AMD Notes

University of Milan
Master in Data Science and Economics

Algorithms for massive datasets

Andrea Ierardi

Contents

List of Figures iii

1 Mathematical preliminaries 1
1.1 Multivariate calculus . . . 3
1.1.1 Multi-variate functions . . . 5
1.1.2 Multi-variable integrals . . . 8

2 Link Analysis 10
2.1 Problems : Deadends and Spidertraps . . . 16
2.1.1 Deadends . . . 16
2.1.2 Spidertraps . . . 18
2.1.3 Spam Farm . . . 20

3 HDFS and Map-Reduce 22
3.1 Distributed File System (DFS) . . . 22
3.2 Map Reduce . . . 24
3.2.1 MapReduce example . . . 25
3.2.2 Use MapReduce to do the same operation of PageRank . . . 26

4 Spark 27
4.1 Spark . . . 27

5 Similar Items 29
5.1 k-gram. . . . 29
5.1.1 Min-Hashing . . . 31
5.1.2 Locally Sensitive Hashing . . . 33
5.2 Similar Items part.2 . . . 35

6 MarketBasket Analysis 37
6.1 Frequent ItemSets/ Market-Basket Analysis . . . 37
6.1.1 Apriori Algorithm . . . 39
6.2 Market Basket Analysis part. 2 . . . 41
6.2.1 PCY-Variant . . . 41
6.2.2 SON algorithm . . . 44
6.2.3 Toivonen Algorithm . . . 45

7 Recommendation Systems 46
7.1 Content-Based recommendation . . . 47
7.2 Collaborative Filtering . . . 50

References 51

List of Figures

1.1 Example of derivative with 1 var . . . 2
1.2 Gradient computing . . . 2
1.3 Example of gradient 3D . . . 3
1.4 Second-order derivative . . . 4
1.5 Hessian Matrix . . . 4
1.6 . . . 5
1.7 . . . 8

2.1 Example of webpages relationship . . . 10
2.2 Connection Matrix . . . 11
2.3 Structure of WWW . . . 15
2.4 Example of deadend . . . 16
2.5 Pruned graph . . . 17
2.6 Reconstructed graph . . . 17
2.7 Spidertrap . . . 18
2.8 Example of Spamfarm . . . 19
2.9 Spamfarm . . . 20

3.1 Distribute File System . . . 23
3.2 Map Reduce operations . . . 24
3.3 Chain of MapReduce operations . . . 25
3.4 vector v cannot fits the RAM . . . 26

4.1 Spark logic . . . 28
4.2 Spark KeyReduce . . . 28

5.1 Intersection . . . 30
5.2 Min-Hashing . . . 31
5.3 Hash Results . . . 31
5.4 Hash computation . . . 32
5.5 Locally Sensitive Hashing . . . 33
5.6 Plot distribution . . . 34
5.7 Properties distance-probability . . . 35

6.1 HashMap . . . 38
6.2 Apriori Algorithms . . . 39
6.3 Apriori pipeline . . . 40
6.4 PCY-variant . . . 41
6.5 PCY-variant adding intermediate scans . . . 42
6.6 PCY-variant adding intermediate scans 2 . . . 42

7.1 Long Tail phenomenon . . . 46
7.2 Utility matrix for a specific User with boolean entries . . . 48
7.3 Utility Matrix without Boolean entries . . . 48
7.4 Cosine distance . . . 49
7.5 Movies Vectors . . . 49
7.6 UV Decomposition . . . 50
7.7 Initial values of matrix U and V . . . 50
7.8 Perturbation of matrix U . . . 51

1

Mathematical preliminaries

LECTURE 25-01-2021

Multivariable and multivariate function are two independent topics that could be mixed together.
Multivariable functions, derivative extends to this kind of function but in a sort of component wise application.
Standard concept from calculus might not of direct application when dealing with multivariable functions.
If i have

$$f(x,y) = \frac{x \cdot y}{x^2 + y^2}$$

I want to compute a limiting function

$$\lim_{(x,y)\to(0,0)}$$

Limiting values are a pair of value since i have two arguments. This kind of limit depends on the way I am approaching the origin. In a standard case we can approaching limit from the left or the right, here I can approach it in many ways. I want to compute this limit approaching linearly to the origin.

$$f(x,y) = \lim_{(x,mx)\to(0,0)}$$

A segment of a line is passing to the origin. So

$$f(x,mx) = \lim_{x\to 0} \frac{m \cdot x^2}{x^2 + m^2 \cdot x^2} = \lim_{x\to 0} \frac{m \cdot \cancel{x}^2}{\cancel{x}^2 + m^2 \cdot \cancel{x}^2} = \frac{m}{1+m^2}$$

I have not a unique answer computing this limit. Use tools and see how they extend to functions having two or more variables. First one is derivatives.
Let's start with partial derivatives.

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$

We fix every argument of the function except for one of them. We obtain a regular function of 1 parameter and we can compute its derivative as usual. The partial derivative of f with reference to its k variable:

$$\frac{\delta f}{\delta x_k} \quad = \quad \lim_{h\to 0} \frac{f(x_1, ..., x_{n-1}, x_k + h, x_{k+1}, ..., x_n)}{h}$$

I can compute as many partial derivative as argument here (n) and I can arrange them one over the other on a vector.

$$[v_i]_n$$

This means that I am considering a vector of n components and generic form depends on a index variable i. When i will write this I will mean this is a column vector. This notation easily extend to a matrix.

$$[a_{ij}]_{n\times m}$$

so returning to the example

$$\left[\frac{\delta f}{\delta x_i}\right]_n = \nabla f$$

This is called the gradient of the function f and we indicate it with ∇f (NABLA). I will read it as the gradient of f. When I think as a partial derivative is it self a function since I have the dependency of all its argument. So most of the time I will have to specify the argument of the gradient:

$$\nabla f(x_1, ..., x_n)$$

As a gradient is a vector, a vector correspond to a direction in the space. This is easy to see that gradient correspond to the direction of the highest increase of my function. We can see it easily using a degenerate case of a function of just one variable.

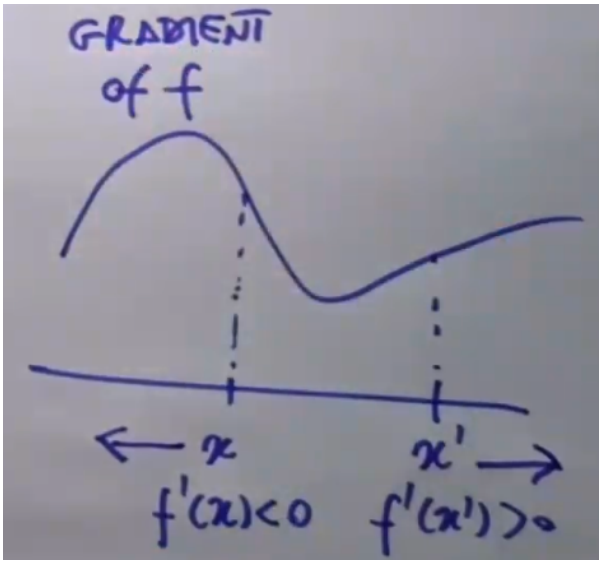


Figure 1.1: Example of derivative with 1 var

Consider point x, we can see function is decreasing and the derivative would be negative: $f'(x) < 0$. Which direction this derivative is describing? When I have something negative i move to the left, so this negative derivative describes the direction going to the left. But going to the left, my function is increasing.
Using x' where $f'(x') > 0$ function is increasing and derivative is positive.
We can see it operationally using more than one argument.

```
In [ ]: gradient(z, 1, 8)

Out[ ]: array([ 0.47182002, -0.0066654 ])
```

and see that it approximately amounts to the vector $[0.472, -0.007]^\top$. We can now check that the gradient identifies the direction of maximal ascent: besides containing a better implementation of `grad`, the following cell generates an interactive graph containing the contour plot of the function g: clicking on any point (x, y) will superpose the vector corresponding to $\nabla f(x, y)$ to the contours, showing that this vector always points towards the nearest local maximum of g.

Figure 1.2: Gradient computing

In the correspondence of a local optima, the gradient nullifies.
We may have mixed second order derivative and if the two var are equal we speak about pure derivative and we use this shorter notation

$$\delta^2 f / \delta x_i^2$$

Note that a second-order derivative might involve either two different variables or a single variable upon which the function is derivated twice. These two cases are referred to speaking of mixed and

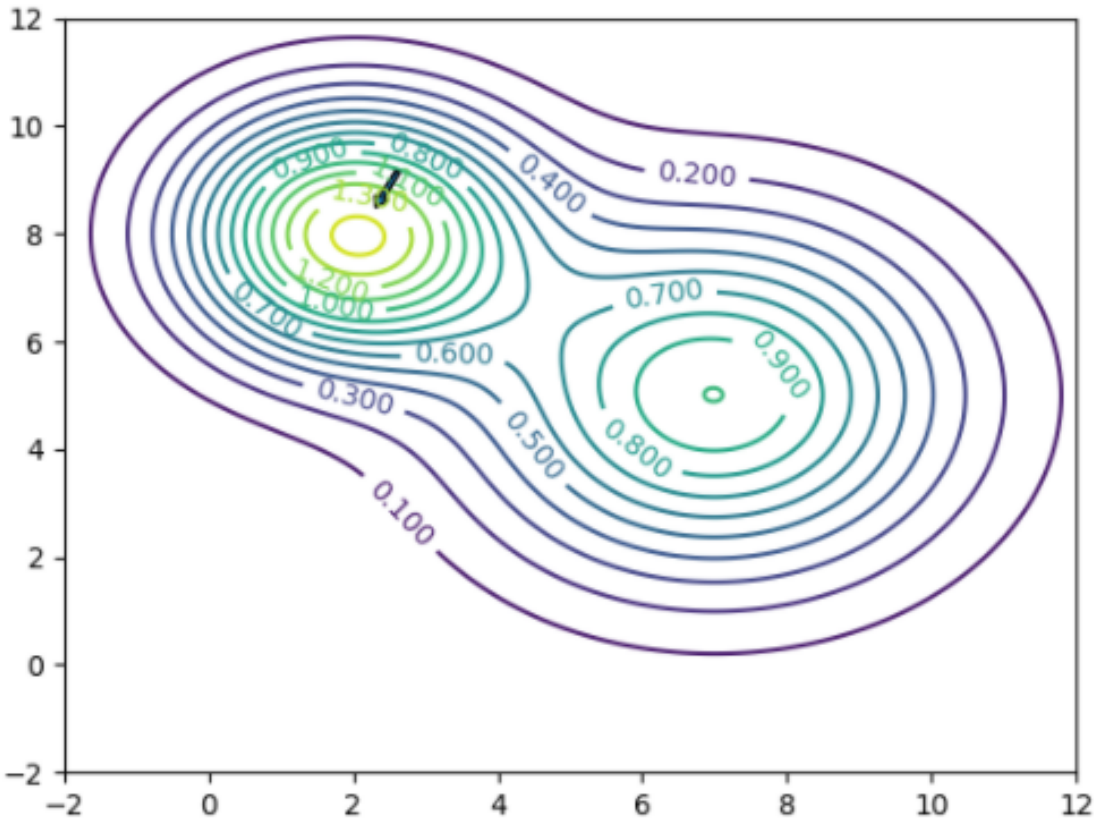


Figure 1.3: Example of gradient 3D

pure derivative, respectively.
If I arrange the second-order partial derivatives in a Matrix I obtain a Hessian Matrix. The determinant of that matrix, together with the sum of trace of that matrix is called Hessian and Laplacian quantities . Hessian and Laplacian quantity help me in finding out if a point that nullifies the gradient is actually a local or global minum of a function.

1.1. Multivariate calculus

Extension of derivatives and integration to multivariate calculus. Function with more than 1 argument. I have a matrix A whose general element will be denoted by a_ij having m rows and n column

$$A = [a_{ik}]_{m \times n}$$

$$B = [b_{kj}]_{n \times p}$$

What happens if I compute the product of A · B?

$$C = A \cdot B \qquad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

What happend if I compute the traspose of C?

$$D = C^T \qquad d_{ij} = c_{ji} = \sum_k a_{jk} b_{ki} = \sum_k a^T_{kj} b^T_{ik}$$

Swapping the two factor I have:

$$\sum_k b^T_{ik} a^T_{kj} = (B^T \cdot A^T)_{ij}$$

Definition 1.3 - second-order partial derivative *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a function, and fix two of its arguments, say x_i and x_j. The second-order partial derivative of f w.r.t. x_i and x_j is the function*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x_1, \dots, x_n) = \frac{\partial \frac{\partial f}{\partial x_i}}{\partial x_j}(x_1, \dots, x_n).$$

Note that a second-order derivative might involve either two different variables or a single variable upon which the function is derivated twice. These two cases are referred to speaking of *mixed* and *pure* derivative, respectively. In the latter case, the form $\partial^2 f / \partial x_i^2$ is generally preferred.

Computation of second-order derivatives via sympy simply requires to specify a further argument for `sympy.diff`, corresponding to the second variable of the derivation (thus repeating the variable in case of pure second-order derivatives though a special syntax for this case can be used).

Figure 1.4: Second-order derivative

Definition 1.4 - Hessian matrix *The Hessian matrix of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the $n \times n$ matrix*

$$\mathrm{H}_f(x_1, \dots, x_n) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x_1, \dots, x_n) \right]_{n \times n}.$$

Definition 1.5 - Hessian *The quantity*

$$\det\left(\mathrm{H}_f(x_1, \dots, x_n)\right),$$

that is, the determinant of the Hessian matrix of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, is called Hessian of f.

Definition 1.6 - Laplacian *The quantity*

$$\Delta f(x_1, \dots, x_n) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2},$$

amounting to the trace of the Hessian matrix of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, is called Laplacian of f.

Figure 1.5: Hessian Matrix

So this means that

$$C^T = B^T \cdot A^T$$

This also holds for the invert

$$C = A \cdot B \qquad C^{-1}?B^{-1} \cdot A^{-1}$$

Giving this, the product of C and the inverse of C must give the identity matrix.

$$C \cdot C^{-1} = (A \cdot B) \cdot (B^{-1} \cdot A^{-1})$$

Matrix product is associative so I can swap terms.

$$(B \cdot B^{-1}) \cdot A \cdot A^{-1}$$

In the $()$ I got the Identity matrix since the product of B and its inverse.

$$A \cdot I \cdot A^{-1}$$

A matrix multiplied by the Identity is the matrix itself so:

$$A \cdot I \cdot A^{-1} = A \cdot A^{-1} = I$$

I would have to prove that if I Swap $C \cdot C^{-1}$ I got an identity matrix. So $C^{-1} \cdot C = I$.

