

APACHE  
LUCENE  
EUROCON



# Adapting Ajax Solr to Compare different sets of documents

Joan Codina-Filba, Barcelona Media-Innovation Center

19-Oct-2011

Presented by

**lucid**  
IMAGINATION



# What I Will Cover

- More than just Facets
- We do research on opinion mining
- To Give an overview of thousands of opinions
- After analyzing the contents of them:
  - **How to explore the key concepts expressed in the opinions.**
  - **Are the key concepts the most common ones, or the most differentiating?**
- Use of Solr and Ajax Solr for a more intelligent data navigation

# My Background

- Joan Codina Filbà
- Barcelona Media – Innovation Center
- Senior Researcher
- Experience on
  - Data => Text => Web => Opinion MINING
  - Integration of NLP tools.
  - Multimodal search
  - Some teaching

# The Challenge

- Companies want to know what does the web 2.0 say about them:
- Too many writers, in many places, not enough time to read.
  - Detect hot topics (Quality, Price, service?)
  - Detect strengths / weakness
  - Knowledge about the market
- Use NLP to detect key concepts in text
- User generated content, noisy, irony, dispersion
- From data to information and knowledge

# Use Case Example

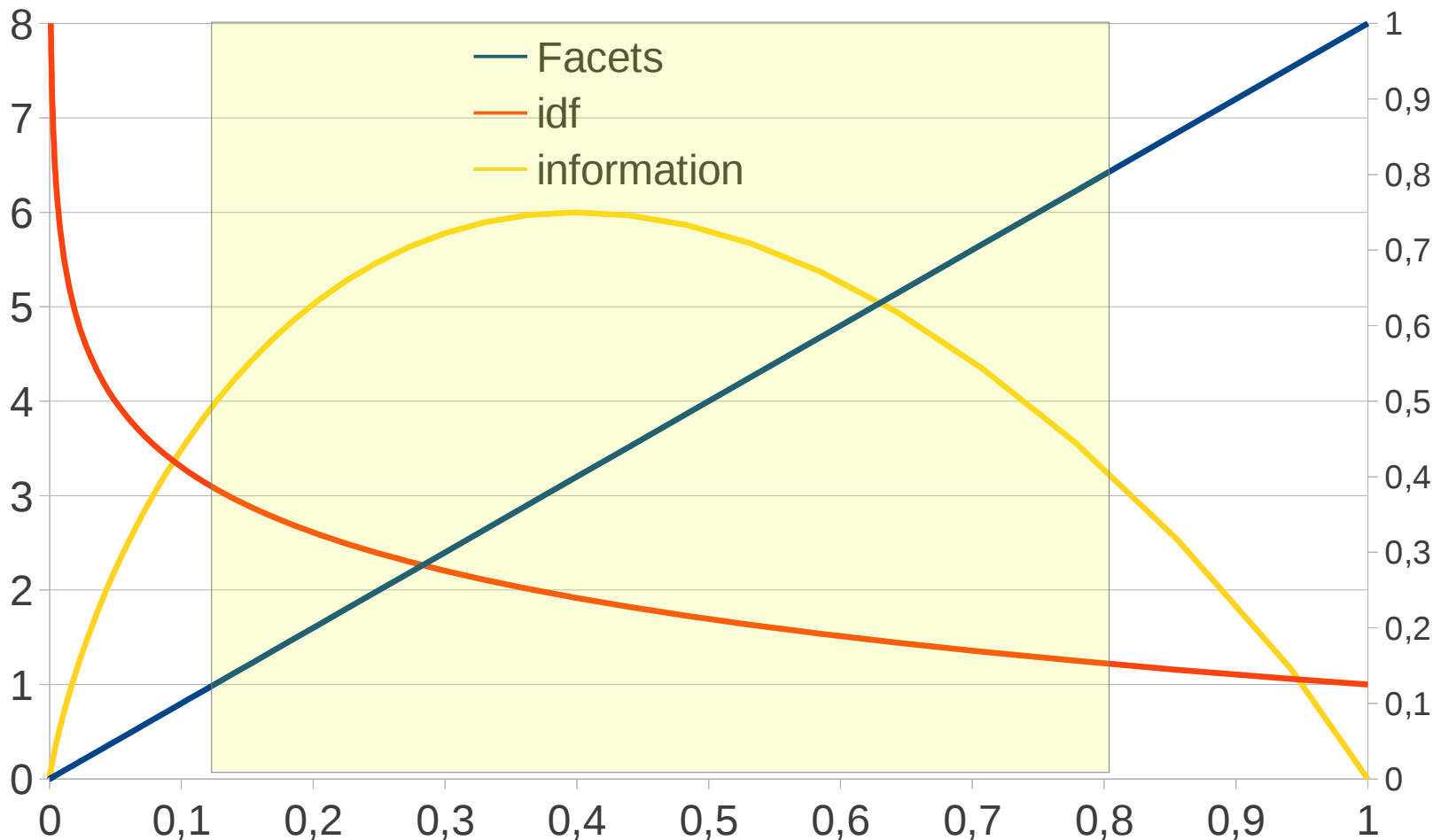
- Analyze the users in MySpace
- Combination of demographic data and analyzed text
- Possibility to "navigate" the data to get "knowledge" of the users.
- Exploring thousands of messages, before sending them to a "blind" , "automatic" text mining tool to find relevant aspects.
- An "OLAP" for data and text, a kind cube view of the opinions

