Introduction

Classification of Grouping techniques

Graph models:

If the data is considered to be represented

through a graph where the vertices are the items or

Patterns and edges are connections between them.

Relational models.

When considering that groups should be

"Cohesive" so that the items of the same

group are closer to each other and the distance between

groups is the largest possible, One of the most extended

is the least square, where the criterion of

cohesion is obtained as the sum total of the distance

from each item to the midpoint (centroid) The method of

The k-means is based on this approach.

Introduction

Classification of Grouping techniques

Density analysis models.

When you consider that a group is a region of

space where the density of items is very high and

It is surrounded by a region of low density.

Introduction

Classification of Grouping techniques

Density analysis models.

When you consider that a group is a region of

space where the density of items is very high and

It is surrounded by a region of low density.

Exclusive groupings, non-exclusive groupings.

All the approaches discussed in the paragraphs

The previous ones start from the hypothesis of non-overlap,

When this hypothesis is relaxed, methods of

grouping that support overlap or

non-exclusive.

Introduction

Classification of Grouping techniques

Fuzzy clusters.

The non-exclusive grouping methods that have

had more 'success' are those who assume that the groups

they are fuzzy sets so that an item can

belong to different groups with a level of

Relevance to each one.

Introduction

Classification of Grouping techniques

Fuzzy clusters.

The non-exclusive grouping methods that have

had more 'success' are those who assume that the groups

they are fuzzy sets so that an item can

belong to different groups with a level of

Relevance to each one.

Agglomerative and divisive clusters.

If part of a grouping in which each item is

a group and new solutions are being built

joining groups in larger groups, we have a

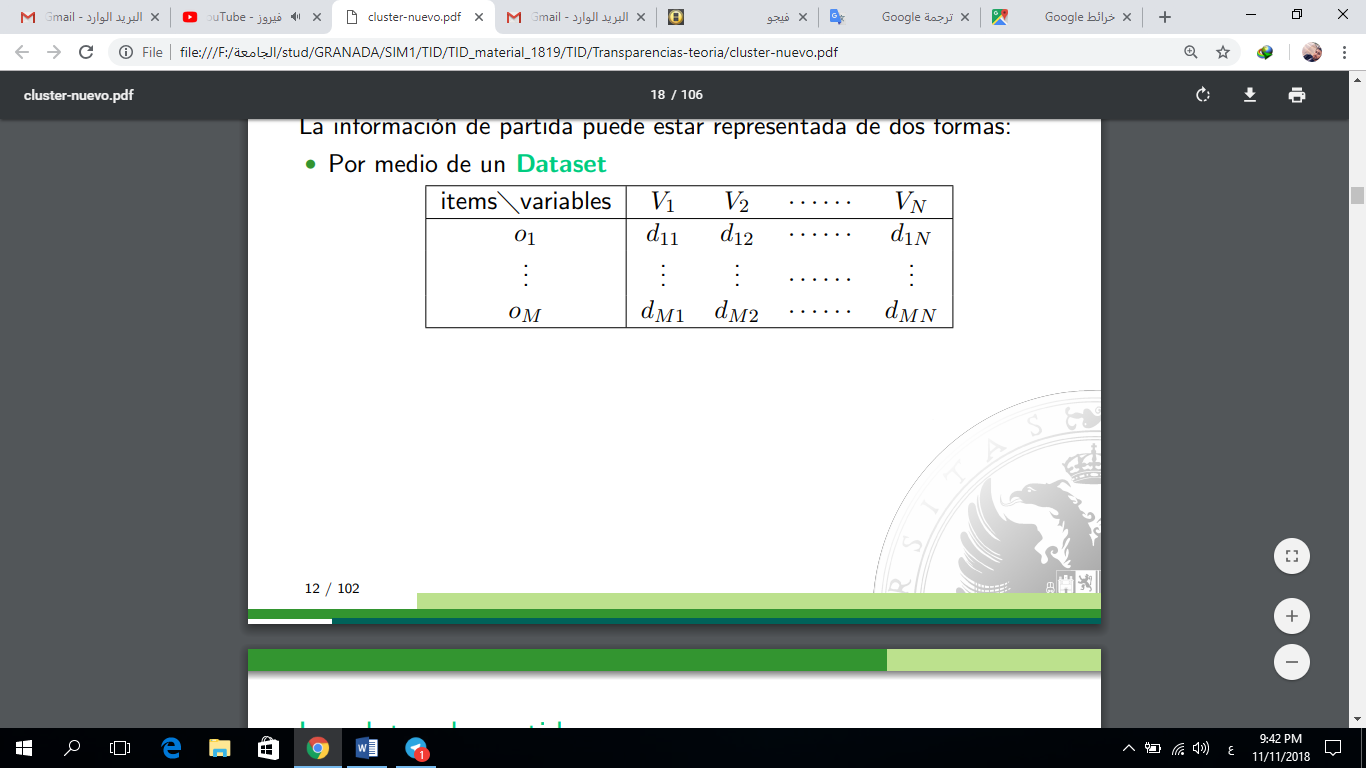
algorithm of agglomerative type, if the process is the

Otherwise, we will have a divisive algorithm.

The starting data

The starting information can be represented in two waysThe starting data

The starting information can be represented in two ways:

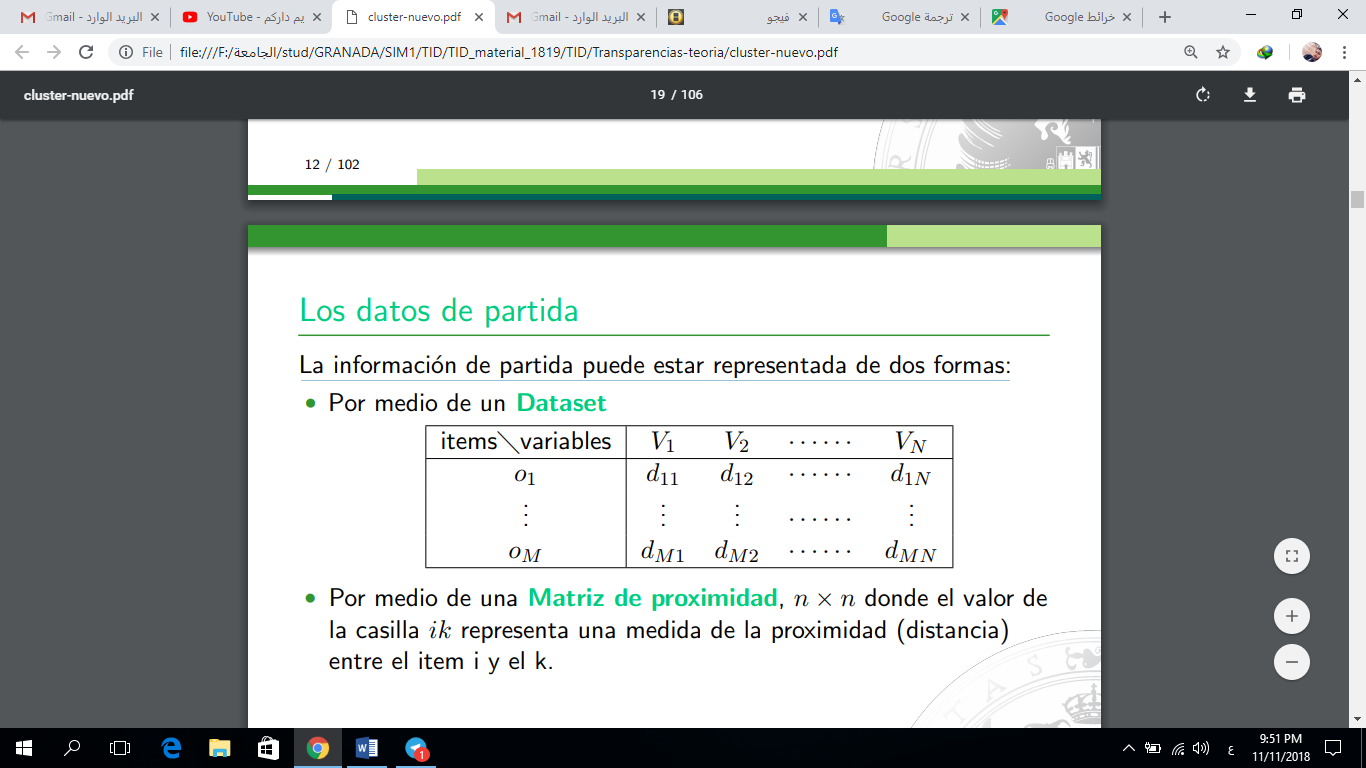
• By means of a Dataset

The starting data

The starting information can be represented in two waysThe starting data

The starting information can be represented in two ways:

• By means of a Dataset



• By means of a proximity matrix, n × n where the value of

the ik box represents a measure of proximity (distance)

between item i and k.

Usually the proximity matrix is calculated from the

data set; but in certain psychometric applications and

Sociological data may be collected directly in

Terms of agreement

Type of data

There are different types of variables to represent an item:

Quantitative are divided into:

With continuous values for example the weight of a

person, or the level of sodium in a soil.

With discrete values for example the number of

computers of a center.

Qualitative variables are divided into:

With nominal or unordered values for example the

color of a floor, or the diagnosis of a

sick.

With ordinal values for example the rank of a military

or the level of severity of a patient. A

Particular case are the binaries

that represent the presence or absence of

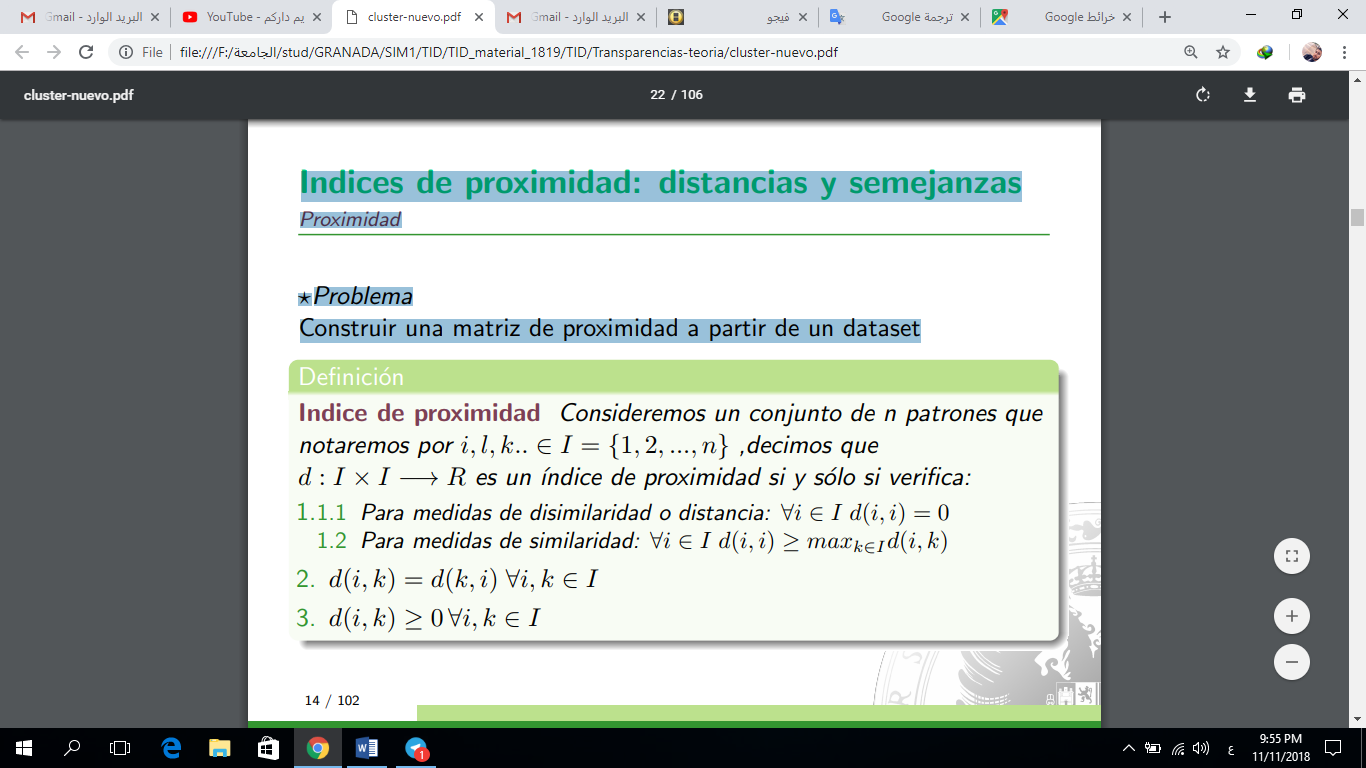
a certain characteristic.