It is possible to show that, given a point $(x_1, \dots, x_n)$ such that $\nabla f(x_1, \dots, x_n) = 0$,

- if $\det\left(\mathrm{H}_f(x_1, \dots, x_n)\right) < 0$, $(x_1, \dots, x_n)$ is a saddle point for f;
- if $\det\left(\mathrm{H}_f(x_1, \dots, x_n)\right) > 0$ and $\Delta f(x_1, \dots, x_n) > 0$, $(x_1, \dots, x_n)$ is a (either local or global) minimum of f;
- if $\det\left(\mathrm{H}_f(x_1, \dots, x_n)\right) > 0$ and $\Delta f(x_1, \dots, x_n) < 0$, $(x_1, \dots, x_n)$ is a (either local or global) maximum of f.

Figure 1.6

1.1.1. Multi-variate functions

In this case we have a multiplicity in the results. A function that return a vector: $f : \mathbb{R} \longrightarrow \mathbb{R}^d$ I can mix multivariable and multivariate functions:

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

How can we deal with derivative when working with such function?
Let's say I have a vector x that depends on a argument t. I can compute derivative component wise.

$$\frac{d\,\bar{x}(t)}{d\,t} = \left[\frac{d\,x_i(t)}{d\,t}\right]_n \qquad \bar{x}(t) \in \mathbb{R}^n \qquad (1)$$

Jacobian: say that we have a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and we have $\bar{x} = [x_j]_n \qquad \bar{y} = [y_i]_m$
So $\bar{y} = f(x)$ The Jacobian is the natural extension of the concept of derivative.

$$\frac{\delta f(\bar{x})}{\delta \bar{x}} \qquad (2)$$

And it is different from what we wrote in (1). In (1) I only have 1 argument. Having several components in the results lead me to a vector of partial derivatives.
In (2) I have two degrees of freedom, several arguments and several components in the results. So this leads to a matrix whose generic component is the partial derivative of one argument of f. This partial derivates need to be done to one of its n arguments.

$$JACOBIAN_f(\bar{x}) = \frac{\delta f(\bar{x})}{\delta \bar{x}} \quad = \quad \left[\frac{\delta f_i(\bar{x}}{\delta x_j}\right]_{m \times n}$$

For example if I have a function that takes a vector and return the sum of its components:

$$f(\bar{x}) = \sum_{i=1}^{n} x_i$$

In this case the Jacobian will be easy to compute because I only have one results. It is not a multivariate function, it's a function of several arguments which however return a scalar. So I will have a row vector.

$$J_f(\bar{x})$$

Givin the fact that I compute the derivative of x_1 wrt to x_1 I obtain 1.

$$\frac{\delta(x_1 + ... + x_n}{\delta x_1}$$

But then I have to compute the derivative of all the remaining variables w.r.t x_1 but those are different from x_1 so I obtain 0.
This holds for all the remaining variables I get as results the column vector whose components are set to 1:

$$J_f(\bar{x}) = (1, ..., 1)^T$$

Another example: change the role of what is a scalar and what is vectorial. In the previous example I have a function that return a scalar. Let's think about a function of one only argument that return a vector and has x as first argument and return x^2 as a second:

$$f(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

Let's compute the Jacobian:

$$J_f(x) = \begin{pmatrix} 1 \\ 2x \end{pmatrix}$$

Now let's see something a bit more complex.

$$A = [a_{ij}]_{m \times n} \qquad \bar{x} = [x_j]_n \qquad A \perp \bar{x} \qquad (independent)$$

$$\bar{y} = A\,\bar{x}$$

What can I say to the derivative:

$$\frac{\delta \bar{y}}{\delta \bar{x}} = \frac{\delta}{\delta \bar{x}}\, A\,\bar{x}$$

If we were dealing with standard calculus, both A and $\bar{x}$ would be scalars and the results would be A. In our case, the result is quite similar. If I consider one of the component of y

$$y_i = \sum_{t=1}^{n} a_{it} x_t$$

What happens if I compute the derivative:

$$\frac{\delta y_i}{\delta x_j}$$

Among the n possible values of t, I also have exactly in one position the j that i have fixed. Think that j in the derivative is 1. All the element of the sum, expect of the first one will not depend on x_1. So that their derivative will always amount to 0. The only exception is when t is actually equal to j. In that case the derivative will not nullify and will be:

$$\frac{\delta y_i}{\delta x_j} = \frac{\delta}{\delta x_j} a_{ij} x_j = a_{ij}$$

$a_{ij}x_j$ is a scalar so standard rule of calculus applied and in the end is equal to a_{ij}.
So the Jacobian is equal to the matrix of those generic element. I shown that is actually the generic element of derivative matrix is equals to $a_i j$. So I can say that:

$$\frac{\delta \bar{y}}{\delta \bar{x}} = \frac{\delta}{\delta \bar{x}}\, A\,\bar{x} = \mathbf{A}$$

In this case we have the equivalent of the result we saw in standard calculus. What about if I have the same information of before but $\bar{x}$ depends on another vector $\bar{z}$. I want to compute the derivative of y over z.

$$\frac{\delta \bar{y}}{\delta \bar{z}}$$

Where $\bar{y}$ (as before) the product of A and $\bar{x}$. So:

$$\frac{\delta \bar{y}}{\delta \bar{z}} = \frac{\delta}{\delta \bar{z}}\, A\,\bar{x}$$

Recal that a generic element of y like y_i is equal to the sum on k of the $a_{ik}x_k$

$$y_i = \sum_{k} a_{ik} x_k$$

So when I compute the derivative, I can distribute the derivative of the y_i sum:

$$\frac{\delta y_i}{\delta z_i} = \sum_k a_{ik} \frac{\delta x_k}{\delta z_j} \qquad (3)$$

If I want to compute the product of A and the vector which I obtain computing derivative of x over z, that will be a matrix product. Computing the generic element:

$$\left(A \, \frac{\delta \bar{x}}{\delta \bar{z}} \right)_{ij} = \sum_k a_{ik} \left(\frac{\delta \bar{x}}{\delta \bar{z}} \right)_{kj}$$

So a_{ik} is the same we have here (3) and the kj element of $\left(\frac{\delta \bar{x}}{\delta \bar{z}}\right)_{kj}$ is equal to this derivative $\frac{\delta x_k}{\delta z_j}$.
So at the end

$$\frac{\delta \bar{y}}{\delta \bar{z}} = \frac{\delta}{\delta \bar{z}} \, A \, \bar{x} = A \frac{\delta \bar{x}}{\delta \bar{z}}$$

We have found again what we expect from standard calculus. A do not depend on z, I can factor it out when I compute the derivative with reference to z with the product of A and x.
Things are not always so easy to do. If I have matrix A and vector $\bar{x}, \bar{y}$ i can call

$$\alpha = \bar{y}^T \, A \, \bar{x}$$

I obtain a scalar and if A is independent of x, y what can I say about derivative of α w.r.t. x?

$$A \perp \bar{x}, \bar{y}$$

$$\frac{\delta \alpha}{\delta \bar{x}} = \frac{\delta}{\delta \bar{x}} \, \bar{y}^T \, A \, \bar{x}$$

If I call $\bar{w}$ the product $\bar{y}^T \cdot A$. We can see that $\bar{w}$ is not dependent on x and applying what we just shown we can see that in the end we obtain $\bar{w}$:

$$\frac{\delta \alpha}{\delta \bar{x}} = \frac{\delta \alpha}{\delta \bar{x}} = \frac{\delta}{\delta \bar{x}} \cdot \bar{y}^T \cdot A \cdot \bar{x} = \bar{w}$$

I get derivative of x with itself which become the identity matrix. Recall that $\bar{w} = \bar{y}^T$ so :

$$\frac{\delta}{\delta \bar{x}} \cdot \bar{y}^T \cdot A \cdot \bar{x} = \bar{w} = \bar{y}^T \cdot A$$

What about $\frac{\delta \alpha}{\delta \bar{y}}$?
Remember that α is a scalar I can transpose it without changing anything.

$$\frac{\delta \alpha}{\delta \bar{y}} = \frac{\delta}{\delta \bar{y}} \alpha^T$$

Remember when I transpose the result of a product I have to consider all the factors in inverse term.

$$\frac{\delta \alpha}{\delta \bar{y}} = \frac{\delta}{\delta \bar{y}} \alpha^T = \frac{\delta}{\delta \bar{y}} \bar{x}^T \cdot A^T \cdot \bar{y}$$

We can applied previous results and then we get:

$$\frac{\delta}{\delta \bar{y}} \bar{x}^T \cdot A^T \cdot \bar{y} \quad = \quad \bar{x}^T \cdot A$$

So derivative wrt x I obtain the left part of $\alpha = \bar{y}^T \cdot A \cdot \bar{x}$. While derivative wrt y^T I obtain more or less what I expected (right part) but I need to exchange the order and to transpose A otherwise I would not have the compatibility between dimensions of the objects.
What happens when x and y are equal?

$$\alpha = \bar{x} \cdot A \cdot \bar{x}$$

It is like multiplying something for the square of a variable. x is not more a scalar but a vector.

$$\frac{\delta\alpha}{\delta\bar{x}} = \frac{\delta}{\delta\bar{x}} \sum_i \sum_j x_i a_{ij} x_j$$

Now, compute derivative for a specific component k. I can bring derivative inside sum and I can factor out a.

$$\frac{\delta\alpha}{\delta\bar{x}_k} = \frac{\delta}{\delta\bar{x}_k} \sum_i \sum_j x_i a_{ij} x_j = \sum_i \sum_j a_{ij} \frac{\delta}{\delta x_k} x_i x_j = \sum_i \sum_j a_{ij} \left(x_i \frac{\delta x_j}{\delta x_k} + x_j \frac{\delta x_i}{\delta x_k} \right)$$

It will nullify every time apart when $k = j$ and I get:

$$= \sum_i a_{ik} x_k + \sum_j a_{kj} x_k$$

$\sum_i a_{ik} x_k$ is the k component of the product between A and the vector x. Second sum is the equivalent of the transpose of A.

$$= \sum_i a_{ik} x_k + \sum_j a_{kj} x_k = (A \cdot \bar{x})_k + \left(A^T \cdot \bar{x}\right)_k$$

So I can say that:

$$\frac{\delta \bar{x}^T \cdot A\bar{x}}{\delta\bar{x}} = \bar{x}^T \cdot \left(A + A^T\right)$$

If A is symmetrical, then it is equal to its transpose and it that case I get $2\bar{x}^T A$.

Proposition 2.8

Given two vectors $\mathbf{x} = [x_i]_n$ *and* $\mathbf{y} = [y_i]_n$*, both dependent on another vector* $\mathbf{z}$*, and defined* $\alpha = \mathbf{y}^\top \mathbf{x}$*,*

$$\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\top \frac{\partial \mathbf{x}}{\partial \mathbf{z}}$$

Proof

By definition $\alpha = \sum_{i=1}^{n} x_i y_i$*, thus for each* k

$$\frac{\partial \alpha}{\partial z_k} = \sum_{i=1}^{n} \left(x_i \frac{\partial y_i}{\partial z_k} + y_i \frac{\partial x_i}{\partial z_k} \right) = \sum_{i=1}^{n} x_i \frac{\partial y_i}{\partial z_k} + \sum_{i=1}^{n} y_i \frac{\partial x_i}{\partial z_k} = \mathbf{x}^\top \frac{\partial \mathbf{y}}{\partial z_k} + \mathbf{y}^\top \frac{\partial \mathbf{x}}{\partial z_k}.$$

Therefore $\frac{\partial \alpha}{\partial \mathbf{z}} = \mathbf{x}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^\top \frac{\partial \mathbf{x}}{\partial \mathbf{z}}$.

Figure 1.7

1.1.2. Multi-variable integrals

The concept of integration à la Riemann naturally extends to functions having the form $f : \mathbb{R}^n \mapsto \mathbb{R}$. Indeed, we can consider the volume of the multidimensional region in $\mathbb{R}^{n+1}$ lying between the graph of f and the hyperplane $x_{n+1} = 0$.

Such multi-variable integrals are denoted using several integration symbols, one per involved variable. When integration is done over a hyper-rectangle, the extremes of each dimension is shown in the lower and upper part of each integral sign, as usual, otherwise the integration region is specified under the sequence of integral signs.

The most simple kind of multi-variable integration occurs when f is separable, that is it can be expressed as the product of n factors, each depending only on one of the considered variables. In this case, the overall process boils down to the product of n standard integrals.

In more complex cases, such as the one which will study in the section on multivariate normal distributions, a variable substitution is often used, e.g. in order to reformulate integration using polar coordinates.

More formally, let $\varphi : \mathbb{R}^n \mapsto \mathbb{R}^n$ denote the transformation to be applied, and denote with $x_1, \ldots, x_n$ the original variables and with $u_1, \ldots, u_n$ the transformed ones (that is, $\varphi(x_1, \ldots, x_n) = (u_1, \ldots, u_n)^\top$). If φ is injective and continuously differentiable, it can be shown that

$$\mathsf{d}x_1 \ldots \mathsf{d}x_n = |\det(\mathsf{J}_\varphi)| \mathsf{d}u_1 \ldots \mathsf{d}u_n$$

Therefore, denoting by $g(u_1, \ldots, u_n)$ the equivalent of $f(x_1, \ldots, x_n)$ after the application of φ, we have

$$\int \cdots \int f(x_1, \ldots, x_n) \mathsf{d}x_1 \ldots \mathsf{d}x_n = \int \cdots \int g(u_1, \ldots, u_n) |\det(\mathsf{J}_\varphi)| \mathsf{d}u_1 \ldots \mathsf{d}u_n.$$

2

Link Analysis

LECTURE 01-02-2021

Link Analysis when speaking about Link Analysis we are studying the relationship. This is intended in the mathematical sense: Cartesian product between two different sets. Classical example is the so called Page Rank Algorithm. It can be used to study generic relationship. First field for Pagerank was Webpages.
Link analysis refers to relationships.
Algorithm $\longrightarrow$ Page Rank $\longrightarrow$ Among web pages find the most relevant.

An example of net using a graph.

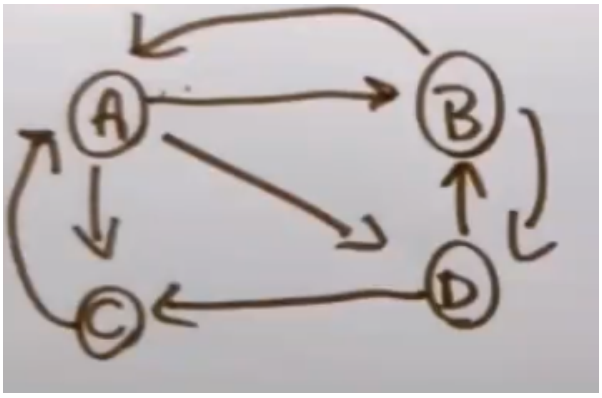


Figure 2.1: Example of webpages relationship

- node is a webpage
- edge is a hyperlink (directed)

Web is hudge! The stimated number of web-pages is $n \approx 10^9$
Since now, we don't define the realm.
We deal with big data when we have an amount of data which is different with small data. Big data from medicine is different from economics big data.
In computer science we speak about big data when we cannot use a single computer in order to process data or even to store data. (10 Tera of data). Here is in main memory problems.
If we want to encode the graph we need a matrix that record what happen to each node in the graph. It will be a square matrix:

$$M = [m_{ij}]_{n \times m}$$

which is of the order of 10^{18}.
With Pagerank I want to find ranking between those pages.