# Solr Facets

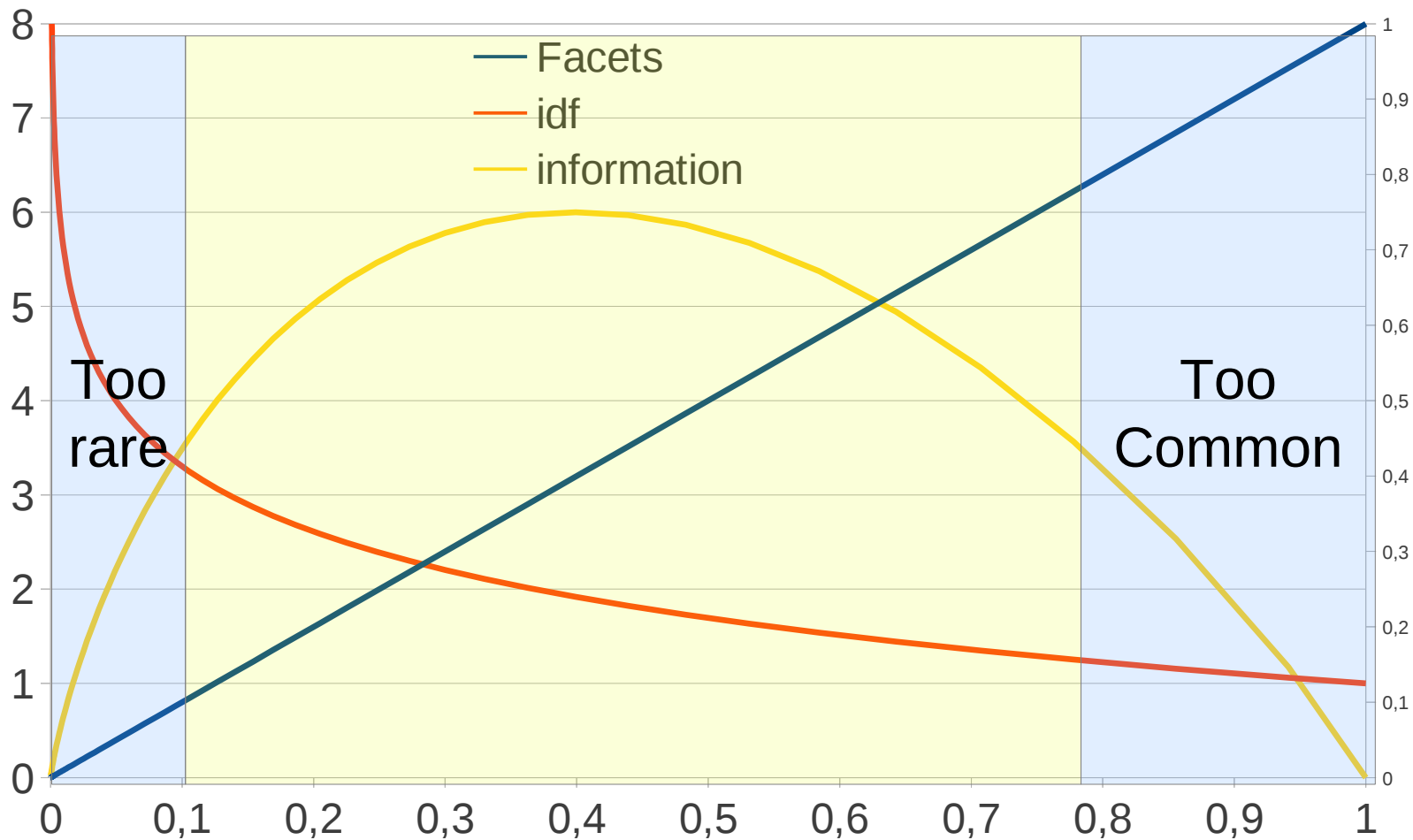
- Simple statistic...
  - For fields with a single value:

```
SELECT field, count(x) FROM table GROUP BY field
```
  - Simple aggregation statistic
  - For text fields ?
- Facets vs IDF
  - The most common facets give less information
  - The rare ones, are not useful

# Information given by Facets



# Information given by Facets



# The Information is in the “not so common” terms

- Doing a search and looking at facets to characterize the data:
  - For most of the queries the most common facets do not change significantly
  - Sorting by frequency:
    - The order almost is the same.
    - Difficult to see the changes
- The information is not on the term's frequency but in the relative changes on these frequencies

# Example

- Exploring MySpace posts
- Demographic Data: Age, Gender, Country
- Text Content

## Current Selection

Viewing all documents!

