as distances. We call *correlation clustering* the clustering applied to correlation values expressed as distances.

4.1 Experiments

We illustrate our approach here with some results using the database of 12 titles, and with some results using the database of our 100 artists (chosen as the most frequent artists appearing on the radio station).
The clustering trees for the 12 titles produced based on the CDDB are illustrated in figures 5 and 6. Each clustering produces a tree. The root node contains implicitly all the titles. The numbers between parentheses indicate the respective min and max distance between two

titles in the cluster. Distances are between 0 (two titles are completely similar, i.e. co-occurred with exactly the same other titles) and 1 (never co-occurred). Since none of the selected titles of this 12 title set actually co-occurred in the radio corpus studied (over 1 year), a cluster analysis based on that surce was not done.

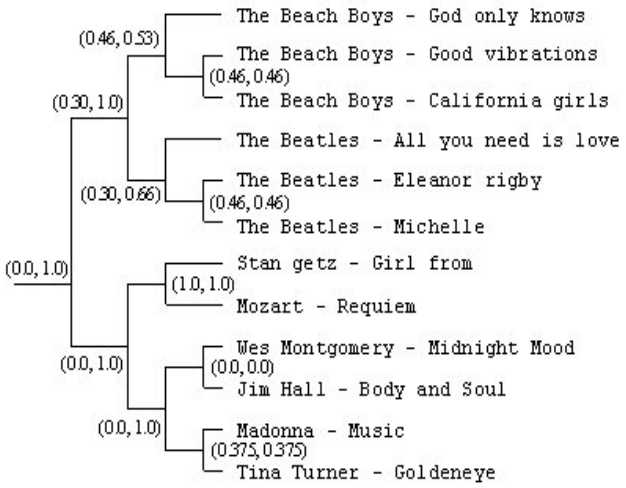

Figure 5 CDDB co-occurrence clustering.

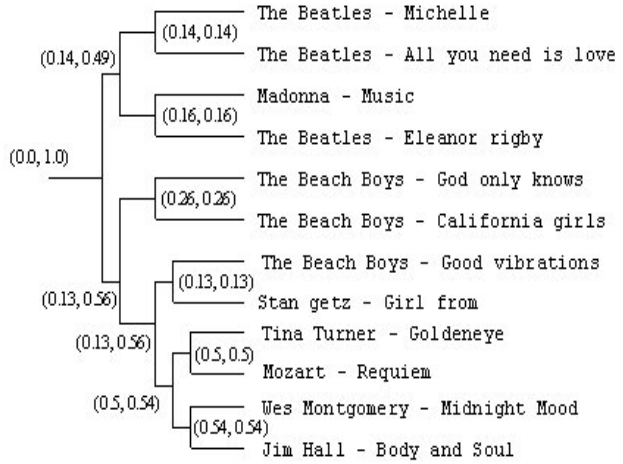

Figure 6 CDDB correlation clustering.

Both figures 5 and 6 show interesting results: specific music genres are quite well distinguished. For instance, the two jazz guitar titles (Jim Hall and Wes Montgomery) are clustered together in the process in both the co-occurrence and the correlation trees. Titles from the same artist tend to be grouped together (The Beach Boys, The Beatles). Distance values are smaller (i.e., clusters are more tight) for the correlation clustering which might therefore be preferable for this kind of similarity analysis. For general genres (Classical) the database is too small here to draw any general conclusion.

Finally, it is interesting to note that the Mozart title is actually clustered with Tina Turner/Goldeneye in the correlation clustering. The distance here is 0.5, making it a meaningful result as opposed to the distance of 1.0 to Stan Getz in the co-occurrence clustering. This detected similarity comes mainly from incidental co-occurrences of the two pieces (individual play lists published on the web), although in this case it can be argued that the symphonic nature of the soundtrack of Golden Eye is somewhat close to the symphonic orchestra playing the Requiem.

Regarding to the clusterings for 100 artists, we tried to check the consistency of similarities by comparing human judgments with the results obtained from CDDB and the FIP radio program. Five persons with a good knowledge of music (Sony CSL, Sony Music) were asked to give their judgment on the extracted similarities. The result of these judgments is shown in table 1..

Leaves of similarity trees (Level 1 clusters) alone (2 artists)	**good clusters**	**wrong clusters**	**unknown**
FIP co-occurrence clustering	70%	25%	5%
CDDB co-occurrence clustering	76%	15%	8%
FIP correlation clustering	53%	43%	4%
CDDB correlation clustering	59%	30%	11%
Level 2 clusters with 3, 4 or 5 artists			
FIP co-occurrence clustering	28%	72%	0%
CDDB co-occurrence clustering	54%	23%	23%
FIP correlation clustering	47%	38%	17%
CDDB correlation clustering	74%	19%	7%

Table 1 Human judgment of generated CDDB and FIP-based similarities

We differentiated level 1 from level 2 according to their meaning: single clusters with 2 artists can be considered as direct similarities whereas clusters with 3, 4, or 5 artists are considered as indirect similarities. For instance, 2 artists seen in a sampler or consecutively in a radio program are directly similar. If one artist is seen in two samplers with two different artists, or twice in a radio program with two different artists consecutively, these two artists are indirectly similar.