Indices of proximity: distances and similarities

Proximity

?Issue

Build a proximity matrix from a dataset



Indices of proximity: distances and similarities

Distance functions

The 'proximity indices are a generalization of other concepts

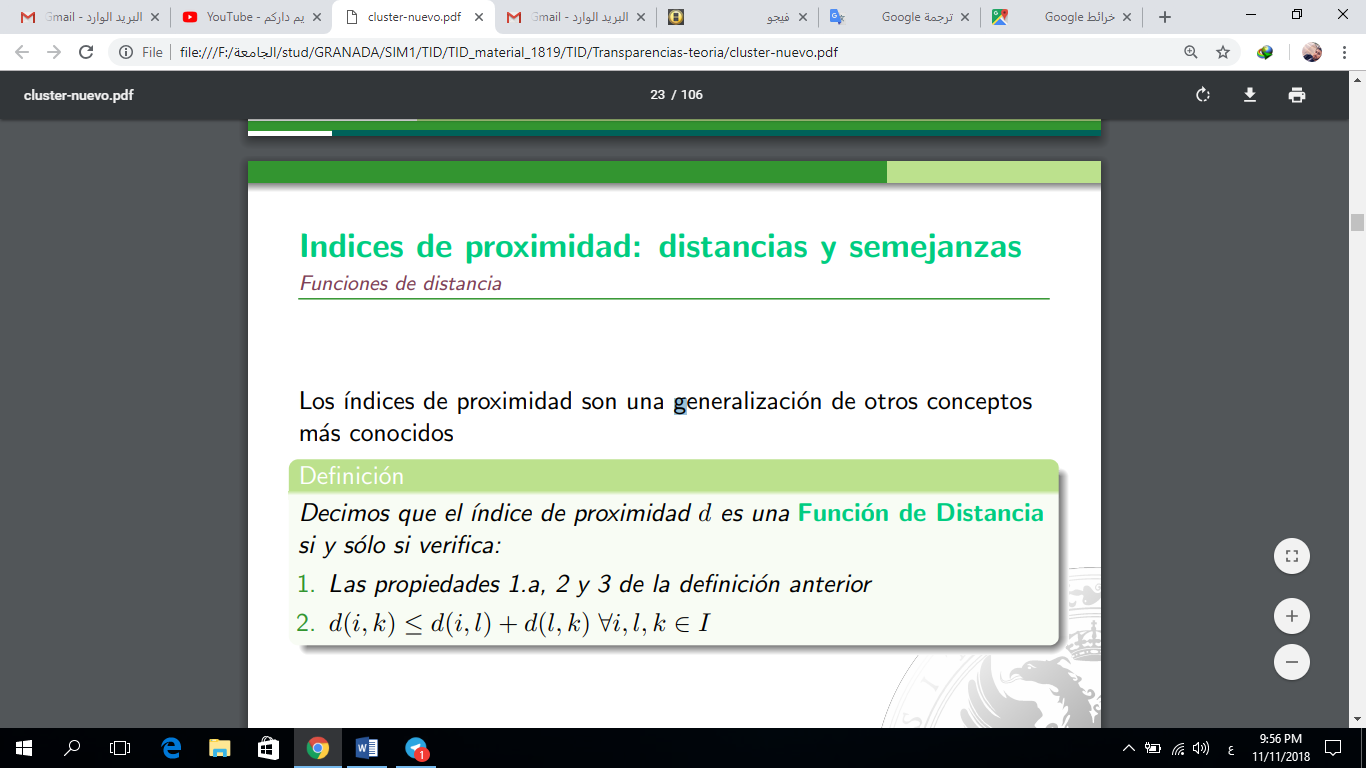
more known

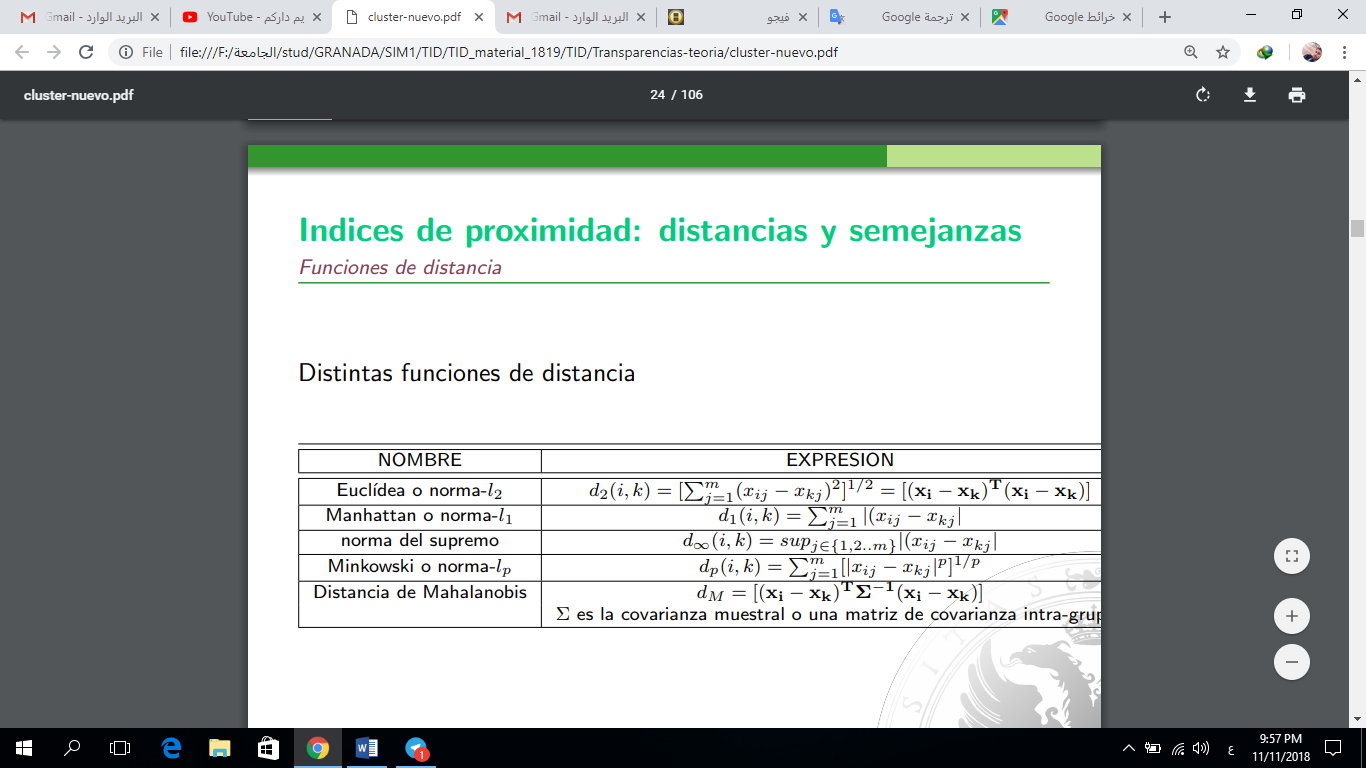
Definition

We say that the 'proximity index d is a Distance Function

if and only if it verifies:

1. Properties 1a, 2 and 3 of the previous definition





Indices of proximity: distances and similarities

Distance functions

• The distance dp generalizes to the others, the Mahalanobis distance

It also generalizes the Euclidean distance.

• There are also distance functions based on the distance of