**content**  
ahora bueno creo decir gente gusta hacer jajaja ke  
mejor mucho nada puede q **solo** soy todos ver  
verdad vida

**gender**  
F M

## Search

(press ESC to close suggestions)

Edit query +

**country**  
AR CL CO DO **ES** HN **MX** NL PE PR SV  
US VE ZW

**age**  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

# Example

- After a query (country = MX)

## Current Selection

Viewing all documents!



## Current Selection

(x) AND country:MX



# Example

- After a query (country = MX)

## Current Selection

Viewing all documents!



Find the 10 differences !!

## Current Selection

(x) AND country:MX



# Example

## ▪ After a query (gender= Male)

### Current Selection

Viewing all documents!

content  
ahora bueno creo decir gente gusta hacer jajaja ke  
mejor mucho nada puede q **SOLO** soy todos ver  
verdad vida

gender  
F M

Search

(press ESC to close suggestions)

Edit query +

country  
AR CL CO DO **ES** HN **MX** NL PE PR SV  
US VE ZW

age  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

### Current Selection

(x) AND sex:M

content  
ahora bueno creo decir gente hacer jajaja ke mejor  
mucho nada puede q sea **SOLO** soy todos ver  
verdad vida

gender  
M

(press ESC to close suggestions)

Edit query +

country  
AR CL CO DE DO **ES** **MX** NI PE PR US  
UY VE YU

age  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

# Example

- After a query (gender= Male)

## Current Selection

Viewing all documents!

**content**  
ahora bueno creo decir gente gusta hacer jajaja ke  
mejor mucho nada puede q **SOLO** soy todos ver  
verdad vida

**gender**  
F M

Search

(press ESC to close suggestions)

Edit query +

**country**  
AR CL CO DO **ES** HN **MX** NL PE PR SV  
US VE ZW

**age**  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

Find the 10 differences !!

## Current Selection

(x) AND sex:M

**content**  
ahora bueno creo decir gente hacer jajaja ke mejor  
mucho nada puede q sea **SOLO** soy todos ver  
verdad vida

**gender**  
M

(press ESC to close suggestions)

Edit query +

**country**  
AR CL CO DE DO **ES** **MX** NI PE PR US  
UY VE YU

**age**  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

# The Information is on the differences

- What do we want to know:
  - Which are the terms that use the Males and/or people from Mexico ?
  - OR
  - Which are the terms that better identify or differentiate them?

# Example

- After a query (country = MX)

## Current Selection

Viewing all documents!

content  
ahora bueno creo decir gente gusta hacer jajaja ke  
mejor mucho nada puede q solo soy todos ver  
verdad vida

gender  
F M

Search

(press ESC to close suggestions)

Edit query +

country  
AR CL CO DO ES HN **MX** NL PE PR SV  
US VE ZW

age  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

Substract them !!

## Current Selection

(x) AND country:MX

content  
bueno creo decir eres gusta hacer jajaja ke mejor mucho  
nada q quien sea solo soy todos va ver verdad

gender  
F M

Search

(press ESC to close suggestions)

Edit query +

country  
**MX**

age  
10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

# Example

- After a query (country = MX)

## Faceted Results: The most used words by the mexicans



## Statistical Results: The words that differentiate the mexicans from the rest of the world



# Example

- After a query (gender = M)

## Faceted Results: The words that Males use most

### Current Selection

(x) AND sex:M

(press ESC to close suggestions)

Edit query +

#### content

ahora bueno creo decir gente hacer jajaja ke mejor  
mucho nada puede q sea **SOLO** soy todos ver  
verdad vida

gender  
M

#### country

AR CL CO DE DO ES **MX** NI PE PR US  
UY VE YU

#### age

10 100 105 **15 20 25** 30 35 40 45 50 55  
60 65 70 75 80 85 90 95

## Statistical Results: The words that differentiate the males from the females

(x) AND sex:M

Edit query +

#### content

2 5 algunos asi cualquier debe dios **ejemplo** gran **historia** kon  
lugar madre metal **mexico** mientras mierda **neta** of palabras  
**pinche** poder punto seguro **wey**

gender  
M

#### country

AF AQ BO **CN CX CZ** DE EG GR  
**HM** IE IL LY MN **MX** NI NO PK  
TG TJ **UM UY VN YU**

#### age

10 100 105 15 20 25 30 35 40 45  
**50 55 60 65 70 75 80 85 90**  
95

# The information is on the difference

- We can compare a subset with the full set:
- But we can compare different subsets:

## Current Selection

Viewing all documents!