The best results are given for clusters composed by two items in the co-occurrence tree, both for the FIP and the CDDB data. This result indicates that artists or titles from the same sampler or appearing consecutively in the radio program show strong similarities. The correlation tree for such clusters gives worse results even if good clusters stil account for more than half of all clusters. Two artists or titles indeed can be similar but the link between them is less evident, example in the previous paragraph: "The Beatles / Eleanor Rigby" grouped with "Madonna / Music" instead of another title from the Beatles.

However, the correlation clustering shows very good results considering bigger clusters. Figure 7 shows two parts of the correlation clustering applied to 100 FIP artists:

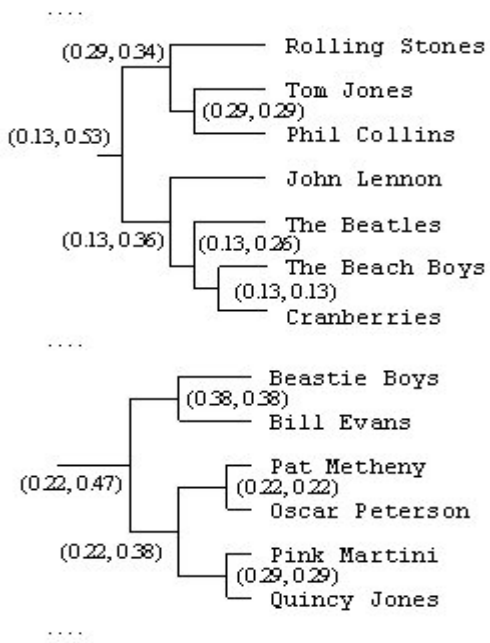

Figure 7 Part of the 100 FIP artists correlation clustering.

The first part showed includes well-known sixties' music and rock music artists, whereas the second part contains jazz artists. The co-occurrence analysis also shows surprising results with "Beastie Boys", a rock music group, clustering with "Bill Evans", a jazz musician. This result could indicate similarities between the two titles that are non-superficial.

The correlation clustering generally indicates that items in a bigger cluster tend to be classified according to their specific music genres, whereas the co-occurrence clustering is better suited for small clusters, indicating similarities between two titles only.

5 Conclusion

We have introduced co-occurrence techniques to automatically extract musical similarity between titles or between artists. The technique yields a distance matrix for arbitrary sets of items. It was applied to two different music sources, and experiments were conducted on various title and artist databases.

These experiments are still in progress, but preliminary results on small databases show that the technique is indeed able to extract similarities between items, as demonstrated by the analysis of the resulting clusters. Basic similarities such as common artist and basic genre are recognized, which validates the technique *per se*.

Characterizing the nature of the extracted similarities is trickier. Besides common artist similarities, two main kinds of similarity relations for CDDB were identified: thematic/genre similarity, and similarity of period (coming probably from the abundance of "best of the year" samplers).

For the radio (FIP), the similarity relations are quite different. Current experiments on a database of 5000 titles show that artist consistency is not enforced as systematically as in the other data sources. Moreover, the similarities are more metaphorical, and in some sense less obvious, and therefore often more interesting. They can be of various kinds: 1) *covers*, e.g. "Lady Madonna" by the Baroque ensemble is close to "Ticket to Ride" by the Beatles, 2) instrument / orchestration (e.g. Eleanor Rigby and a Haydn quartet, 3) based on title names or actual meaning of the lyrics (e.g. Kiss - Prince close to Le Baiser - Alain Souchon).

Besides scaling-up these experiments to larger databases, future work will focus on the integration of these different sources of similarity, and their actual use in EMD systems.

References

[1] Belkin, N. (2000) Helping people find what they don't know. *CACM* Vol. 43, N. 8, August 2000, pp. 58-61.

[2] Scheirer, Eric D. (1998) "Tempo and beat analysis of acoustic signals", JASA, 103(1).

[3] Lepain, Philippe (1999) Polyphonic Pitch Extraction from Musical Signals, Journal of New Music Research, 28:4.

[4] Rossignol, Stéphane & al (1998) "Features extraction and temporal segmentation of acoustic signals", Proc. ICMC.

[5] Gouyon, F. Delerue, O. Pachet, F. (2000)"Classifying percussive sounds: a matter of zero-crossing rate ?" Digital Audio Effects Conference, Verona (It).

[6] Shardanand, U. and Maes, P. (1995) Social Information Filtering: Algorithms for Automating "Word of Mouth". Proceedings of the 1995 ACM Conference on Human Factors in Computing Systems, pp. 210-217.

[7] Schütze, H. (1992) "Dimensions of Meaning", Proceedings of Supercomputing '92, pp. 787-796, Minneapolis, MN.

[8] Lowe, W. and McDonald, S. (2000) "The direct route: Mediated priming in semantic space", Proceedings of the 22nd Annual Conference of the Cognitive Science Society.

[9] E.Diday, G. Govaert, Y. Lechevallier, J. Sidi, Clustering in pattern recognition. In *Digital Image Processing*, page 19-58, J.C. Simon, R Haralick, eds, Khiwer edition, 1981.