two probability distributions and distance functions based

in the correlation coefficient that apply to the space of

variables, not the items.

• Euclidean distance is the most intuitive and works very well

when you have "compact" and "isolated" groups.

Indices of proximity: distances and similarities

Distance functions

• The main drawback of all Minkowski's metrics, in

general, is that its use gives a great weight to the variables with values

very large, this problem is solved by normalizing

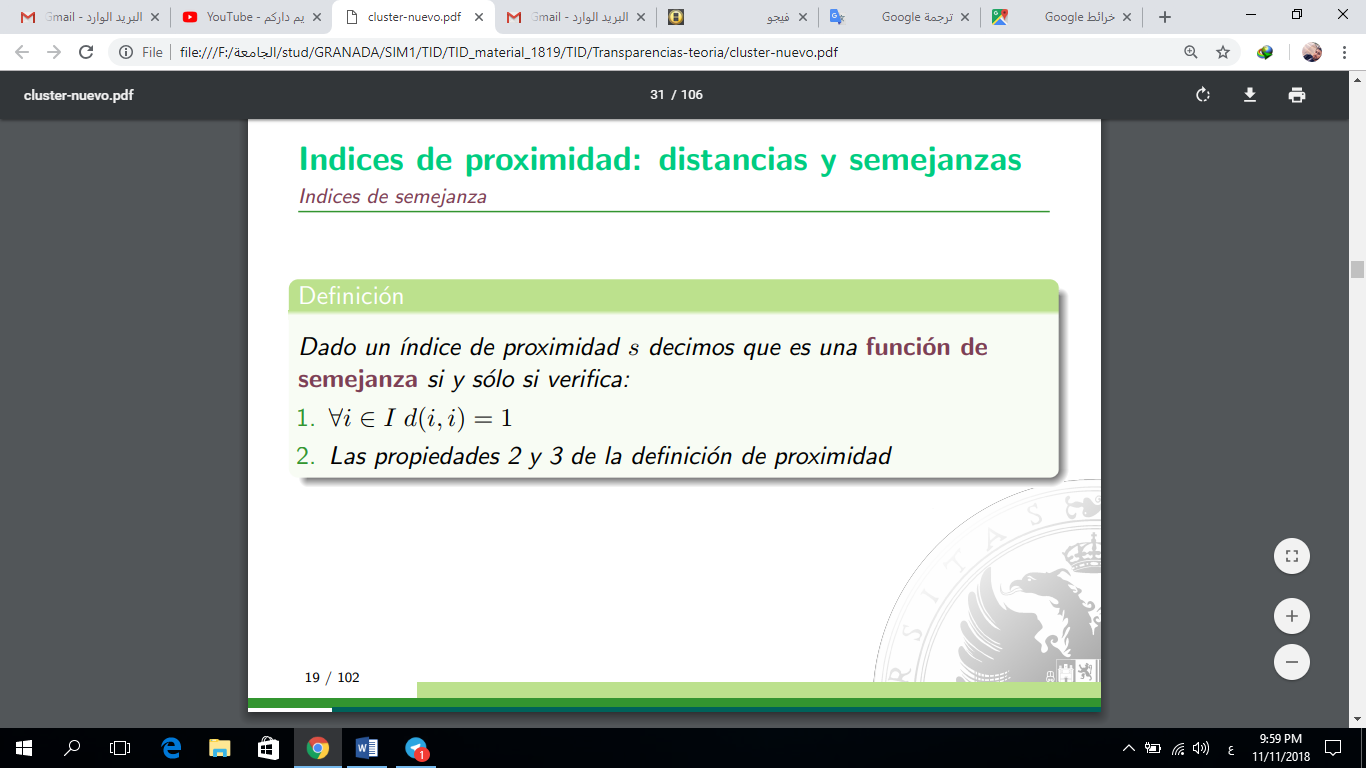
previously the variables.

• Another problem presented by continuous variables is the possible

existence of correlation between them, this problem can be mitigated

using Mahalanobis distance or previously reducing the

space.