Here value are floating point values. We need a big memory to store this amount of data. We will develop some method to deal with this pages. We will going to a distribute setting:

- Distributed storage: data divided in chunks
- Distributed computation: I can use CPU of that computer to parallelize the computations.

The important of a web page is related to the internal of a web page. There were before Google other search engine but they were related to the searching of a keyword in a web-page. I write Paris and engine will search Paris. I could figure out which are the most written words in the page and I get some how to cheat.
Founder of Google found an Algorithm that doesn't rely on the content of each page but on the hyper-links from one page to another one. A web-page is important is link with more important pages. This is not a acceptable definition since it will lead to infinite recursion. I will have to check the importance of the neighbours of a node, then I check the importance of that node and this is a recursion that will never stops.

So how do we get to an acceptable definition of **Importance**?
The basic idea is thinking of random process : random surfing. We are dealing with a stochastic process. If i have the probability distribution i will find the importance of that web-page (higher is better).
We will use **Connection Matrix**: We don't have any loop. This process is stochastic in any column.

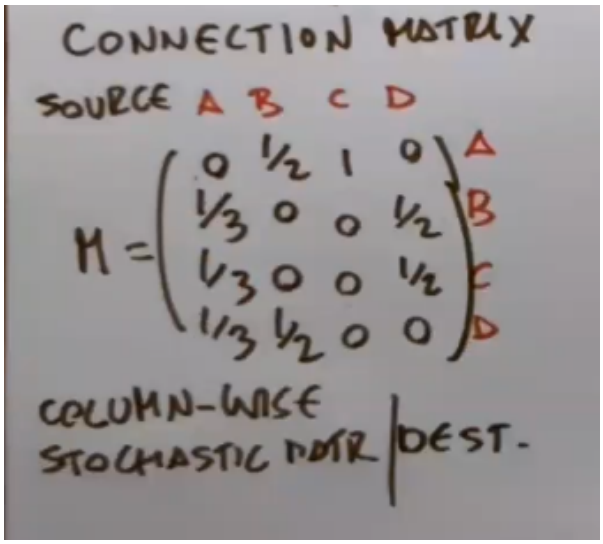


Figure 2.2: Connection Matrix

The fractions will sum up to one in a column.

$$m_{ij} = P(i \to i)$$

These will be the probability of transition from i to j

$$v_j(t) = P(\textit{sitting in } i \textit{ at time } t)$$

The probability of sitting in the node j at time t.
We know use the total probability theorem. Sum of the probabilities having travel to j

$$v_j(t+1) = [P(i \to j|\ \textit{sitting in } i \textit{ at } t) \cdot P(\textit{sitting in } i \textit{ at } t)] =$$

Where $P(\textit{sitting in } i \textit{ at } t) = v_i(t)$
What about the conditional probability? We hypothesis there is no relation of addition of new links. Probability is independent on time

$$= \sum_i P(i \to j) v_i(t) = \sum_i m_{ji} v_i(t)$$

This is the definition of matrix vector product:

$$\bar{v}(t+1) = M\bar{v}(t)$$

If I start from initial value v_0 and iterating n times, do i get something? Yes.
The first thing is a numerical method called the **Power Method** that allow to find a eigen vector of a square matrix.

Recall that

$$Ax = \lambda x$$

Where x is the **eigenvector** and λ is the **eigenvalue**
We call it **eigenpair**.

We have a matrix A:

$$A = [a_i j]_{n*m} \qquad (\lambda_1, x_1), ..., (\lambda_n, x_n) \qquad \lambda_1 \geq \lambda_i \quad \forall i$$

Standard results tells us that the set of eigenvector is a set of an orthogonal base.

$$x_1, ..., x_n \quad \textit{is a basis}$$

If i take any vector I can write as the sum of the vectors in the basis.

$$v_0 = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$$

$$v_1 = Av_0 = A(\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n) = = \alpha_1 A x_1 + \alpha_2 A x_2 + ... + \alpha_n A x_n = \alpha_1 \lambda_1 x_1 + \alpha_2 \lambda_2 x_2 + ... + \alpha_n \lambda_n x_n$$

$$v_2 = Av_1 = A(\alpha_1 \lambda_1 x_1 + ...) = \alpha_1 \lambda_1^2 x_1 + \alpha_2 \lambda_2^2 x_2 + ... + \alpha_n \lambda_n^2 x_n)$$

$$v_k = \alpha_1 \lambda_1^k x_1 + \alpha_2 \lambda_2^k x_2 + ... + \alpha_n \lambda_n^k x_n)$$

Factoring λ:

$$\lambda_1^k (\alpha_1 x_1 + \alpha_2 (\frac{\lambda_2}{\lambda_1})^2 + ... + \alpha_n (\frac{\lambda_n}{\lambda_1})^k x_n)$$

$$\lim_{k \Rightarrow +\infty} v_k = \lim_{k \Rightarrow +\infty} \lambda_1^k \alpha_1 x_1$$

As k becomes big as all v_k became close to right of this last equation.
This mean that Power Method works.

$$det(A) = \sum_i a_{ij} c_{ij} = \sum_j a_{ij} c_{ij}$$

$$det(A^T) = \sum_i a_{ij}^T c_{ij}^T = \sum_i a_{ji} c_{ji} = \sum_j a_{ij} cij = det(A)$$

I swapped i and j we get that:

1. $det(A^T) = det(A)$

Another process:

$$det(A - \lambda I) = 0 \quad \Leftrightarrow \quad \text{det}(A - \lambda I)^T = 0 \quad \Leftrightarrow \quad \text{det}(A^T - \lambda I) = 0$$

where I is the identity matrix.

1. $det(A^T) = det(A)$)
2. A and A^T have the same eigenvalues (λ

If A is a row-stochastic if i sum the elements in a fixed row, i have to obtain 1. $\longrightarrow \sum_j a_{ij} = 1$

$$A\bar{1} = \left[\sum_j a_{ij}(1)_j\right]_n = \left[\sum_j a_{ij}\right]_n = [1]_n = \bar{1}$$

scalar 1 and vector 1 is an eigen pair of A

$$A\bar{1} = 1 * \bar{1} \quad (1, \bar{1}) \textit{ is eigenpair of A}$$

1. $det(A^T) = det(A)$
2. A and A^T have the same eigenvalues (λ)
3. $(1, \bar{1})$ is eigen-pair of row-wise stochastic matrices
4. 1 is eigenvalue of col-wise stochastic matrices

If A is row-stochastic, also A^k is row-stochastic
Proof by Induction
Base: $k = 1$ Obvious Step: $A^k isr.s \rightarrow A^{k+1}$ is r.s

$$A^{k+1} = A^k \cdot A$$
$$\alpha_{ij}^{k+1} = \sum_s a_{is}^k \cdot \alpha_{sj}$$
$$\sum_j a_{ij}^{k+1} = \sum_j \sum_s \alpha_{is}^k \cdot \alpha_{sj} = \sum_s \alpha_{is} \sum_j \alpha_{sj} =$$

where $\sum_j \alpha_{sj} = 1$

$$= \sum_s \alpha_{is}^k = 1$$

Result is 1 because is row-stochastic.

1. $det(A^T) = det(A)$
2. A and A^T have the same eigenvalues (λ)
3. $(1, \bar{1})$ is eigen-pair of row-wise stochastic matrices
4. 1 is eigenvalue of col-wise stochastic matrices
5. A is row-wise stochastic, A^k is row-wise stochastic

The first eigenvalue of A col-wise stochastic matrix is 1.
There is not existence of eigenvalue higher than one

Proof by absurd.

$$\exists \lambda > 1 \textit{ is eigenvalue of A}$$

$$\lambda \textit{ is eigenvalue of } A^T$$
$$A^t \bar{v} = \lambda \bar{v} \rightarrow (A^T)^k \bar{v} = \lambda^k \bar{v}$$

$$\sum_j (a^{k^T})_{ij} v_j = \lambda^k v_i \quad v_{max} = \max_i v_i$$

$$\sum_j (a_{ij}^{k^T}) v_{max}$$

Suppose G is high

$$\sum_j (a_{ij}^{k^T}) v_{max} > G$$

Then,

$$\sum_j (a_{ij}^{k^T}) > \frac{G}{v_{max}}$$

a transpose is row-wise, then a transpose to the power of k is row-wise, then

$$\sum_j (a_{ij}^{k^T}) = 1 > \frac{G}{v_{max}}$$

$$1 > \frac{G}{v_{max}}$$

I can select G as big as I want. But it is absurd!
Then $\lambda \leq 1 \rightarrow$ the first eigenvalue is 1. QED.

$$v(1) = Mv(0)$$

$$..$$

$$v(t) = Mv(t-1)$$

$$..$$

$$v = 1 \cdot v = Mv$$

v will be the vector contain the PageRank value of the portion of pages in which I am considering

Markov Chain is the simple stochastic process we can think of.

LECTURE 02-02-2021

Vector changes each time. We need a initial vector v_0

1. $v_0 = [\frac{1}{n}]_n$
2. $t = 0$
3. $v_{t+1} = Mv_t$
4. $t += 1$
5. if $||v_{t+1} - v_t|| > \varepsilon$ goto 3
6. return v_{t+1}

Passing a number of iteration of step 3, the vector will converge. So we compute the norm of the difference and if this value is greater to a certain value we go to 3.
How many iteration i need to perform to get to the threshold? The problem is that M is huge! We will see how to manage to save M.

Do we have knowledge about the structure of WWW ?
It is said that it recall a bow-the. So has this form:

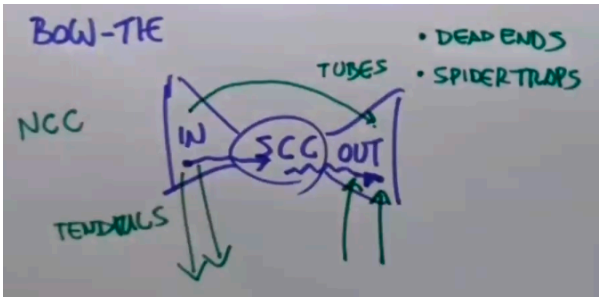


Figure 2.3: Structure of WWW

In the center we have Strongly connected component(SCC). What about In and Out bound components? I can find in IN that has at least one path that connect In with SCC.
The Out, I will find always a path from SCC to Out.
It is possible to reach Out component from the In without passing through the SCC component. These kind of paths are called **Tubes**.
Tendrails are edges that goes from In componenent and they are not goind neither to SCC neither to Out.
Not connected component(NCC): pages that are locally connected.

2.1. Problems : Deadends and Spidertraps

There might be two problems:

- Deadends
- Spydertraps

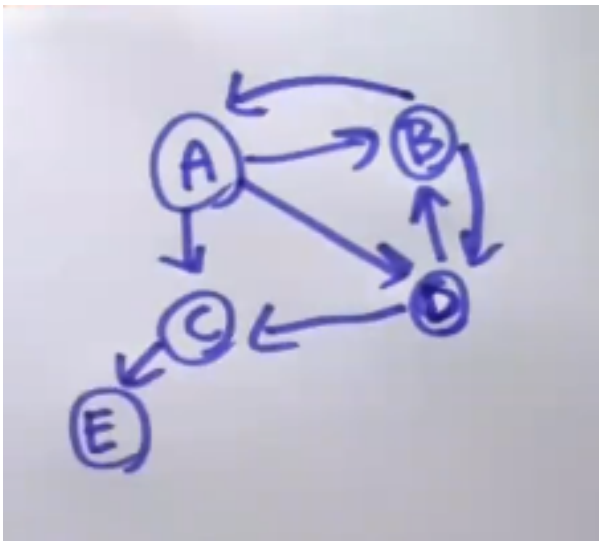


Figure 2.4: Example of deadend

In this examples E is a deadend since it has out-degree equals to 0.

2.1.1. Deadends

What about the column of E? I will have 0 in the column of E. But a **Null column** in a matrix means that the matrix cannot be a column-wise stochastic matrix.

If I would run the Algorithm in this graph anyway? You will notice that as iteration number grows i will have a link in the probability mass. The random surfing will have a huge individual that are distributed across all the nodes.
After a finite number of iteration i will loss a number of individuals. After some iteration the vector v_t will be a vector in which all components will be set to 0.

We have a specific solution: transform the graph into a graph without a deadend. How to do it?

We could simply delete the E node. When i remove E I will have to remove all node to E. Note that the resulting graph will end up in a dead end on C. We could iterate the pruning process and then we get a final graph with A, and D nodes without out degree equal to 0. It may be show that the convergence will be equal to $\frac{2}{9}$ for A, $\frac{4}{9}$ for B and $\frac{3}{9}$ for C.

I can somehow reconstruct the original graph and propagate to the re-added nodes the page-ranks computed. The page rank is evenly distributed to the page rank of the remaining nodes. A has $\frac{2}{9}$ that is evenly distributed to three nodes so : $\frac{2}{9} \cdot \frac{1}{3}$. Only one goes to C.
For B we have $\frac{3}{9} \cdot \frac{1}{2}$. In fact, only 1 of the two out link goes to C.
Summing up the two thing we get that the pagerank of C is:

$$\frac{2}{27} + \frac{3}{18} = \frac{13}{54}$$

For E there is nothing to compute. Since edge from C to E is the only one. Then the pagerank of E is equal to pagerank of C. $\frac{13}{54}$

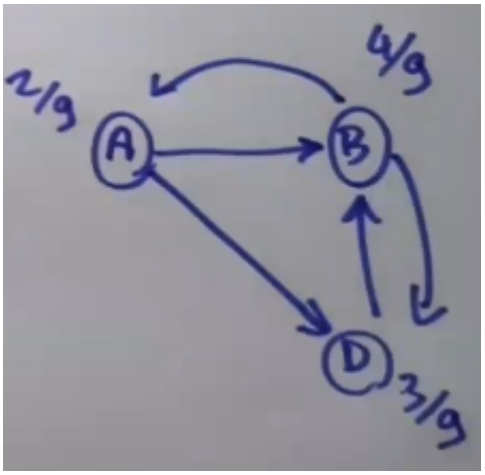


Figure 2.5: Pruned graph

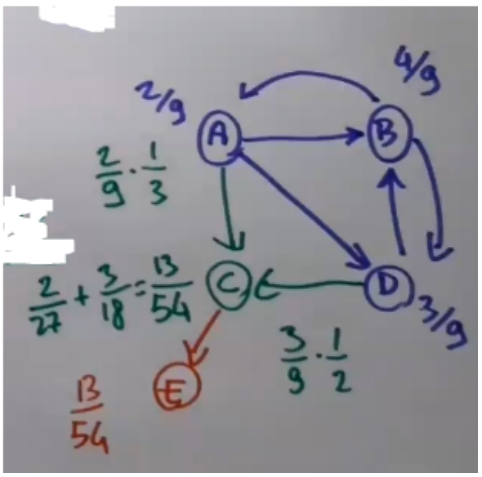


Figure 2.6: Reconstructed graph

But know we do not have a distribution since the sum of the fractions will not sum to 1 but is higher.