## Search

(press ESC to close suggestions)

Edit query +

## Content

Male

ahora bueno cosas creo decir dios eres gente  
gusta hacer jajaja ke mejor mismo mucho mundo  
nada otra otro puede q quien sea solo soy  
todos va ver verdad vida

Female

ahora alguien amor bueno cosas creo d gente  
gracias gusta hacer jajaja jajajaja k ke mejor mucho  
nada puede q s sea siempre solo soy todos va ver verdad  
vida

## Country

Male

AQ AR BO CA **CL CO** CU DE DO EC ES FR GT MX NI PE  
PK UM US UY **VE** VN YU

Female

**AR** BA BO CL CO CR CU DE DO EC ES FR GT **HN** IT JP LU  
MX **NL** PA PE PH **PR** SV UK US **VE ZW**

## Age

Male

15 20 25 30 **35 40 45 50** 55 60 65 70  
75 85 **90 95**

Female

10 **100 105** 15 20 25 30 35 70 75 80 95

# The information is on the difference

- We can compare by gender, doing different queries

## Current Selection

(x) AND rAge:45

### Content

Male

ahora bueno CREO **gane** gracias gusta mejor **metal** mundo n otra  
otro quien **saludos** siempre todos va verdad â

Female

ahora **alguien** **amigos** amor bueno **dios** eres espero  
**están** gracias **hacer** **ii k ke** mejor **mismo** **mucho** mundo  
nada otra **puede** **sea** solo soy **tener** **tienes** todos ver  
verdad **vida**

### Country

Male

BR CR DO **ES** FR IT MX PE VE

Female

AR CA CL CO DE **PH** PR US UY VE

### Age

Male

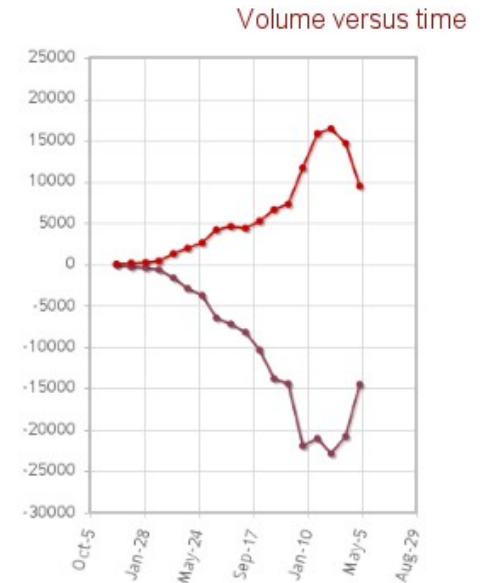
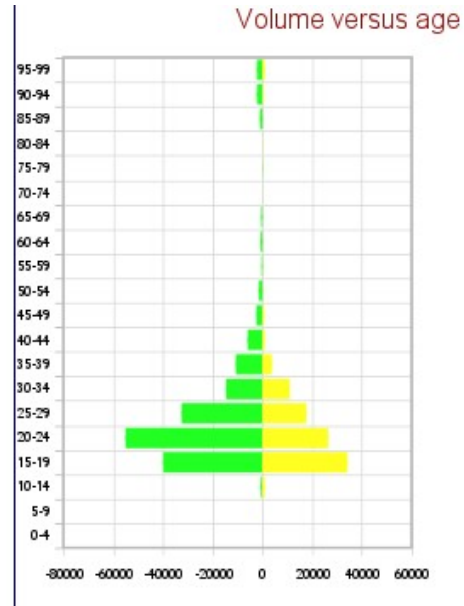
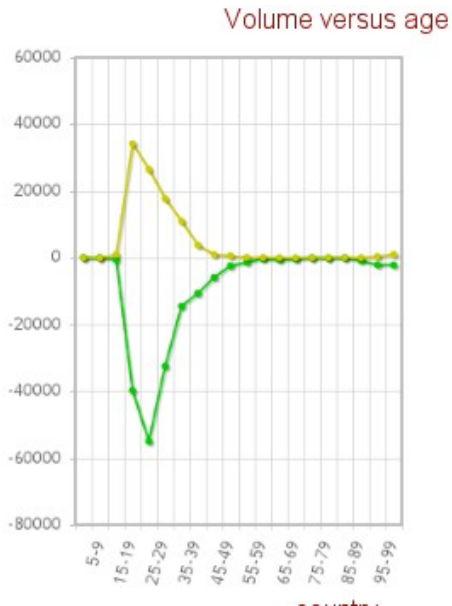
**45**