Indices of proximity: distances and similarities

Indexes of similarity

Definition

Given a 'proximity index' we say that it is a function of

similarity if and only if it verifies:

1. ∀i ∈ I d (i, i) = 1

2. Properties 2 and 3 of the definition of proximity

A 'similarity index can be obtained from a distance:

∀i, k ∈ I s (i, k) = 1 - (d (i, k) / D) where D = maxi, kd (i, k)

Indices of proximity: distances and similarities

Indexes of similarity

Most of the similarity indices, not based on distance, are

have defined for items whose variables are binary. If we consider

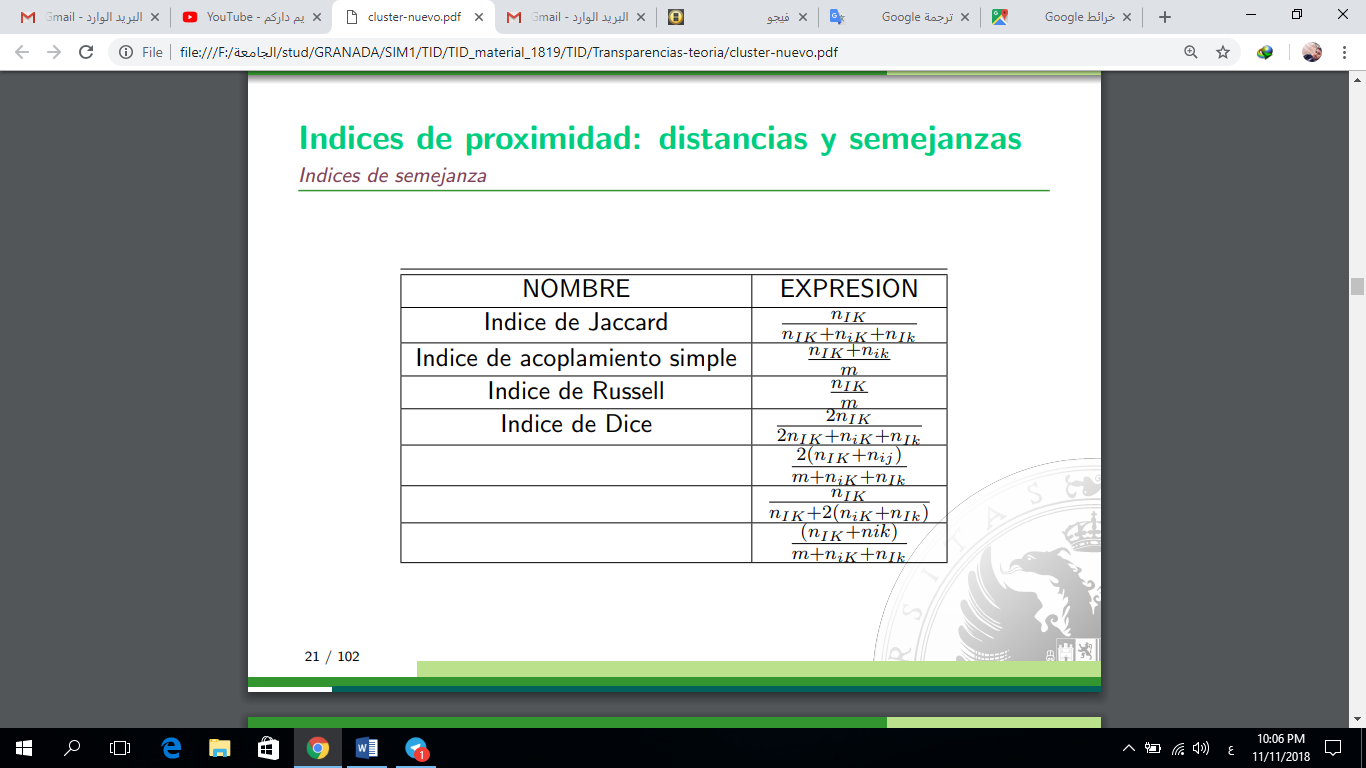
the items xi and xk, formed by m binary variables,

- nIK number of variables that take the value 1 in xi and xk

- niK number of variables that take the value 0 in xi and xk

- niK number of variables that take the value 0 in xi and 1 in xk

- nik number of variables that take the value 1 in xi and 0 in xk



Indices of proximity: distances and similarities

Important considerations

Both distances and similarities are used to

get the proximity matrix of a set of items that is

the starting point for a grouping process.

Each of the approaches corresponds to a type of variable

• Distances will be used in the presence of continuous variables, and

can be used with integer and even ordinal values assimilable to

whole, with care.

• 'Similarity indices are appropriate when working with

binary factors and can be used with nominal variables not

ordinals transforming them into a set of binary factors.

• When it comes to nominal values you can use relationships

of previous diffuse similarity

Indices of proximity: distances and similarities

Important considerations

It is important to keep in mind that it can be problematic

mix approaches directly, when you have several types of

variables It will be necessary to establish combinations of distances and / or

Conveniently standardized similarities

Preparing data and selecting distance are crucial

in the grouping processes. It is usual for

results depend a lot on these two points and for

Both of each type of problem. Usually the

selection of both is a long process that involves

several essays

Hierarchical clustering techniques

Basic ideas

• A hierarchical grouping is a succession of partitions

"Nested"