2.1.2. Spidertraps

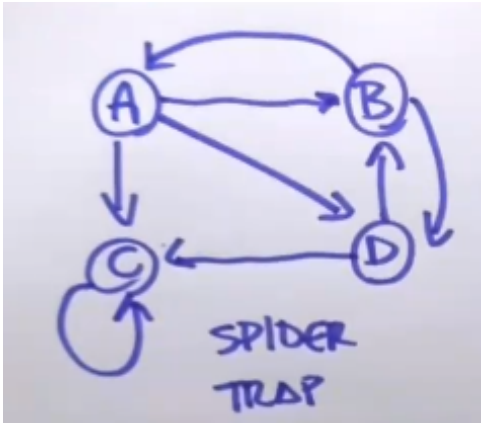


Figure 2.7: Spidertrap

We introduce a self loop that consist of a Spidertrap. Since WWW spider is a historical name from crawling (processing pages automatically). Why it is a problem? Being in this loop, once any of random surfer it caught here it will never exit. This mean that the equivalent of the portion of mass probability will be confined in C. I will have more probability concentrate on C. So after some iteration the vector refers to C will be close to 1 and for the other nodes will be close to 0.
Teleport/Taxation solves this problem. The vector v will not computed only using the matrix but using a parameter β

$$v_{t+1} = \beta M v_t + (1-\beta)\frac{1}{n}\bar{1}$$

where $\beta \in (0,1)$, so is $\beta = P$*(Keep on with random surfing)* β is a number between 0 and 1. It is the probability of keep in on with random surfing.

(1)	P*(sit in* i *at* $t+1$ *| keep random surfing)*$*$	$= Mv_{t+1}$
(2)	P*(Keep random surfing)*$+$	$= \beta$
(3)	$P($*sit in* i *at* $t+1$ *| teleport* $)*$	$= \frac{1}{n}$
(4)	P*(teleport)*	$= (1-\beta)$

where

- $Mv_{t+1} = (1)$
- $\beta = (2)$
- $\frac{1}{n} = (3)$
- $(1-\beta) = (4)$

So at the end this is the first equation of v_{t+1}. Why it should avoid Spidertrap? At each iteration I have the probability of C will be teleport to the other nodes. In this way I will counter balanced because It is likely it will be teleport out of the trap.
Typically $\beta \approx 0.8$, since if it is close to 0 the process will degenerate random process (left part of the vector equation will be negligible).

There is a huge problem: user rely on human language and it is uncertain.
If write Jaguar? An animal, but maybe I am refering to the car or Mac OS.

Another example: we saw her duck
Animal or person ducking. It may refer to past of cutting something in two part.
How can we solve this uncertainty? If we have knowledge of the user. If I know a person is a zoologist is easily gonna search for the animal.

We don't want a unique pagerank or several hyper personalised to the subject but maybe we can use small vector to rank the results. We can rank topics and build a PageRank vector. If i have more info about the user and his prefered topics we can build the vector. I have to decide the topics. There is a pre-existing web page: www.curlie.org.

We have to assign person to topics and I know a vector that is a boolean vector in which is 1 if page is about a topic and 0 if not.

Topic s

$$S_i = \begin{cases} 1 & \textit{if page } i \textit{ is about the topic} \\ 0 & \textit{otherwise} \end{cases}$$

$$v_{t+1} = \beta M v_t + (1-\beta)\frac{1}{|S|}S \qquad |S| = \sum S_i$$

where last part is used for normalising the vector.
In this case I set B and D as pages describing topics.

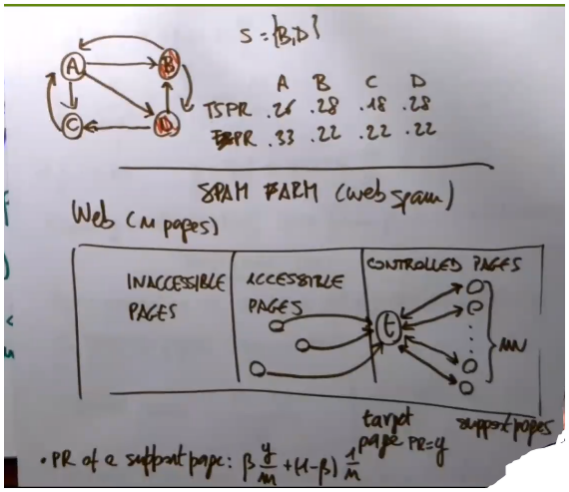


Figure 2.8: Example of Spamfarm

B is 0.28 which higher than the initial pagerank while C has a reduced Pagerank value.

2.1.3. Spam Farm

System to allow users to avoid cheating and boosting the Rank of pages. There are several methods. In this suppose we have n pages in WWW. So any point in the rectangle is a page.

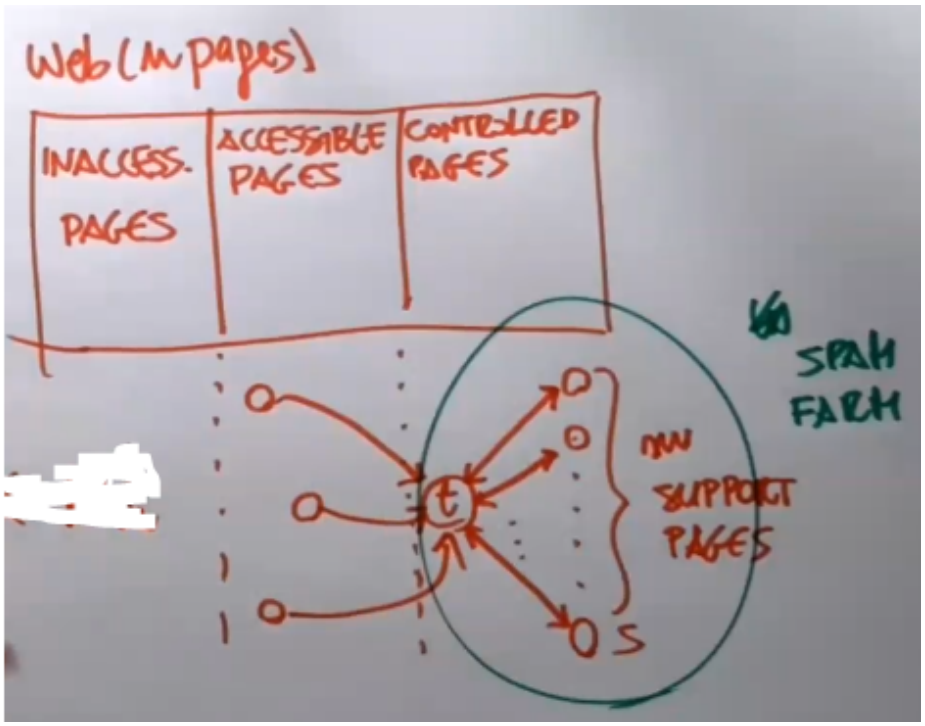


Figure 2.9: Spamfarm

We have three partitions:

- Inaccessible Pages: pages on which webspammer has not full access but can somehow modify the content (adding hyperlink). In this way a spammer can boost a Pagerank referring to this page in the editing of the hyperlinks.
- Accessible pages: pages on which webspammer has not access at all.
- Controlled Pages: pages in which a webspammer has total controls. He is the owner of the pages.

To give effectiveness to a spam farm i have to add links from accessible to my target. If these pages has higher PageRank it can be transferred to t. Why?

Suppose PageRank of t is y
$PR(t) = y$

PageRank of support pages + taxation
$PR(S) = \beta\frac{y}{m} + (1-\beta)\frac{1}{n}$

Adding info about PageRank of t
$PR(t) =$

- x (from accessible pages)
- $\beta(\beta\frac{y}{m} + (1-\beta)\frac{1}{n}$ $\forall$ support pages (in this case we have m support pages)
- $(1-\beta)\frac{1}{n}$ (teleporting) $\rightarrow$ NEGLIGIBLE

So for 3^{rd} I can write:

$$y = x + m\beta(\beta\frac{y}{m} + (1-\beta)\frac{1}{n}) =$$

$$= x + \beta^2 y + \frac{\beta(1-\beta)}{n}\, m$$

$$y(1-\beta^2) = x + \frac{\beta(1-\beta)}{n}\, m$$

$$y(1-\beta^2) = x + \beta(1-\beta)\, \frac{m}{n}$$

$$y = \frac{x}{1-\beta^2} + \frac{\beta}{1+\beta}\, \frac{m}{n}$$

Suppose $\beta = 0.85$ (typical value).

$$y = 3.6x + 0.46\, \frac{m}{n}$$

$$PR(t) = 3.6x + 0.46\, \frac{m}{n}$$

Crawling process might find a farm and not considering it. We can use a variant of the teleport and detect if page is important. We can modify the topic sensitive page rank. We are selecting a set of pages that are trusted. So i can obtain the Trust Rank.

Trust Rank TR

I can compage Page Rank with Trust Rank.

$$\textbf{Spam mass} = \frac{PR - TR}{PR}$$

If this ratio close to 1, the TR is almost 0 compare to PR $\Rightarrow$ it is not an important page.
If this ratio close to 0, the TR is almost comparable to PR $\Rightarrow$ page can be trusted.

3

HDFS and Map-Reduce

LECTURE 08-02-2021

We have used the web as example but PageRank can be used to compute any kind of relation.

3.1. Distributed File System (DFS)

Problem: i cannot use main memory to calculate pagerank because Matrix is too big to be store in RAM and also storage.
The answer is buy bigger storage but this is not a real solution since hardware is expensive. Better buy more computer and distribute them as a cloud. Buy 100 PC in which with hard disk of 1 Tb. We have the power of 100 Tb. I cannot rely on the same architecture we have in the single computer. We putting a another level over my 100 PC. I will have a software layer that will introduce a **Distributed File system (DFS**.
It is an abstraction of several machine in which you operate as a single machine but the content is distributed over several machines.
Concept is not new:

- 1984 the introduction of NFS.
- 2003: basis of Google file system (exactly what we need).
- 2007: Hadoop release by Apache with two components:

  => HDFS: devote to storage
  => MapReduce: process what is in HDFS

Hardware breaks! Using a huge number of computer we will expect failure. There is a big difference between DFS and the standard file system. The standard one you use it for different operations.
Files can be written ones, in Hadoop we can think about appending but not operate in the middle of a file.

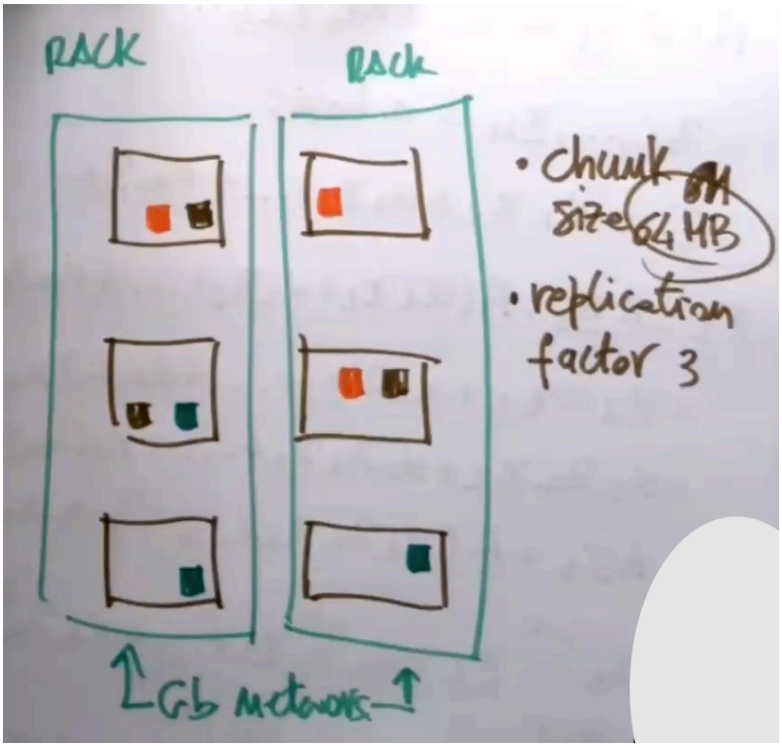


Figure 3.1: Distribute File System

We have rack that contains different chunks. The connection between rack should use a GB network. If i want to dump some content to a file i need to take into account:

- **Chunk size**: file you access are not likely stored entirely in consecutive spaces. The space in Hard Disk is subdivided in stock of different size. When i store file OS break the file in different blocks in different part of memory. We need a meta data to restore the file content and order. We speak about chunk since the size of these blocks is different from the one in a DFS. The default is 64 MB but is a choosable parameter.
- **Replication factor**: i have different chunks in the storage (brown small square). You have to expect hardware failure (PC die). The only way is to performing some kind of replication. I will look to store same data in different chunks. The replication factor tells us how many copies of my file I will have to store in different computers (in this case for example 3).
  I could have also network failure, so that's why 3 replicas: two in the same rack and another one in another rack if we face network failures. We have the concept of distance: distance PC means delay in reaching information.

We have **distributed computation**: CPU of the machines in each racks and i can use for processing.

3.2. Map Reduce

MapReduce is a framework.
Example of documents: your DFS contains different documents. We have divided these files in chunks. In HDFS all the information is organized a sequence of key, value pair:

key-value pair (k, v)

key can be associated to two or more values. We could have series of key value pair.
Each chunk contain a couple (k_1, v_1).