Female

**45**

# Data Charts

- We can apply differences to charts



# Data Charts

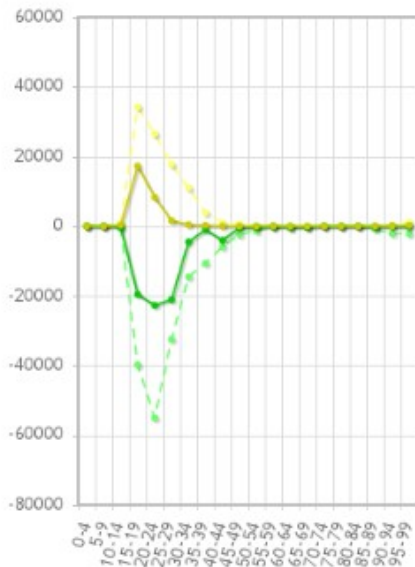
- We can apply differences to charts: And show the difference with the global set

(x)AND country:MX

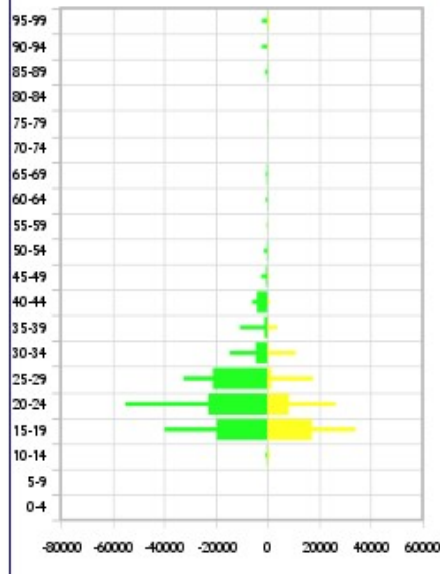
Search

Edit query +

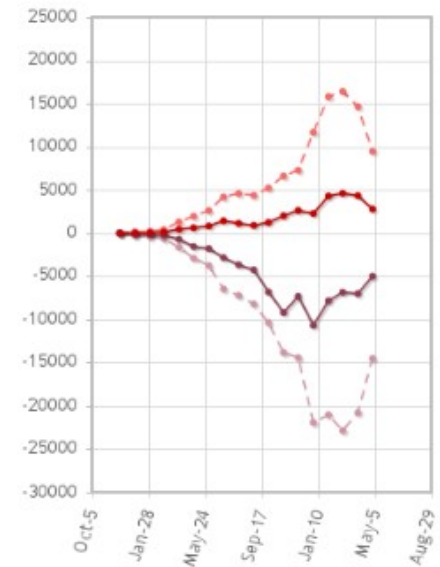
Volume versus age



Volume versus age



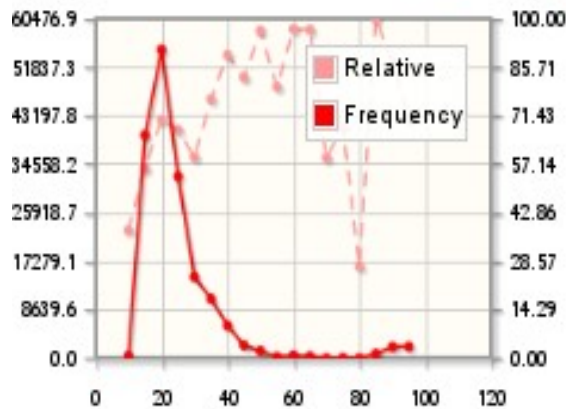
Volume versus time