• Each group of items belonging to a certain partition

is totally included in some group of the following partition

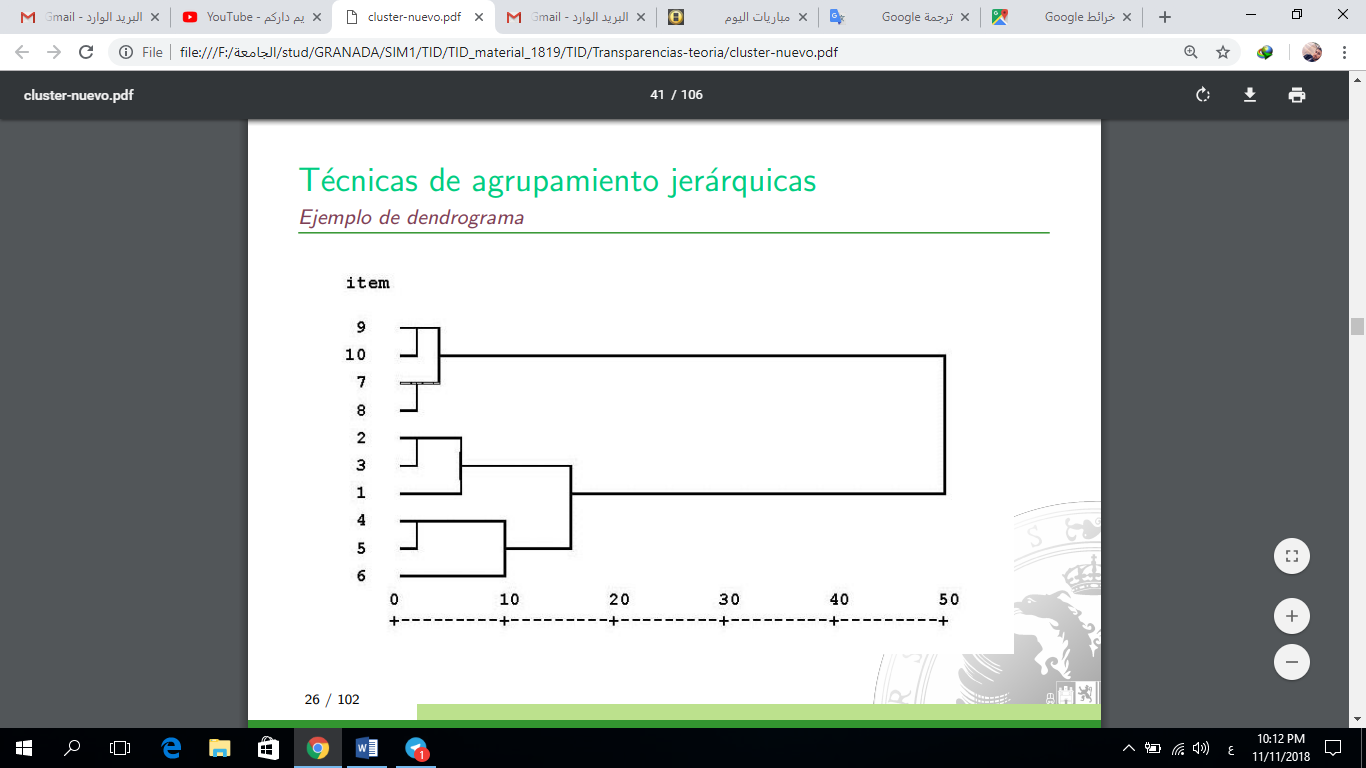
• This structure has a very intuitive graphical representation that

It is called Dendrogram. Where is it presented as they go

uniting the different patterns in groups

• Obviously the union criterion is obtained from the matrix of

distance, through algorithmic processes



Hierarchical clustering techniques

Algorithms

Most of the algorithms are of agglomerative type, starting from

a partition in which each item is a group are obtained

new partitions joining groups with each other.

Graph-based approach

It is considered that each item is a vertex of a graph and they go

generating partitions, connecting vertices of less distance,

Two forms appear:

Simple link grouping (Single-link clustering) Groups are

obtained by looking for the related components of graph and

It ends when all vertexes are connected.

Complete link grouping (Complete-link clustering)

groups are obtained by looking for subgraphs

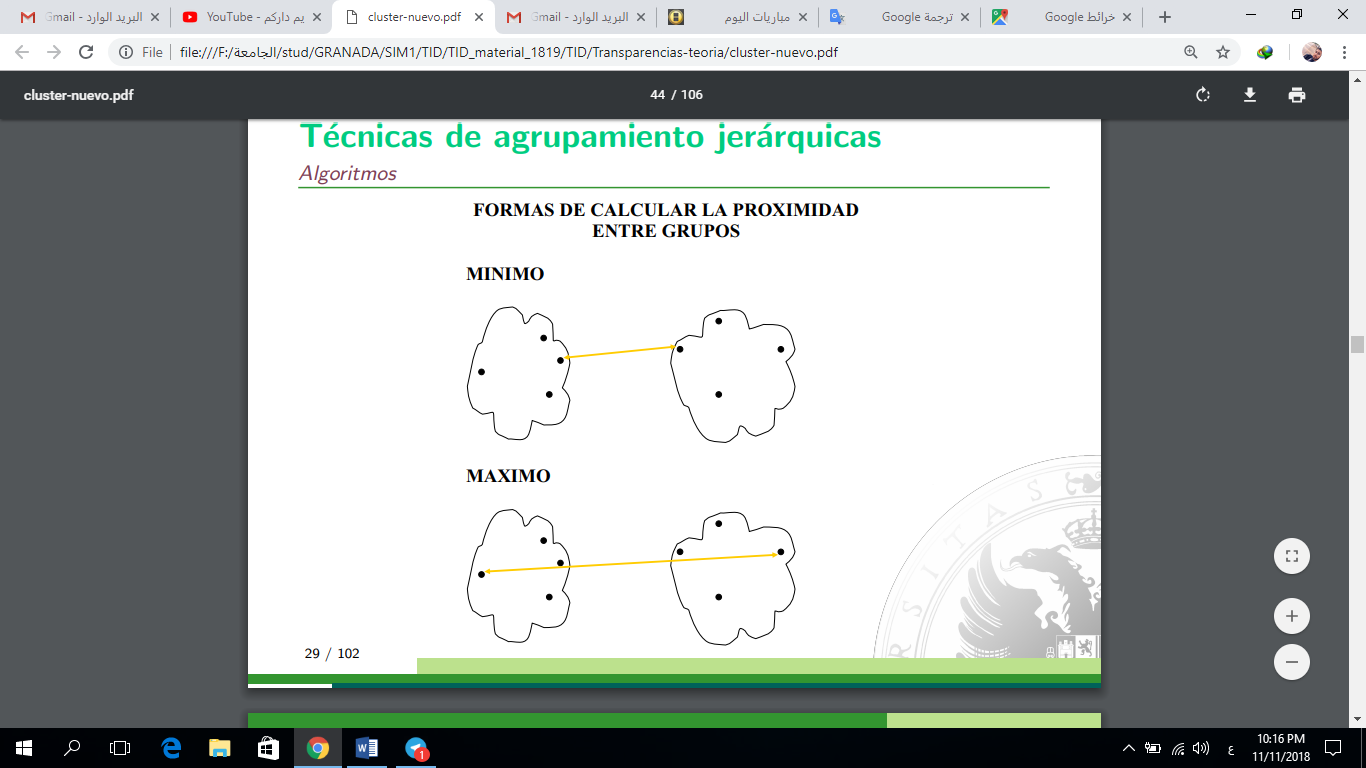
completely connected (clicks)

Hierarchical grouping techniques

Algorithms

WAYS TO CALCULATE PROXIMITY

BETWEEN GROUPS



Hierarchical clustering techniques

Algorithms

? Algorithm of Jhonson. General form of Lance and William

1. Let m = 0, Dm = D, starting distance matrix,

Cm = {{1}, .., {n}} the initial grouping, L (m) = 0 the level at

which is this grouping.