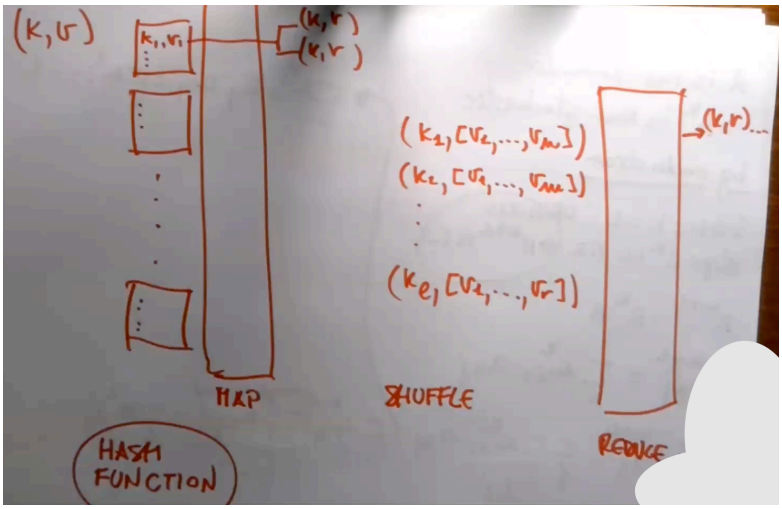


Figure 3.2: Map Reduce operations

- **Map task**: Given a chunk and a computer running in that chunk i will compute a process which is called **Map task**.
  Each map will consider separately every pair of key-value and it will process singularly each key-value pair somehow. The only constrain is that for each value-pair will output 0 or more key-value pair. This is done in parallel in each computer that contains at least one chunk of the file.
- **Shuffling phase**: Once all map task finished the computation there is a **Shuffling phase**. These phase see each result in different computer and will aggregate together the value of a common key and is done for all keys. After this operation it will never happen to have the same key into different of this pairs. To be sure it will be applied an hash function.
- **Hash function**: It take argument to a function an mapped to a bucket. If I consider all possible of each argument, it will approximately uniforming mapped at random to all buckets. I will have to fix number of buckets and the hash function will map keys to computer. So each pair is sent to a computer which is identified by taking key and using it as hash argument. So each pair with same key will be deliver to the same machine. In this way a will have a balanced the numbers of pair to each computer. Now, we need to move information about the network. Once each aggregate pair sit in the same computer another operation will start.
- **Reduce task**: On each of the machine a reduce task will run. It will process the information somehow and will output another key-value pair.
- **Storing**: at each operation results are temporarily stored in a tmp file. Final step all results will be merged in a file stored in the Distributed File system (DFS).

The problem of HW will not only impact in final step or any of these step. I could restart the process but the problem is in one machine. There a master that use a sort of integrity check in the overall DFS. If one machine dies, it will ask to another Machine to process that chunk. The failure of master process will cost to re-executing all the process.

3.2.1. MapReduce example

How DFS will organise the storing of a document?
I could split files row-wise and use each row as a value in key-value pair. Fo example we have a increasing number:

$$(1, \ell_1)$$
$$(2, \ell_1)$$
$$..$$
$$(n, \ell_n)$$

Suppose we have pair in a chunk. Each of the map task will be the same thing: map task split line in each line in each occurrence of the space and we get sequence of words. For each word we get a key-value pair. We have the word as key and the number the occurrence for each word in the line.

$$(i, \ell) \overset{MAP}{\longrightarrow} (w, 1) \quad \forall w \in \ell$$

Now I Will have the shuffling face: for each word I will build a vector of values. As any key value was 1, the vector will contains only ones. If i count the number of ones I will get the frequency of w in the document.

$$(w, \underbrace{[1, ..., 1]}_{n_w}) \overset{REDUCE}{\longrightarrow} (w, n_w)$$

n_w is the number occurrency of w which is the length of the vector of ones.

Map reduce framework is less power full that what we could achieve by standard computational framework.
I can run several MapReduce jobs in chain.

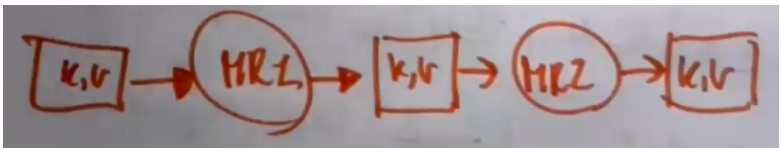


Figure 3.3: Chain of MapReduce operations

3.2.2. Use MapReduce to do the same operation of PageRank

$$A = [\alpha_{ij}]_{n \cdot m}$$
$$v = [v_k]_n$$
$$P = A \cdot v \qquad P_i = \sum_{j=1}^{n} \alpha_{ij} v_j$$

A cannot be stored, but v can or cannot be stored (we could have 8 Gb to store v).
Suppose then we can store v in the main memory.

How to organise the matrix? As a sequence of key-value pair. We distrubited the storage of the matrix as triple (i, j, α_{ij}) for example $(7, 103, 84)$.

$$(i, j, \alpha) \xrightarrow{MAP} (i, \alpha_{ij} \cdot v_j)$$

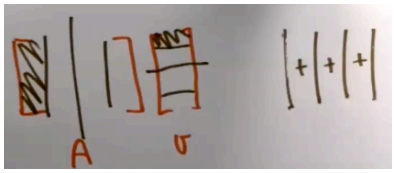
$$(i, [\alpha_{i1} \cdot v_1, \alpha_{i2} \cdot v_2, ..., \alpha_{in} \cdot v_n]) \xrightarrow{REDUCE} (i, S) \; = \; (i, P_i)$$

where $S = \sum_{i=1}^{n} \alpha_{ij} v_j$ which is the product of the component.
Once MapReduce job finished I will store in a file of DFS a element for each of the component of the results. It will be a pair in which first element is the number of the components and second the components.

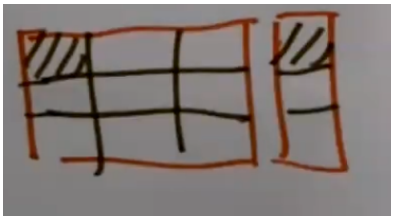
Some possible **Issues**:

- What happens if one of the component $\alpha_{ij} = 0$?
  I will output $(1, 0)$. Zero is the neutral element for sum so this null element will not change the result. If matrix is sparse (as WWW) then, we can exploit it.
- You cannot expect the list ordered, but this is not a problem since I'm considering it as commutative operation. When you look at the file do not expect sorted components.

Hypothesis about v does not fix main memory? If v too big, i might store half of the element of v and the left half of A. I am still able to compute the products. I can simply double the job MapReduce without incurring in main memory problems. If i consider top most of v and left most of A I will get the same number of element of v. So I will have to consider blocks of matrix A.



(a) v cannot fits the RAM



(b) Solution for v cannot fits RAM

Figure 3.4: vector v cannot fits the RAM

4

Spark

LECTURE 09-02-2021

4.1. Spark

Spark transparent way of performing computation of files in the cloud. Several way of interfacing to both local storage and cloud buckets.
Hadoop perform a transformation of a file DFS into another DFS file. But for Spark this is not true. It introduce what it a DFS but it builds up an abstraction to link to different ways of implementation. This abstraction is called **Resilient Distributed Dataset** (RDD).
Once you load a dataset, Spark builds an object that represent the Distributed Files and using Object oriented programming we can do transformations. Why Spark? It is newer and more powerful (faster). It implements RDD in a cleaver way with respect to Hadoop. When you use RAM it is intensive time in accessing but Spark implements it as a cache system.

Spark support Scala, Python and Java languages.
Actually, even in term in primitive, Spark is richer than Hadoop. Spark not only Map and Reduce but also other operations:

- **Transformation**: map RDD to another RDD.
- **Action**:take a RDD and process them and the result is not a RDD but something returned to the program itself.

Hadoop is routed in pair key-value and also the output has the same form. Spark can work with pair but also with single values.

Direct Acyclic Graph (DAG) in Spark: abstraction of operations we are executing.
I am hiding parallelization, just to show the logic. We could have several operations like x, y, z, g, f, h, i

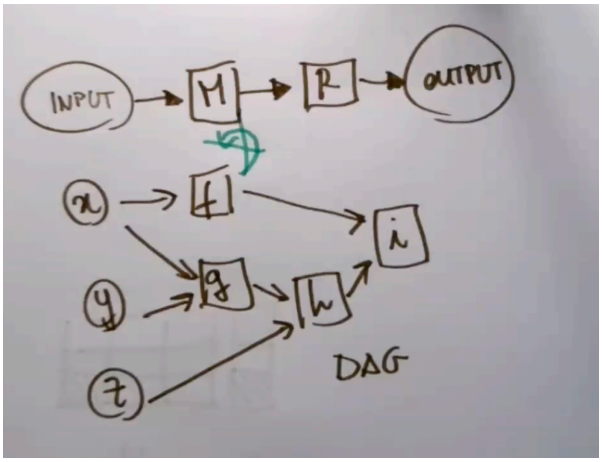


Figure 4.1: Spark logic

When i perform complex operation, the master process build this graph that is a DAG to not have cycle. If I got a cycle the operation is not robust.

Calling KeyReduce function, given a set of keys i will get the keys set and as value a vector corresponding to the count of how much time that key is is the original list of pairs. Then, I will use the vector of 1 as input of my lambda function that execute the sum.

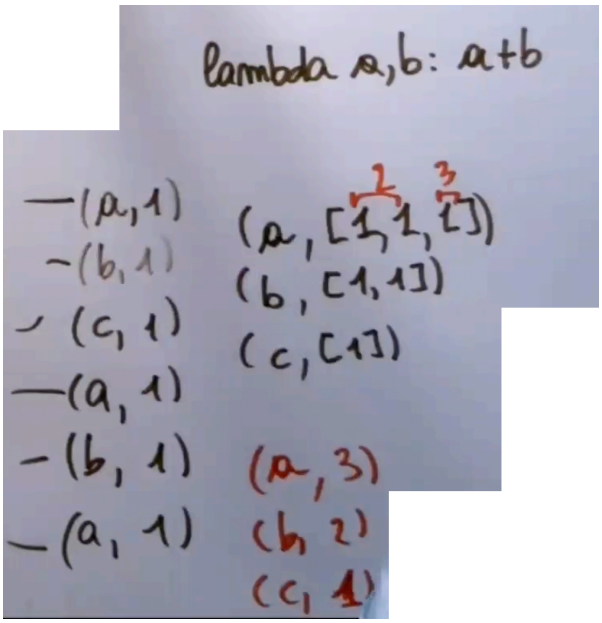


Figure 4.2: Spark KeyReduce

5

Similar Items

LECTURE 15-02-2021

5.1. k-gram

A lot of object and I want to detect a pair of similar. For example for documents is rather easy, I scan the documents and stop as I get a different character then I say the two document are different. The problem is that more difficult to search for similarity (approximate similarity). Example for plagiarisms.

How do we encode documents? there are string way representation is not the best one since we have a sort of order of characters. I plagiarism occurs but in a different paragraph. The idea is representing the document in a certain form of atom: **K-gram**. I might have a value of k >= 2. The simplest form of k-gram is based on character. If I have Good morning and I want to build a 4-gram I will have each combination of the string with 4 characters.
4-grams: *{'Good', 'ood ', 'od M',..}*.

I will shift to the right as the length of the string. The order of occurrence is a key factor, we won't consider the order.

The document will be represented as a set of k-grams

$$DOC \equiv \textit{set of k-gram}$$

We use **Jaccard similarity** to find the similarity between two objects.

$$SIM(s,t) = \frac{|S \cap T|}{|S \cup T|} \qquad \in [0,1]$$

The bigger is the interception, the higher will be the similarity. I am weighting this with the cardinality of the union. This measure is always between $[0,1]$
If I want to measure the similarity of the same object:

$$SIM(S,S) = \frac{|S|}{|S|} = 1$$

The similarity is 0 if the set is an empty set is the result of two sets that are disjoint

$$S \cap T = \varnothing \quad \rightarrow \quad SIM(S,T) = 0$$

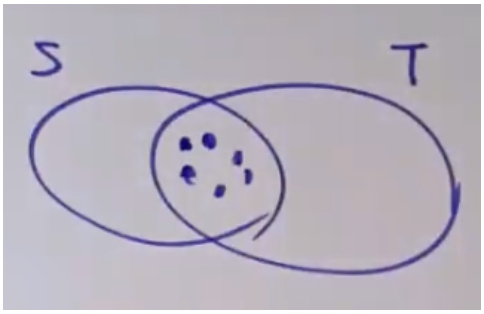


Figure 5.1: Intersection

How big is k? If k is 1, the 1-gram is a string with one character: I will check similarity for a lot of pair of documents.
If k big (length of the document), similarity would get only to find equal document. I need a criterion to find k in a equilibrium spot.

For example
My doc are e-email, k = 5 is good.
I have 26 different chars and how many k-grams I have? 26^5 k-gram which is $\approx 14\ Milions$
A email contains 1000 chars typically and using these 14 M the probability it occurs in my doc is very small. Why I am saying that? The length of document is upper bound of the number of k-gram which describe it. If I have number of k-gram that describe a document which could have a lot of possible k-gram, the probability that appears in all my doc, is very low. Scientific article the length is higher, so k = 9 would be preferable. Textual document is not going to be described by k-gram but there is a pre-processing phase (stop-words, stemming and lemming.

How we know how to get to there encoding using the k-gram using the element of the k-gram. If I have a big order document I will face two problems: space and time.

Space: how do I encode all of them encouring in space problem? I can build a characteristic matrix in which I will have documents in the columns and k-gram in the rows.

	D1	D2	D3	D4
a0	1	0	0	1
b1	0	0	1	0
c2	0	1	0	1
d3	1	0	1	1
e4	0	0	1	0

Table 5.1: Characteristic Matrix

k-gram is contained in the doc D_1 and D_4.

$$SIM(D_1, D_2) = 0$$

$$SIM(D_1, D_3) = \frac{1}{4}$$

$$SIM(D_2, D_4) = \frac{1}{3}$$

The number of k-gram is higher than documents, this matrix will no fits the RAM.

5.1.1. Min-Hashing

We can consider the **Min-Hashing**: function in which rearranging Characteristic Matrix. I will consider the permutation of the rows to fits the RAM. Problem is that simulating a permutation of n element such $n!$ will be generated is complex.

We consider a sort of trick: consider an Hash function from set of k-gram and buckets such that they are equal. For example in this case I have 5 arguments to my hash function and 5 possible buckets. Hash function guarantee to approximately distribute uniformly arguments and the buckets. If that happen, if I have as many buckets as argument, each many buckets is associated with one and only one argument. If I start the sequence of images of this k-gram I got the permutation that is *"a,c,e,d,b"*.