# Data Charts

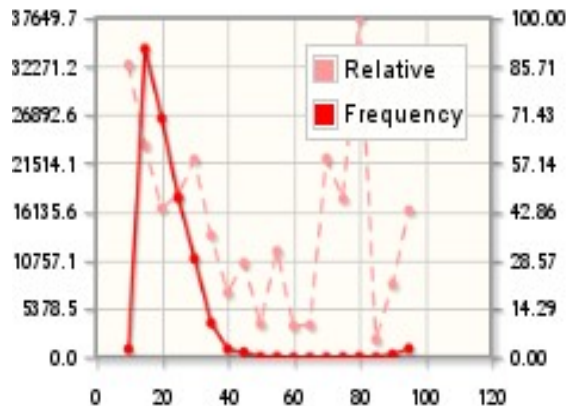
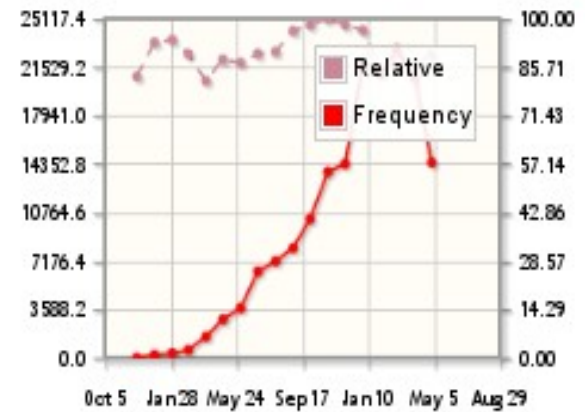
- Data Charts can show the differences

volume by age and gender

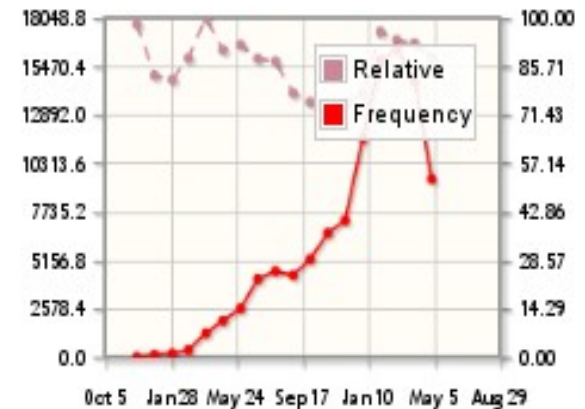


Males

volume by time and gender



Females



# Data Charts

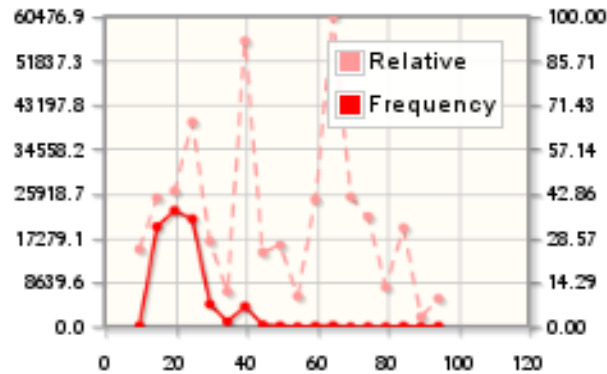
- Even through queries

Current Selection

(x)AND country:MX

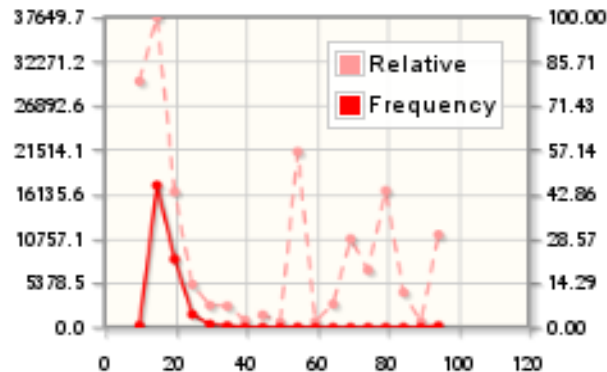
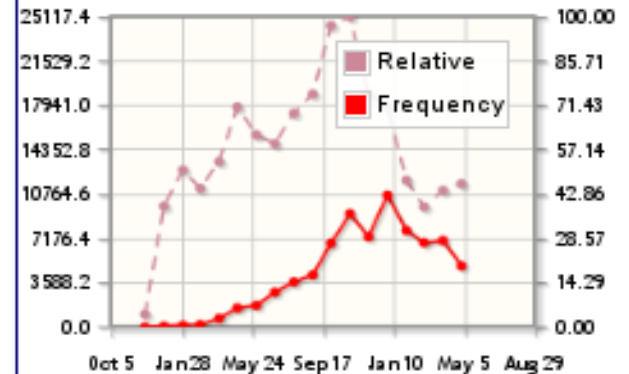
ESC  
(press ESC 1  
  