2. Let R and S be those groups of Cm that have minimal distance:

◦ L (m + 1) = Dm (R, S)

◦ Form a new group K = R ∪ S. Make

Cm + 1 = Cm ∪ (R ∪ S) - R - S and transform the matrix Dm of the

Following way.

• Delete the row and column of S and assign the row and column from R to K.

• For all T belonging to Cm other than K, do:

Dm + 1 (K, T) = (2)

a (R) Dm (R, T) + a (S) Dm (S, T) + bDm (R, S) + c | Dm (R, T) - Dm (S, T) (3)

3. Make m = m + 1

4. If all the items have been joined, otherwise go to 2

Hierarchical clustering techniques:

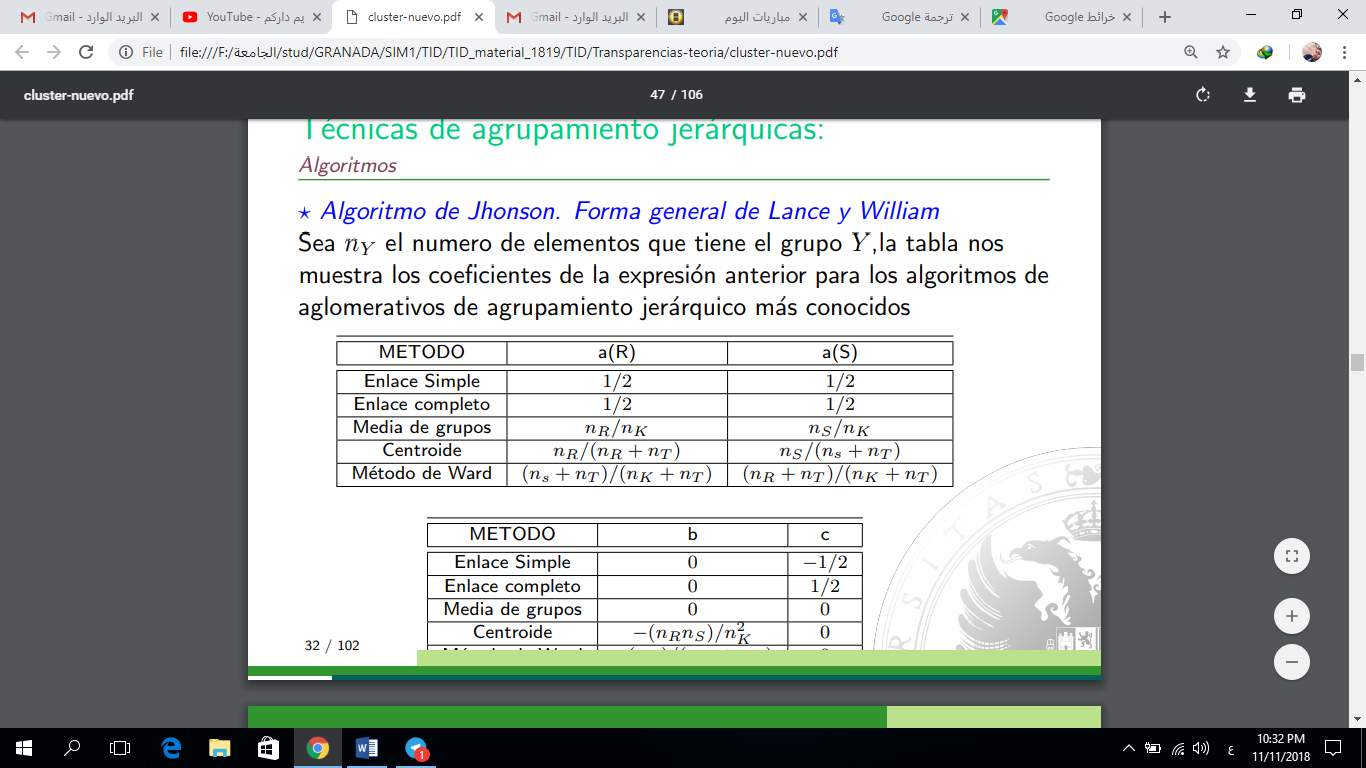
Algorithms

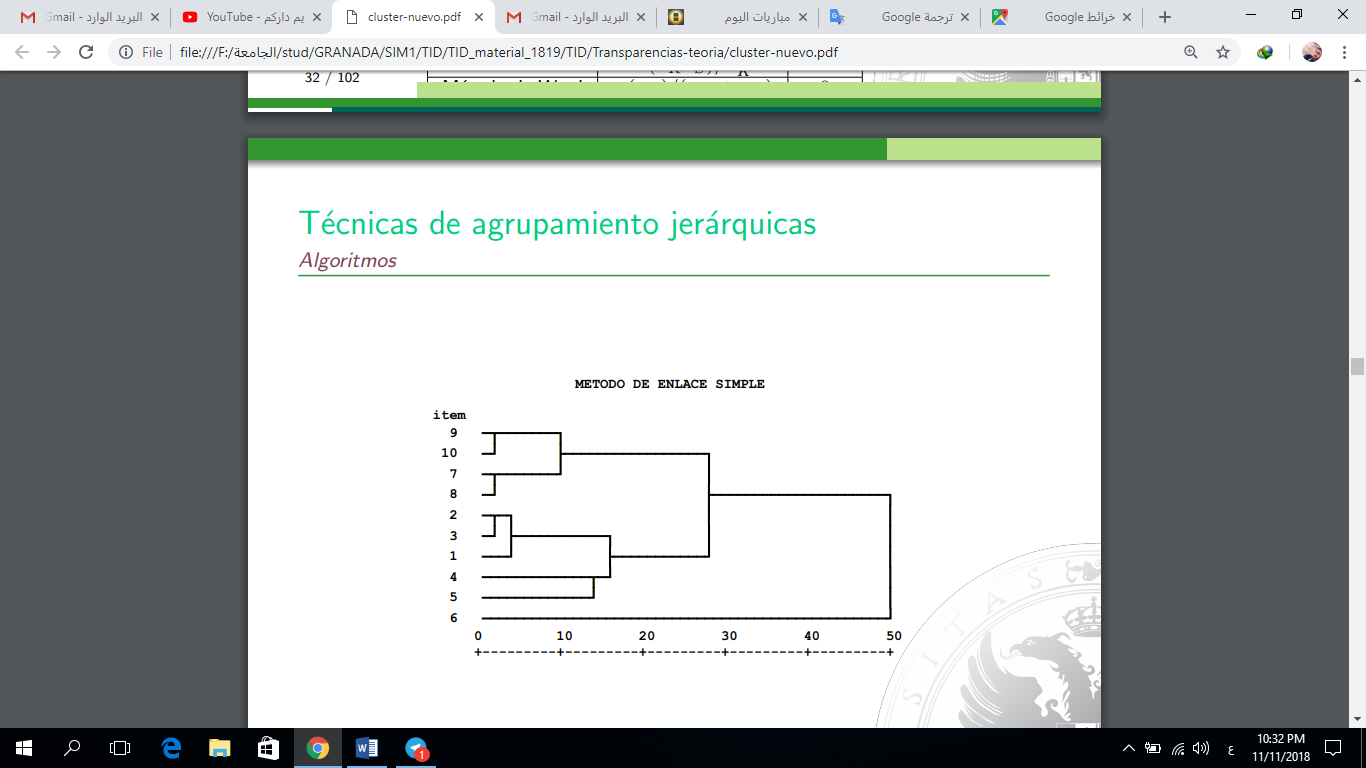
? Algorithm of Jhonson. General form of Lance and William