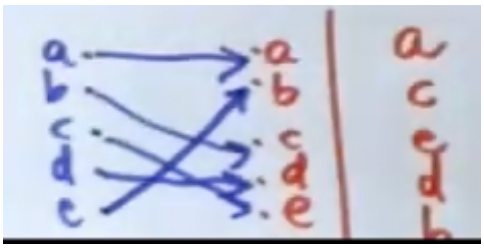


Figure 5.2: Min-Hashing

$$h_1(r) = r + 1 \mod 5$$

$$h_2(r) = 3r + 1 \mod 5$$

Once I have the permutation I will be able to compute the row of the new matrix that is called **Min-**

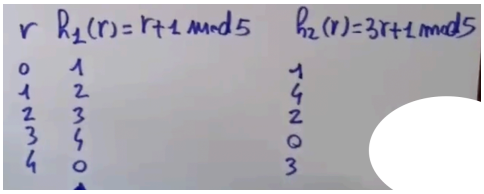


Figure 5.3: Hash Results

Hashing Signature. It has as many matrix as the characteristics but now refers to the hash function I choose. So I will get the order based on the result of the hash function. Most of the time I will get a pure permutation of the rows. How to fill the signature matrix? I will basically scan the characteristic rows of a document a stop at the time I get 1. In the Characteristic Matrix I have 1 in the first position

	D1	D2	D3	D4
h1	3	2	1	2
h2	0	2	1	2

Table 5.2: Min-Hash Signature Matrix

for D_1 but now is no more the first so I got $0 - 0 - 1$ which is the 3^{rd} position. I stop at 2 for D_2, 1 for D_3 and 2 for D_4. It will never happen to have 0 in a column, that would mean that the document is empty. If I do the same for the hash of the second I would get 0,2,1,2.

When I compute h_1 to each value of the rows I have computed a possible way of permutating the rows of Characteristic Matrix. For example considering the h_2 It is equal to the Characteristic Matrix but we rearrange the rows. How to get Min-Hashing? I scan since I got a value of 1 in a cell. A simply record then the values: 0,2,1,2. This values correspond to the hash function h_2.

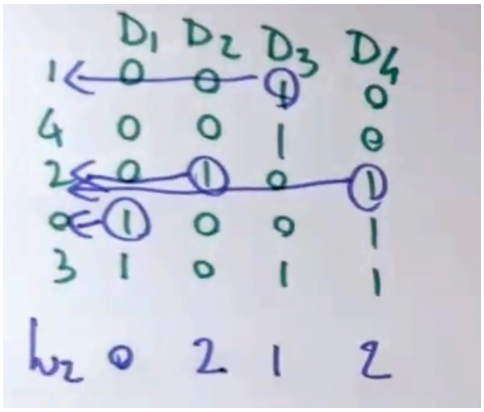


Figure 5.4: Hash computation

Now we will see a Mathematical property of this is computing an approximation of the similarity of the two documents. Considering Table 5.2

$$\widetilde{SIM}(D_1, D_2) = 0$$

$$\widetilde{SIM}(D_1, D_3) = 0$$

$$\widetilde{SIM}(D_2, D_4) = 1$$

These values are computed by dividing the columns value for each documents. I have 0 and 0 as approximation, I have $\frac{1}{4}$ and 0 as approximation, I have $\frac{1}{3}$ and 1 as approximation. Which value do I give? If I have chosen the permutation at random the probability agree in exactly one row correspond to the Jaccard similarity of the two documents.

$$P(h_i(D_l) == h_i(D_m)) = SIM(D_l, D_m)$$
$$P(h(D_i) == h(D_j)) = SIM(D_i, D_j)$$

When I am scanning the Characteristic Matrix I have different possibilities:

- x-type rows $\Rightarrow$ $1\ 1$
- y-type rows $\Rightarrow$ $1\ 0 \quad || \quad 0\ 1$
- z-type rows $\Rightarrow$ $0\ 0$

$$h(D_i) = h(D_j) \leftrightarrow \textit{ first non-z-type row is an x-type row}$$

$$P(h(D_i) = h(D_j)) = \frac{x}{x+y}$$

$$SIM(D_i, D_j) = \frac{x}{x+y}$$

So:

$$SIM(D_i, D_j) = P(h(D_i) = h(D_j))$$

The second problem is related with time. How much time do I have to invest? I have to check all the possible pair of documents. If I have $1M$ of docs I will have Example of binomial

$$\binom{n}{2} \approx \frac{n^2}{2}$$

If I have $1M$ docs

$$10^6\ docs \qquad \binom{10^6}{2} \approx \frac{10^1 2}{2} \qquad \textit{possible pairs}$$

Suppose it took for a CPU for operation a time like: $1\ \mu\ s = \ 10^{-6}\ s$
How much time do I need?

$$\frac{10^12}{2}\ \cdot\ 10^{-6}\ s \approx 5.78\ \textit{days}$$

5.1.2. Locally Sensitive Hashing

A procedure to make it faster it is called **Locally Sensitive Hashing**(LSH).
As output of my procedure False Positive(FP) and False negative (FN). FP my procedure would be faster but it may suggest that a pair of docs are similar even if they aren't in reality.

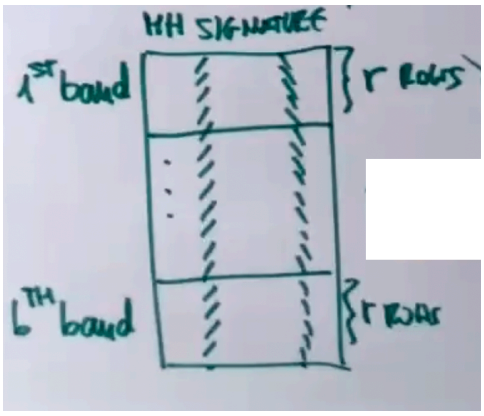


Figure 5.5: Locally Sensitive Hashing

$$b \cdot r = n$$

Dividing the matrix in b number of bands. r is the number of rows in each band. I could use different hash function and I would select document if at least of my band the two sub column map the same quantity.

$$P(\textit{ same val in a row}) = s$$

If I am interested in:

$$P(\textit{ same value in a row}) = s^r \qquad = P(\textit{ same value in one band})$$

$$P(\textit{ not having same value in a band}) = 1 - s^r$$

In all my bands:

$$P(\textit{ not having same value in all band}) = (1 - s^r)^b$$

What happens of the complement?

$$P(\textit{being equal at least in one band }) = 1 - (1 - s^r)^b \qquad = p(s)$$

If I try to plot this function? I obtain a sigmoid function of s. How can we find the s^*?

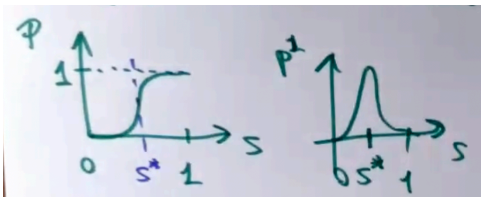


Figure 5.6: Plot distribution

I would get the right plot taking the first derivative (Gaussian like).
I would have to check the second derivative of p and set it to 0.

$$p^{''}(s) = 0$$

This happen if

$$\frac{1 - s^r}{s^r} = \frac{r}{r-1} \cdot (b-1) \tag{5.1}$$

It is difficult to solve this equation, but we can see that we can get a good approximation of the best solution:

$$s^* = \frac{1}{b}^{\frac{1}{r}}$$

Substituting in left of (5.1) we get that in the left:

$$\frac{\frac{1}{1-b}}{\frac{1}{b}} = \frac{\frac{b-1}{b}}{\frac{1}{b}} = b - 1$$

and in the right we have $\frac{r}{r-1} \cdot (b-1)$

At the end:

$$1 = \frac{1}{r-1} \qquad \textit{BUT THIS IS NOT TRUE}$$

It's not true but it is a good "approximation" of s^*.
I will fix $t \in [0,1]$ and I solve a system of equation

$$\begin{cases} b \cdot r = n \\ t = (\frac{1}{b})^{\frac{1}{r}} \end{cases}$$

I might choose t close to 0 or to 1 and this would be the threshold to which a pair of document is similar or not. t close to 1 I would have few pair of documents (higher number of FP) t close to 0 I would higher number of pair of documents similar (higher number of FN). If I have FP I could check it, while for FN I could not because I don't know if they are chosen or not.

All the things can be applied to different concept. The key thing is to have a **distance** concept. We can calculate distance in different ways (Euclidean, Hamming).

5.2. Similar Items part.2

LECTURE 16-02-2021

There is a property for functions I can use in the same way of LSC. We used it to reduce number of documents in which we search for similarity.
A family F of LSH functions is set to be $(d_1, d_2, p_1, p_2) - sensitive$: $\forall\ f\ \in F$:

- $d(S_i, S_j) \leq d_1) \quad \rightarrow \quad P(f(S_i) = f(S_j)) \geq p_2$
- $d(S_i, S_j) \geq d_2 \quad \rightarrow \quad P(f(S_i) = f(S_j)) \leq p_2$

If the distance of two element is $\leq d_1$ then the probability they are will be selected is $\geq p_1$

We can plot this properties: On y we have probability of being selected, on x we have the distance.

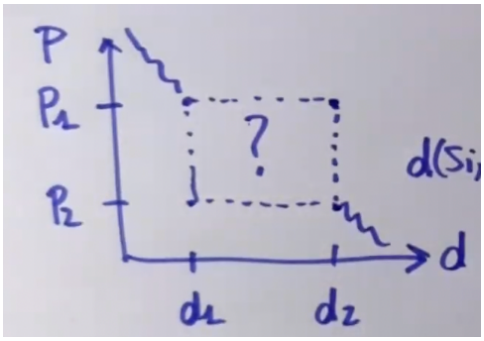


Figure 5.7: Properties distance-probability

I don't know what happen in this rectangle. But i know if $\geq$ of d_1 or d_2.

$$F = \textit{Family of min-hash}$$

We are speaking about similary and if i take Jaccard similarity and i compute the complement

$$d(S_i, S_j) = 1 - SIM(S_i, S_j) \qquad \textit{Jaccard dist}$$

If I know the Jaccard distance is smaller than d_1?

$$d(S_i, S_j) \leq d_1$$

$$1 - SIM(S_i, S_j) \leq d_1$$

$$SIM(S_i, S_j) \geq 1 - d_1 \qquad \leftrightarrow \qquad P(f(S_i) = f(S_j)) = 1 - d_1$$

The probability is now dependent on d_1.
I shown that my family is:

$$F \textit{ is } (d_1, d_2, 1 - d_1, 1 - d_2) - sensitive$$

I can apply an amplification of this sensitiveness: I start with F and order this family:

$$F = \{f_1, f_2, ...\} \qquad (d_1, d_2, p_1, p_2) - sensitive$$

When distance of two document is smaller than d_1 but we get the prob higher than p_1 but p_1 is small. I fix AND-constructor fixing r.

$$AND - construction: \textit{ fix } r \in \mathbb{N}$$

Family is linked with r of the function. I am interested if a function applied to different documents return the same value. So i check if the function applied to S is equal to the same function applied to T:

$$F^{AND} \ni f^{AND} \equiv (f_1, f_2, ..., f_r)$$

$$f^{AND}(S) = f^{AND}(T) \leftrightarrow \forall i \; f_i(S) = f_i(T)$$
$$d(S,T) \leq d_1 \;\rightarrow\; P(f_i(S) = f_i(T)) \geq p_1 \qquad \forall \; i = 1, ..., r$$

$$P(f^{AND}(S) = f^{AND}(T)) = P(\forall i \; f_i(S) = f_i(T)) = \prod_{i=1}^{r} P(f_i(s) = f_i(T)) \geq p_1^r$$

I know that regardless of i, it is greater than p_1. Same applied for p_2:

$$d(S,T) \geq d_2 \quad \rightarrow \quad P(f^{AND}(S) = f^{AND}(T)) \leq p_2^r$$

Maybe i know p_2 but it is not satisfactory. What if p_2 is no low? But p_2^r will be closer to 0.
We can do a similar construction that will compute this function f^{OR} each time i get the same value.

$$OR - constructor$$
$$f^{OR}$$

These select a copy of a pair of doc if for at least one I got the same value

6

MarketBasket Analysis

6.1. Frequent ItemSets/ Market-Basket Analysis

The typical application comes from marketing. I want to find out if two kind of product is going to be purchase together. A frequent set is a set of documents in which I maybe want to check for plagiarism. I might use frequent item sets to build the so called **Association rule**.

$$I \rightarrow j$$

Where I is the set of ITEMS and j is the single item.
If in a basket we have all items of I, maybe j will be contained in I.
I need a measure of goodness of the item set.
The **Confidence of the rule**

$$CONF(I \rightarrow j) = \frac{supp(I \cup \{j\}}{supp(I))}$$

where supp function check how much element are in a basket. Once i get the association rule with high confidence i should check the **Interest of a rule**:

$$INT(I \rightarrow j) = CONF(I \rightarrow j) - \frac{supp(\{j\})}{tot.baskets}$$

Maybe there are special item that tend to be frequent independently to the set of item (for example when i go to the grocery i always get water bottles $\Rightarrow$ So Water will tend to be in a lot of baskets). Note that the second element is the number of baskets that contain j divided by the total number of baskets. I might have different situation that may depends on that rule.

- $\approx 0 \rightarrow I$ has no influence on item j
- $> 0 \rightarrow$ baskets with all item of I tend to contain j
- $< 0 \rightarrow$ baskets with I tend to NOT contain j
  this could happen for example for two sodas: soda A and B. Soda A will always substitutes soda B so baskets with I tend to NOT contain j.

$$J \textit{ is frequent} \qquad \forall j \in J \qquad J \setminus \{j\} \rightarrow j$$

This mean for example, if i have 4 items that tend to be purchased, when i see a set of 3 of them, it would be likely that the customer will purchased the 4^{th} of that set.

Why we are speaking about this in our course? We could have massive set in a big supermarket. We can hypothesis some properties:

- set of baskets: too big for RAM
- each basket: fits RAM
- n diff. items $\rightarrow \binom{n}{2} \approx \frac{n^2}{2}$ $\binom{n}{k} \approx \frac{n^k}{k!}$
  $n = 10^5$ $\frac{10^{10}}{2}$ 4 byte (of counters)
  $2 \cdot 10^{10} byte =$ 20 Gbytes
  But we haven't 20 GB of RAM.

I need a way to organise my counters but also to access properly to them. Each item will be identify by integers numbers.

I have a matrix $n * n$, and we can access to it in constant time. But i will have a symmetric ma-

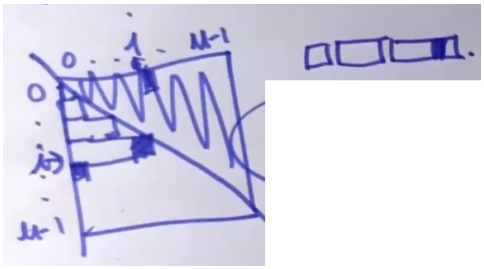


Figure 6.1: HashMap

trix so i can just use the half of this matrix. But this will be a waste. Moreover, do I have the possibility of working with a triangle matrix? But I could consider each rows of this triangular matrix and build a rectangle from it. From this we get a vector in which we concatenate each row and I have no waste of memory. But i luck on accessing the items.
How can we access to the pair (i, j) I can use this formula

$$k = (i-1)(n - \frac{i}{2}) + j - i \tag{6.1}$$

I might still waste memory space for the sparsity problem. I might consider a different approach. As we speak in PageRank, we stored the matrix as a set of triples: (i, j, c). To access to it I will not organise it linearly but using index of an HashMap

$$(i, j, c) + HashMap \tag{6.2}$$

This access is controlled, but we have the problem of the hash in which we could get the same value from different keys. So, if i don't have RAM problem i would use Formula 6.1, otherwise I would prefer Formula I will use 6.2 How do we use this way of organising counter?
a **Apriori method** that is based on monotonicity property.