Edit query +

volume by age and gender

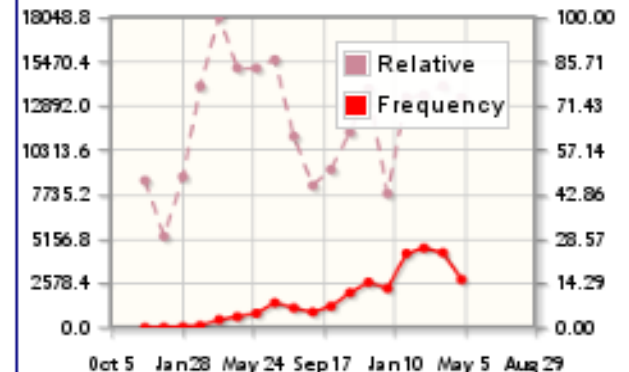


Males

volume by time and gender

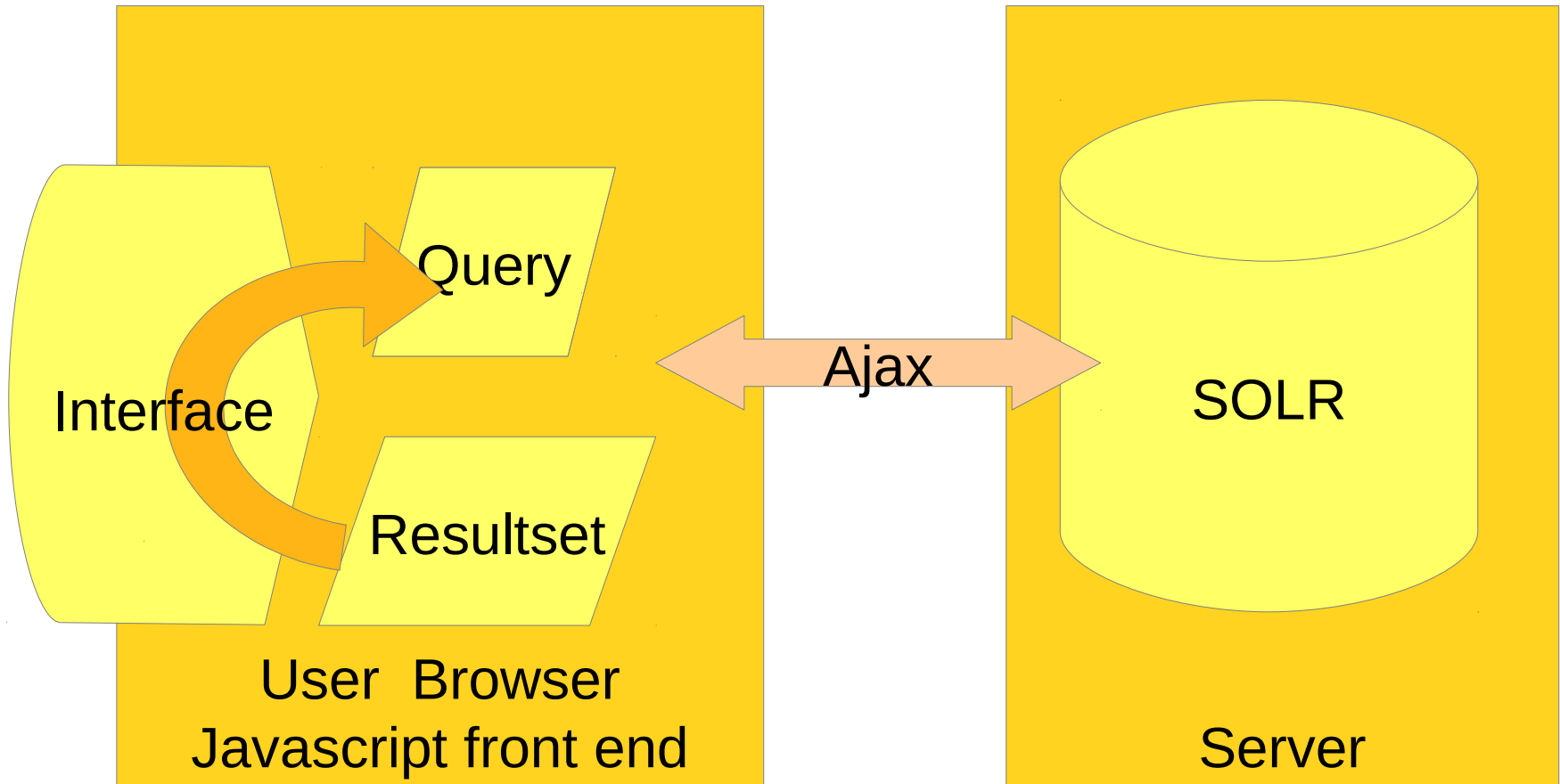


Females



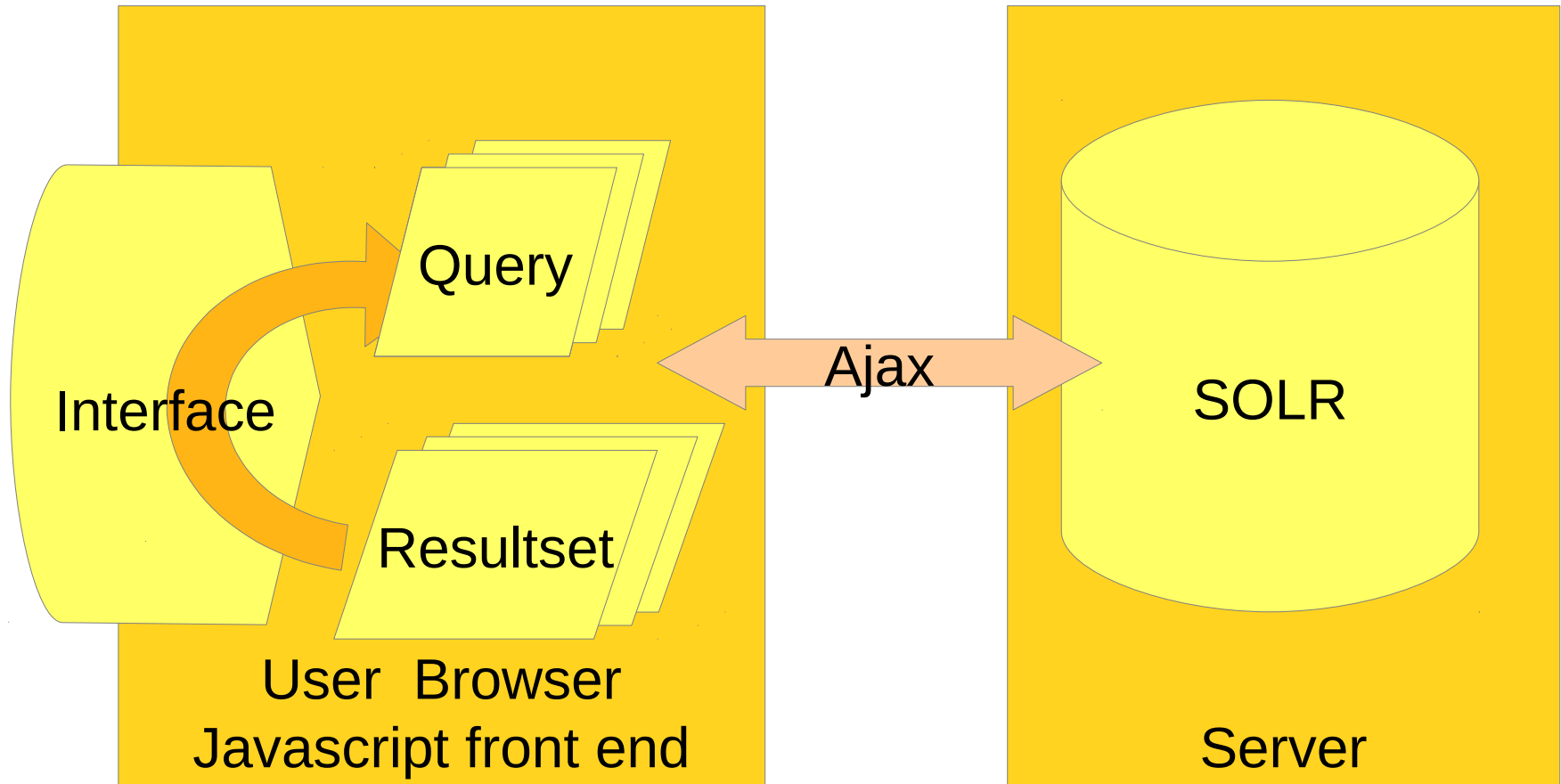
# Ajax Solr

- Single Query / Resultset



# Ajax Solr Extension

- Multiple Queries / Resultsets



# Ajax Solr Extension: Multiple Sets

- Each set has its own query
  - A query is composed of different query terms
  - The terms that define the set, can not be deleted
  - Query terms can be copied/deleted through the different sets
- Widgets can view/compare data from the different result sets.

# Ajax Solr Extension: Multiple Sets

Set 0

Independent widgets  
Base Data

Set 1

Widgets related to this set  
Own Filter Query

Set 2

Widgets related to this set  
Own Filter Query

# Conclusions

- Facets give basic statistics
- The information is on the differences
- Ajax Solr can be adapted to manage different sets of data
- Charts and graphics help us to have an overview of the data: information.

# Sources

The code is not ready to be published

- Was done for SolrJs
- Migrating to AjaxSolr (Lucene-Solr 4.0)

But it will be...

- Links
  - <http://www.barcelonamedia.org/personal/joan.cordinale/en>