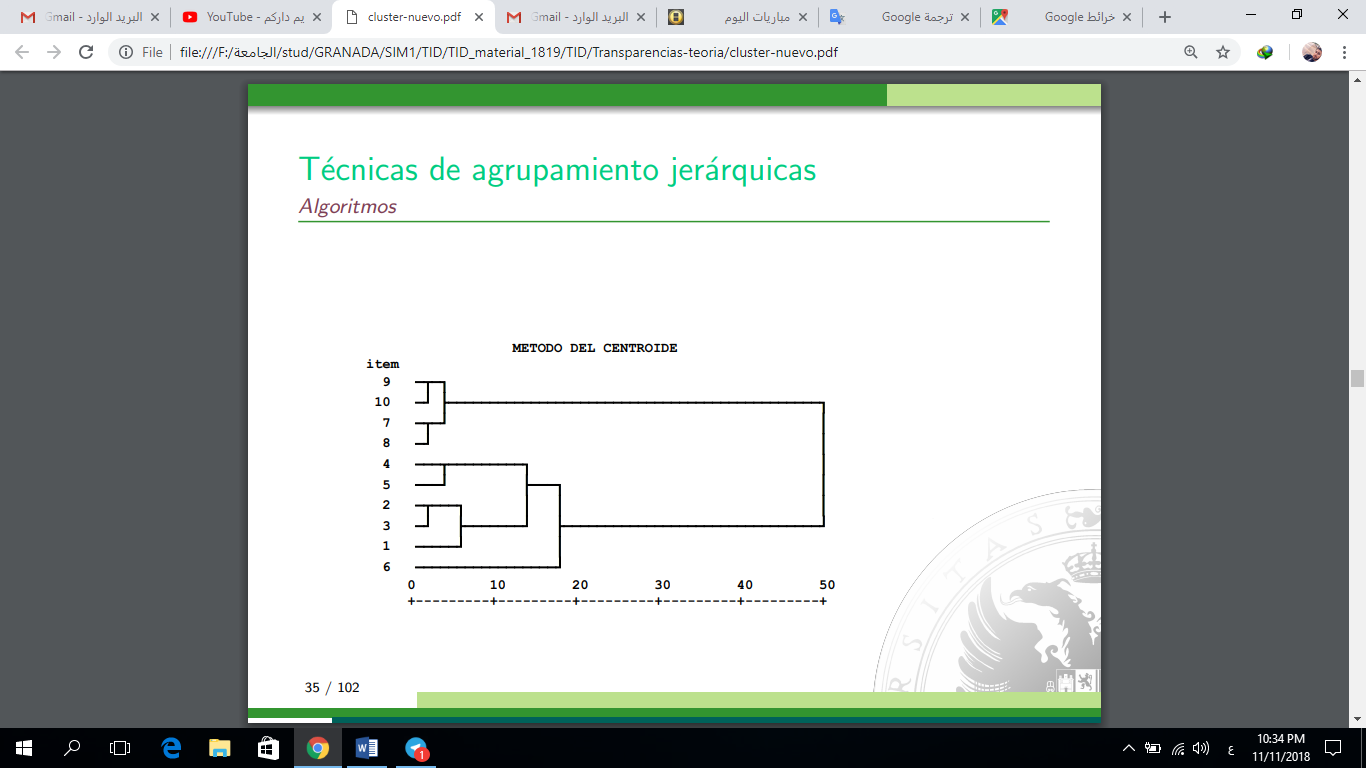
Let n be the number of elements that group Y has, the table

shows the coefficients of the previous expression for the algorithms of

most known hierarchical clustering agglomerates







Partial grouping and hierarchical grouping.

- When the groups to be obtained are disjoint and cover

the whole set of items is said that the grouping

It is partitional.

- When a hierarchy of groupings is obtained

"nested" partitions, in such a way that each group

One level is divided into several at the next level. Is

structure is called jer'arquico grouping and has

a very intuitive graphical representation called

dendrogram