6.1.1. Apriori Algorithm

How do we use this way of organising counter?
a **Apriori method** that is based on monotonicity property.

$$\textit{Monotonicity} : I \textit{ freq} \quad \rightarrow \textit{each } J \subseteq I \quad \textit{is freq}$$

$$supp(J) \geq supp(I)$$

If a set I, then a subset J needs to be frequent to. This is because $supp(J) \geq supp(I)$ How Apriori algorithms is structured?

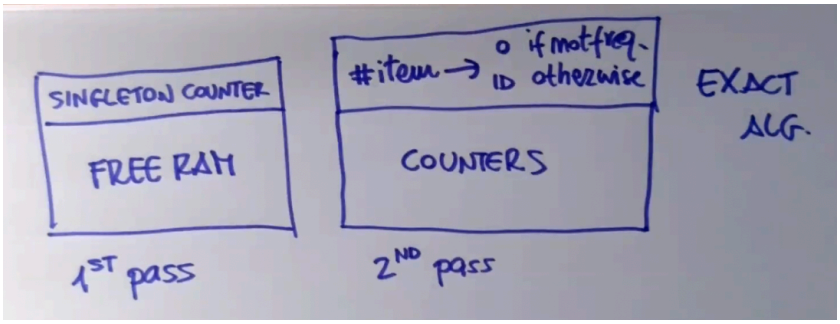


Figure 6.2: Apriori Algorithms

First rectangle represent all the available memory. I will scan all basket files analysing single basket. For each item in a basket i will increase a counter. We cannot have a counter for each pair. So that, when i have compute the 1^{st} pass i know how many basket contain that singleton. I can check if it is higher or lesser than my support threshold. All the rest of memory is free. Before going on, I will perform a small operation to my counter: I will assign item to an integer (that counts how many times it appears).

$$\textit{Number of items} = \begin{cases} 0 & \textit{if singleton not freq} \\ ID & \textit{otherwise} \end{cases}$$

This numerical ID indexes as index either to access my triangle matrix or my triples base representation. Using triangular matrix I would raise space.
Example: If first item is not frequent i will assign zero, up till i find a singularly frequent item and assign it 1, the next will be 2 and so on. The amount of RAM is the same of 1^{st} pass.
Now all the main memory all the remain memory will be devoted to the structure of my counter (triples or triangular matrix). But how the algo works during 2^{nd} pass?
If i results in a frequent item, then the (i, j) pair will results as a frequent (each subset of the pair have to be frequent).

Note that this algorithm is Exact algorithm: i have not FP, no FN. Why? If i have FN, happen to be frequent but this would not happens to for monotonicity property. FP if some of the pair would actually not be frequent, but this cannot happen because i check that the value of each counters is actually greater than threshold value. So if i output value, the support is for sure greater than the threshold value.

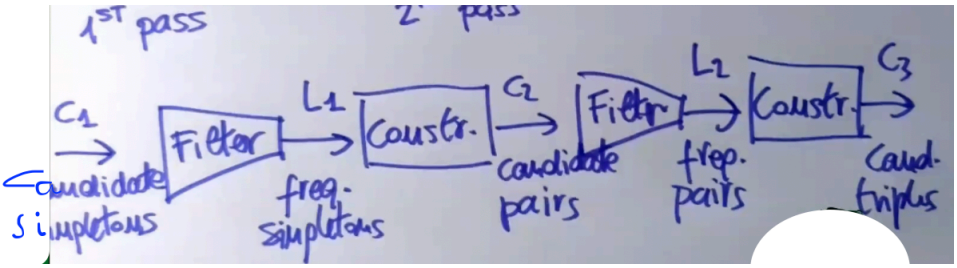


Figure 6.3: Apriori pipeline

$$(i, j, k, c)$$

I start filtering all my singleton candidates in C_1 to get the new frequent singletons in L_1. Via a construction procedure I arrive with a set of candidate pairs (generated from counters update in 2^{nd} pass). I have to check if this counter are actually frequent. Then i check if my triples are frequent. Another construction with candidate triples. I would do another filtering till I get a results. So we fix a cardinality order to define how much passage of construction we are going to do. I will be able to find set of frequent item of any cardinality. The cardinality depends on the cardinality of the basket we desire in output.

6.2. Market Basket Analysis part. 2

LECTURE 22-02-2021

We saw 2 scans of RAM. We used a small fraction of all available memory in 1^{st} scan to analysis singletons counters.As I get the counters I could check in the 2^{nd} scan whether or not they are frequent or not. Spare more memory to counting the frequency for each set, so i discarding pair that are not interesting.

6.2.1. PCY-Variant

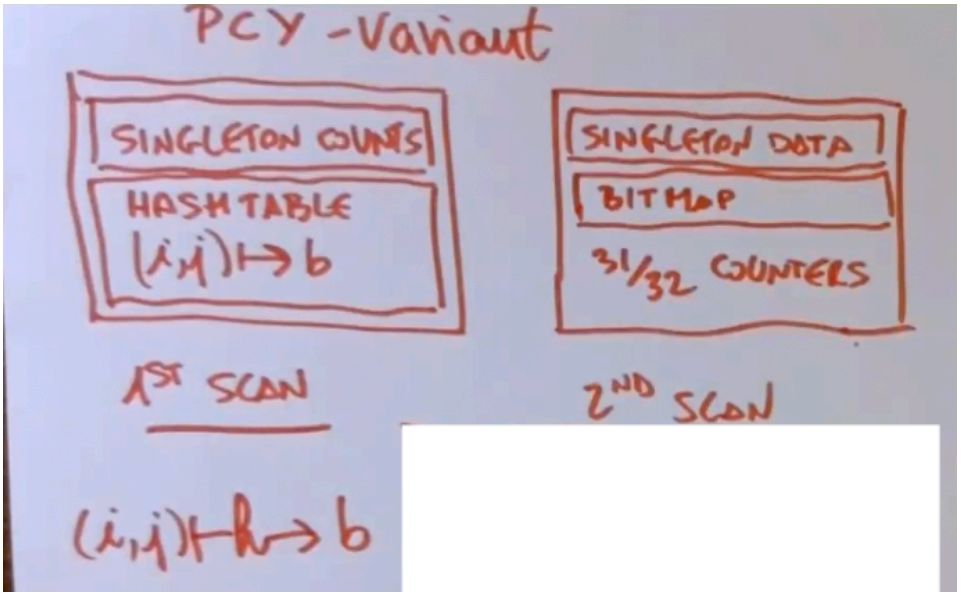


Figure 6.4: PCY-variant

This consider that in the first scan i consider a small fraction of available memory. Most of the RAM is not used at all. We want to farther reduce the number of pair. Think that you have a hash function h that map a pair to a bucket b:

$$(i, j) \to h \to b$$

How many buckets do I choose? I want to fill the remaining memory with a set of b counters. The idea is that i want to associate each pair with a counter. So the first phase is actually equal to the Apriori algorithm. Scans all bucket, check how many times they appear in the basket but before passing to another variant, it build all pair in the basket and increase the corresponding counter. What happen at the end of this scan? If a pair is frequent, the corresponding counter in the hash table has been increase that is higher than my support threshold. If is lower, the accumulating of different pair is lower than the support threshold. Looking at the counters, the only think I could say is that if any pair max a counter that is higher than my support threshold i can't say for sure they are frequent yet.

2^{nd} SCAN
I can keep this structure in memory, but it used it all. I just to check that the value is higher than the support threshold, so i can replace each value with a Boolean number that state if a counter is higher or lower to the support threshold (Using a bitmap). A typical dimension of a integer is 32 bits, now I am using only 1 bit. This mean is that the free space is approximately equal to $\frac{31}{32}$ then the original one. If at least one of the two item is not singularly frequent I will not consider them. Then, I will have to check if both item in a pair are frequent and they map to my BitMap to one. I will be able to filter out no frequent pair. No False Positive and No False Negative using this technique.

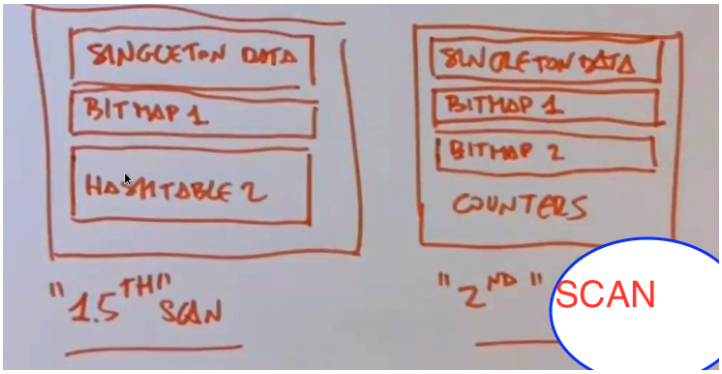


Figure 6.5: PCY-variant adding intermediate scans

I might add an intermediate phase but considering another hash function. This is the $1.5th$ scan. In 2^{nd} scan i will have two bitmaps. I have to pay this with another scan. I am free to add as many hash table I want. I will have to find a trade off. Increasing bitmap will have the effect of diversifying the frequent pair but i will have the cost of many scans. I do also have another possibilities: different filters but does not require more access.

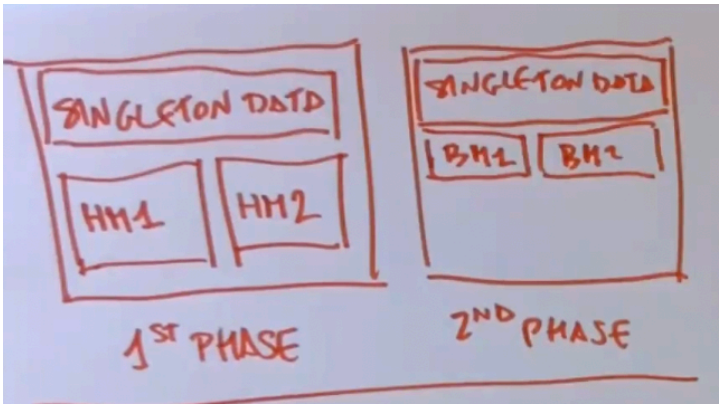


Figure 6.6: PCY-variant adding intermediate scans 2

As you build your singleton data, instead of filling up with an hash, we fill the memory with two portion: two hash-map.
In second phase? I have my singleton data but i will have 2 BitMap that together take much space as took here Figure 6.5. Now I can apply an intermediate version applying a second scan checking if they are frequent. I have more filter but same space and time. I might divide it in more of Hash Map and BitMap: as usually we have to find a good trade-off.

An interesting alternative of the Apriori algorithm is based on sampling: **Sampling Baskets**.
You have the possibility of processing a small sample of the baskets. This way has interesting advantages:

- not just pairs
- only requires 2 scans
- Partial results: not finding all the results

As I am sampling i can have ALL IN RAM! So I can have not just pair and require only 2 scans. How do I practically sample my files? It depends on the structure and technology.
If i want to fix a fraction p between 0 and 1.

$$p \in (0,1) \qquad m = \#\, baskets$$

Can i get first $p \cdot m$ baskets? It depends on how that file is created. Each time I will generate a random number and check if is greater or not to my threshold, in this way I will properly generate a sample. If the original threshold:

$$threshold \quad s \rightarrow p \cdot s$$

I need to scale also the threshold. The fact that I am not using all the available data will result on False Positive and False negative.
What about FP and FN? There is a difference between FP and FN. If i get a FP I am able to scan if it really a frequent set.
For FN is not possible but i can make something and lower the likelyhood of FN. So I could lower the ratio of the suggested threshold. But this also increase the number of FP.

$$threshold \quad s \rightarrow p \cdot s \rightarrow 0.9\, p \cdot s$$

6.2.2. SON algorithm

I have some algorithm which exploit the sample techniques. This algorithm perform an interest hypothesis: organise baskets in chunks. I will process chunk of baskets.
Also, I will fix a fraction p that will tell me the fraction of baskets that each chunk contains. If p is 0.1 i have $\frac{1}{10}$ of total number of baskets in each chunk. I will use a chunk as a set of data.

$$p \in (0,1) \qquad EXACT \quad ALGORITHM$$

- baskets: $\frac{1}{p}$ chunks
- $\forall$ chunk find frequent sets with threshold $p \cdot s$
- merge results: set of candidates
- filter out FP

I will perform a second scan and I will count the occurences of each candidate sets. I have already a first scan although distributed considering each chunk separately. This is an exact algorithm, since the FP are filtered out on the last step.
It is easy to show that I don't have False Negative. Let me say I have FN set I. It is not frequent. So if i check the $supp(I)$:

$$supp(I) \geq s \qquad \forall j \; supp_j(I) < p \cdot s$$

We refer to $supp_j$ as the support in the chunk j.

$$s \leq supp(I) = \sum_j supp_j(I) < \sum_j p \cdot s = \not{p} \cdot s \cdot \frac{1}{\not{p}} = s$$

I can express the support as the sum of the support in each chunk. As we say before is smaller than $p \cdot s$. Summing the chunk so i can divide by the number of chunks $\frac{1}{p}$ to get s at the end.
I say that s is **strictly lower** than **itself**, which is **not possible!**. So i have no FN.

SON algorithm with MapReduce
If I have a baskets file stored in a distributed file system which support the MapReduce framework I can easily a MapReduce job that apply this algorithm.

- M1: read chunk, find frequent sets with adjusted threshold
  $\forall$ found candidate F, emit $(1, F)$
- R1: read $(1, [F_1, F_2, ...])$
  emit its input with no duplicates.
- M2: read chunk and set of candidates
  $\forall$ candidate C emit $(C, n(C))$ (calculate number of occurrences of set C)
- R2: aggregate by sum, emit $(C, 1)$ if $\sum \geq s$

Where M refers to a Map and R to a reduce. s is the support threshold. SON algorithm I don't use sample in reality, I am processing all basket, while is not true with the Toivonen Algorithm.

6.2.3. Toivonen Algorithm

Each time i run it it will return frequent item sets or a message that it was not able to find any of it. However, as this algorithm rely on sampling.

- Fix $p \in (0, 1)$, sample $m \cdot p$ baskets
- Find frequent items candidates in sample with thresholds < $p \cdot s$ (slightly less, e.g. $0.95\ p \cdot s$)
- Build **Negative Border**
- Perform a full scan on baskets
  - no set $\in$ NEG. BORDER if frequent $\rightarrow$ output all candidate after FP filtered
  - otherwise $\rightarrow$ NO answer

$$s \in \textit{ NEG BORDER} \longleftrightarrow s \textbf{ freq. in sample}$$

$$\forall \textit{ IMMEDIATE SUBSET } T, \quad T \textit{ is not frequent in the sample}$$

$$set \quad S,\ T \textit{ is immediate subset of } S \quad if \qquad \exists\, i \in S, T = S \setminus \{i\}$$

Let's see and example:
items a, b, c, d, e
candidates $\{a\}, \{b\}, \{c\}, \{d\}, \{b, c\}, \{c, d\}$
NEG. BORDER $\{e\}, \{a, b\}, \{a, c\}, \{a, d\}, \{b, d\}$
They belong to the negative border because they don't belong to the set of candidates. But we see if we remove any of this two we get back to the singleton a or b that are candidates. I can build all possible pair adding a,c, a,d and b,d (b, c doesn't because it is already a candidate as c,d). We could check if neg. border contains triples.

7

Recommendation Systems

LECTURE 23-02-2021

I want a system that is able to suggest me something: for example a merchandise.
It born with online purchased.
For example in books: in the real word, what happen in the so called *brick & mortar libraries*
How many different kind of book you think to find in a medium size library? Suppose 10^3 books. When we consider and online store? How many books is amazon able to sell?
10^6 books
Why you don't have a top10 selling books? We have a phenomena called :
Long Tail Phenomenon

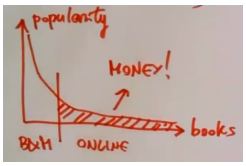


Figure 7.1: Long Tail phenomenon

On the x-axis i have items : books and i want to sort them about popularity (a lot of users purchase that book). I will typically observe an exponential curve like that divide in two parts: traditional libraries (B & M) and online. B & M is able only to manage top sellers since he has to store all the books. In the Online work can also focus in also less sold book. Why interested in the less sold? Because we have more money!
You need to be able to properly suggest this item to your users. So a recommendation system that automatically is able to suggest books to users.

We can focus on a **Utility Matrix**.
Column and rows refers respectively to items and users: So we have HP refering to Harry Potter, Tw

	HP1	HP2	HP3	Tw	Sw1	Sw2	Sw3
A	4			5	1		
B	5	5	4			5	
C				2	4		
D		3					3

Table 7.1: Utility Matrix

for Twilight, Sw for StartWars. A, B, C, and D are users.
The number in cells are the ratings for each book. Most of the entries are the same. This matrix is spars! We are interested in the empty cell to fill with a number that is high. So it is better to suggest

something the user might like.

We want to find a way to fill the empty space in the utility matrix and I have to ways to proceed:

- **Content-Based**: we use properties of the item in order to perform recommendations
- **Collaborative Filtering**

7.1. Content-Based recommendation

The idea is working with profiles. If the profile are close enough I will suggest to the user.

- Profiles
  - Items : year, director,actors, genre, rating
  - Users: ”discovery” for features

Let's focus on the items of the profile. For instance i might consider the year, director, actors and also other info which are vague and harder to catch (genre, public ratings). For year and ratings are described by numbers, what about director and actors? We could build a vector with Boolean components. First 100 components of vector that describe movies are Boolean vectors and they are all set to 0 but one is set to 1 that highlights the director (or the actor). This is called One-hot encoding.

What about the case you don't have this information? Think about recommendation of documents. In this case you must someone have a *discovery* process for the so called features. For example we have a Textual corpus:

- removing stop words
- TF $\cdot$ IDF score: term frequencies x inverse document frequency

$$f_{ij} = \text{\# } \textit{Occurencies of word } i \textit{ in doc } j$$

Once I know the frequency of f I will be able to compute the Term Frequency:

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

where max f_{kj} is the maximum frequency in the document.

$$IDF_i = \ln \frac{N}{n_i} \qquad \textit{where } \ln \textit{ is } log_2$$

IDF is the inverse document frequency. Which is the inverse of the frequency in which a word appear in different documents (n_i) divided by the number of total documents (N) .

$$TF \cdot IDF_{ij} = TF_{ij} \cdot IDF_i$$

TF is high if a term appear a lot in a document, but the IDF will be high if a term i will no be frequent in a lot of documents. So asking the two high mean focusing on a word and document that is not use so much in all documents and it is used in a singular document and there it is used a lot.
After removing stop words I may want to compute the score and choose a subset that maximise the TF. IDF score:
n terms with highest TF. IDF score.
I could use Jaccard similarity but i can use numerical vector as in the previous examples of movies. I can use the same strategy: I just have to compute the term that are not stop words and i place a component in the vector which these words, the value will be 0 or 1 if the term is contained in the document.

What about users? Consider a utility matrix that is Boolean. I want to describe a user U with a numerical vector which describe the movies. Once I have built the profile for my items I can join the info of the profile with the utility matrix. For each Item I will have a profile which is a vector (right in

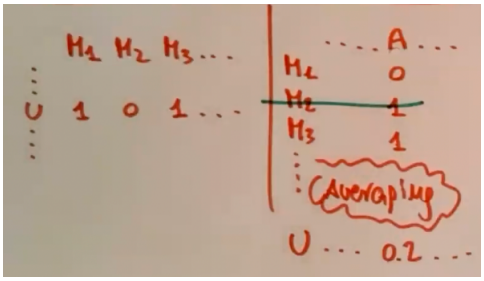


Figure 7.2: Utility matrix for a specific User with boolean entries

the Figure 7.2). This will describe if an Actor A plays a in certain movie (M1, M2,...). The idea is that i will filter these profiles according to the vector in the utility matrix in the row on user U. User U didn't like M2, why would I want to consider the content of that movie to describe him? I will simply metaphorically delete that row. I will have a Boolean value for the author A and then I will compute the average. (In this case 0.2) and I can do that for all the columns. I will get a numerical vector that identify the preference of the user U. In 20% of movies user U watched, the actor A plays in that movies.

When utility matrix has not Boolean entries the situation will get complex.

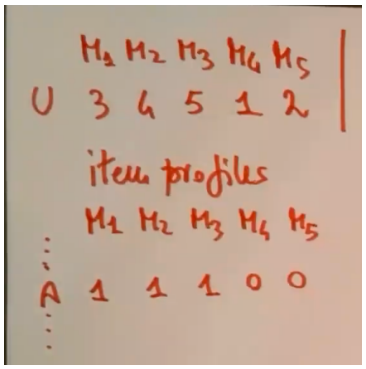


Figure 7.3: Utility Matrix without Boolean entries

We get matrix $U = [3, 4, 5, 1, 2]$ and profile $A = [1, 1, 1, 0, 0]$.
First step is computing the average rating of user U:

$$avg(U) = 3$$

Now I will filter the columns which is the movies in which i do have 1 (in matrix A) and now I will compute the difference between Utility matrix values and the average value.

$$\frac{(3-3)+(4-3)+(5-3)}{3} = 1$$

We get 1, what does it mean? It mean that if I only consider movies in which actor A played, the User U gives a rating which is over the average. The sign of the result is the interesting thing.
In the Boolean case I had number between 0 and 1 that tells me the frequency of the presence of a given actor in movies by the user. Here I get a value positive of negative that identify if the ratings are below or above the average value.
When i Have a vector that describe a user and more vectors that describes items. There is a distance that is suitable here which is the **Cosine Distance** If I consider two vector x and y, there exists a angle θ between the two. The inner product of this vector is a real number equal to the product of the norm of x vector, the norm of y times the cos of the angle.

$$x \cdot y = ||x|| \cdot ||y|| \ \cos\ \theta$$

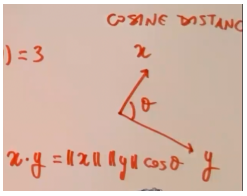


Figure 7.4: Cosine distance

I can express this cosine as in the equation below:

$$\cos\theta = \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}$$

Now, in the case of everything is Boolean which is the information of the product $\sum_i x_i \cdot y_i$? x and y are always 0 or 1, so then it will be the summation of the number of times in which both are equal to 1. So I am considering a case in which an actor plays a movie.
On the denominator we have actually the square root of the number of actors that play in a movie. The idea is that we have a user U (Figure 7.5), several movies M1, M2, M3 and M4. The close the vector are the smaller the angle will be between them. So considering user U, i will consider the closer movies, so the movies with a small θ. Note that this kind of distance does not take in account the length of the arrows.

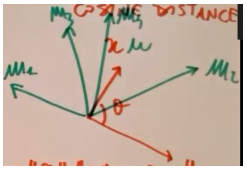


Figure 7.5: Movies Vectors

7.2. Collaborative Filtering

The other method to solve a Recommendation system problem: **Collaborative Filtering**.
More or less find similar user and the typical similarity involving the rows of the utility matrix. I identify a user with the item he/she rated. In order to say that two rows are similar I need a similarity measure. We spoke about having a distance is equivalent of finding the similarity. When I want to maximise similarity, I want to minimise distance. Which kind of distance? It depends on how I want to describe my users:

- **if users:** rows in U $M \rightarrow$ *cosine distance*
  - We have the problem of missing values(since Utility matrix is sparse);
  - A solution could be replace $N/A \rightarrow 0$;
- **if users:** sets of items $\rightarrow$ *Jaccard distance*
  - I have a lot of Variations

Maybe utility matrix is sparse I might prefer to compute clustering using extra-knowledge (group books about the same genre) or using classical clustering techniques. If I replace all rows in UM that correspond to user in the same cluster I will shrink the matrix reducing the rows. What happen after i found similar user? I will select a user to which i want to suggest items, then I select user similar to it and compute the average of entries for similar user and I will estimate the rating that user gave to items as the average of the evaluation of similar user have done.

We will have a last approach which is called **UV-decomposition**.

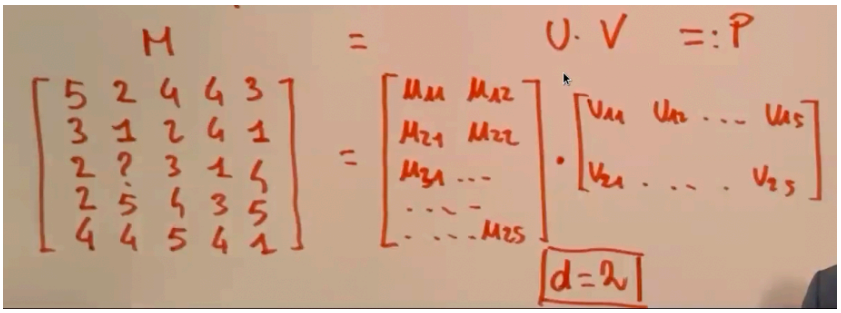


Figure 7.6: UV Decomposition

I want to find a Matrix U having two columns and a Matrix V such that their product is equal to M. If I am able to find numerical values such that their product is equal to M, this product is called P. As entries P are similar to entries M I can estimate the missing entries in matrix M which are represented by $?$ symbol. We have a degree of freedom which is equal to 2: $d = 2$ but we can choose other values. I want to link users in matrix U with movies in matrix V. How to do this operation? We need to initialise these matrices. We initialise in a common value which is one. The idea is that we are going to perturb

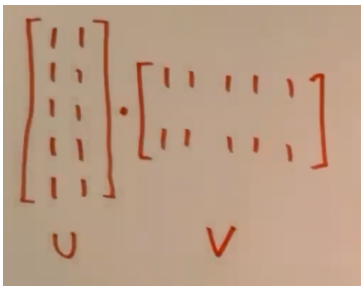


Figure 7.7: Initial values of matrix U and V

some element of U or V and this is done to get a results which is closer to M. In order to proceed we need to compute the distance between two matrices. How? With the *RSME*.

$$RMSE(P, M) = \sqrt{\frac{1}{n}\sum(P_{ij} - m_{ij})^2}$$

I will take generic element P_{ij} and compute the square difference with the corresponding element m_{ij}. We can do it since we know element of M but in not all cases we have a value for m_{ij}, so we want to compute an error.
Now perturb U to find an x which is the best way i can modify the top-left element in matrix U. What happen to the result?

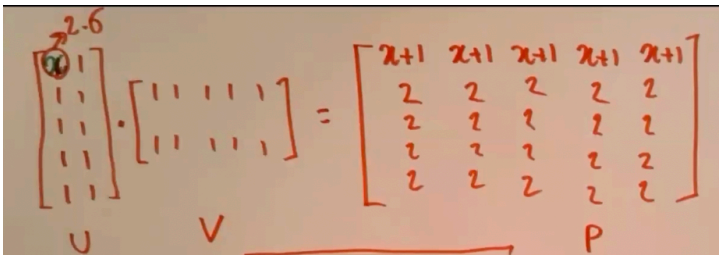


Figure 7.8: Perturbation of matrix U

This is my P and if I want to compute the RMSE refering to M matrix in Figure 7.6:

$$(5 - (x+1))^2 + (2 - (x+1))^2 + ... + (3 - (x+1))^2 + k$$

We stop till the end cell of the row and on k we have a constant that does not depend on x I want that the new matrix U bring me a P that is close to M, to I want to minimise the RSME. To compute the min I compute the difference w.r.t x:

$$\frac{\delta}{\delta x} = 0 \qquad\qquad x = 2.6$$

How Algorithm proceed? I will iterate each item also for V till the end changing each time elements of U and V.

General formula: suppose i want to get u with respect to x: $u_{rs} \to x$ I can do it using:

$$\sum_j (m_{rj} - p_{rj})^2 = \sum_j \left(m_{rj} - \left(\sum_{k \neq s} u_{rk} \cdot v_{kj} - x \cdot v_{sj} \right) \right)^2$$

What happen when i get the derivative of this quantity?

$$\frac{\delta}{\delta x} = 2 \cdot \sum_j \left(m_{rj} - \left(\sum_{k \neq s} n_{rk} \cdot v_{kj} - x \cdot v_{sj} \right) \right) \cdot v_{sj} = 0$$

If you solve this equation by x we get in the end that:

$$x = \frac{\sum_j \left(m_{rj} - \sum_{k \neq s} v_{kj} \right) \cdot v_{sj}}{\sum_j v_{sj}^2}$$

Once you have decided the rows and columns index of the matrix U you can apply this formula and you can use the same formula for matrix V. After some repetitions and when the RSME of $V \cdot U$ is similar to M we are able to predict the missing value of M. Performing this iterative process as several things alike ML. We have to choose which element to perturb, criterion for converge, overfitting and all the things for ML algorithm. Most of the time they do apply here.