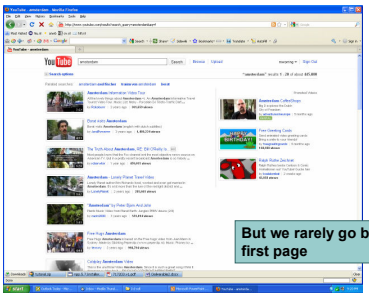
Multimedia Analytics: Exploration of Large Multimedia Collections

Marcel Worring

With contributions from
Ork de Rooij, Daan Odijk, Cees Snoek, Koen van de Sande

Informatics Institute
University of Amsterdam

The Internet



YouTube
every minute
>50 hours of video uploaded

But we rarely go beyond the first page

Broadcasting archives

BEELD EN GELUID

Sound and Vision Archive

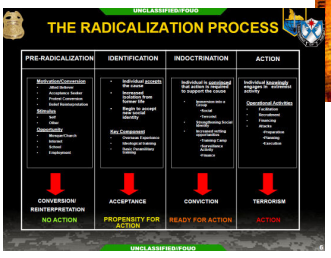
Yearly 15.000 hours video added

Forensics

10 Tb is a large set of potential locations for child porn, but how to find it?

We need to find all needles

Intelligence

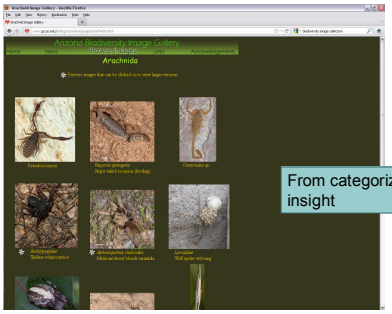


Hsdl.org

publicintelligence.net

Not obvious what your are looking, for any clue would help

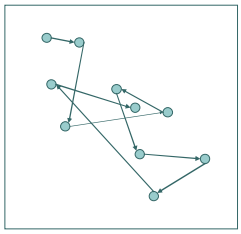
Science



From categorization to insight

The basis: ranking of data

Some query defines starting point and order

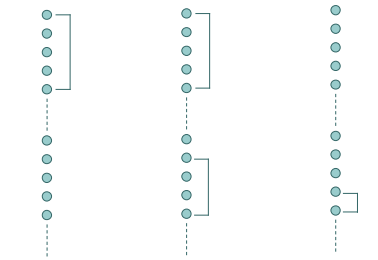


An image/video collection

Result

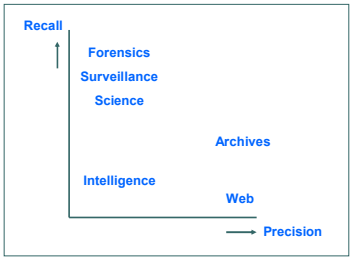
Tasks

Most of the techniques are based on providing a ranking, different applications different criteria of success.



Precision versus recall

Search requirements in retrieval vary



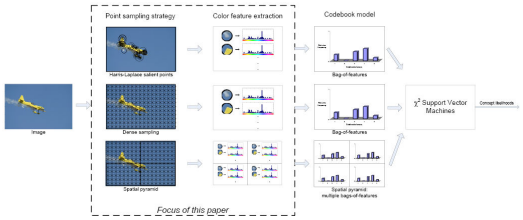
"Insight"

What is a category?

- Definition
  - A class or group of things, people, etc. possessing some quality or qualities in common; a division in a system of classification
- So for images
  - Could be in the content or in the metadata
- Current focus
  - The visual content

References: van de Sande, CIVR 2008, Snoek Trecvid 2008

Concept detection



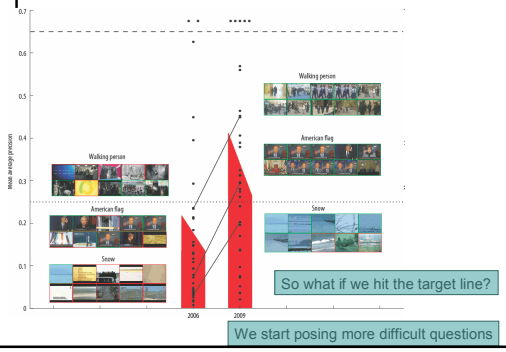
MediaMill

One of the best detection pipelines

50 - 500 concepts
Wildlife, Face,
People marching, ……

Ref: Snoek IEEE Computer 2010

Quality is steadily improving

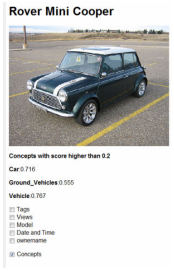


So what if we hit the target line?

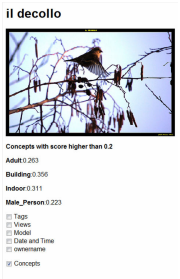
We start posing more difficult questions

Multimedia Analysis

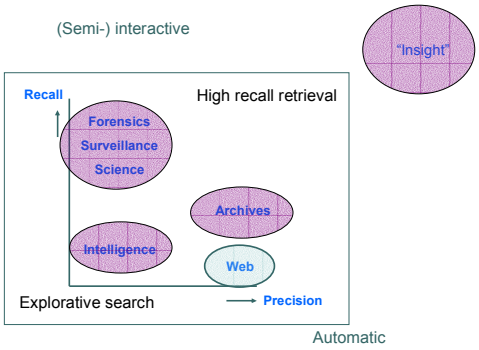
Sometimes it works



Sometimes it doesn't



Automatic versus interactive



Man versus machine

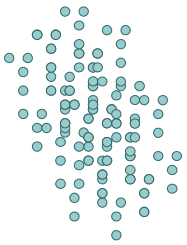
What humans are good at

Recognize

Interpret context

Associate

What machines are good at



Bulk processing

Multimedia Analytics

- Aims
  - To combine the best of both worlds
- For
  - Datasets that are too large to be handled by humans alone and to complex to be handled by machines only
- With the constraint
  - That we need to see an image before we can make a judgment on it

Chinchor, Thomas, Wong, Christel, Ribarsky CG&A 2010

Definition

Multimedia Analytics
=
Multimedia Analysis
+
Visual Analytics

Visual Analytics

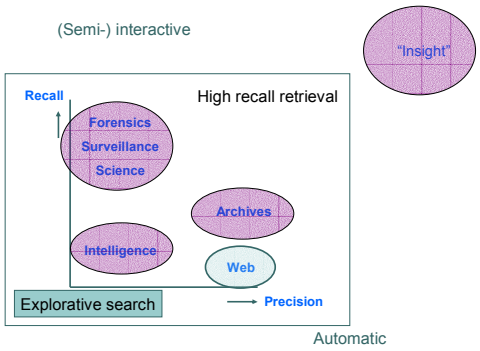
Visual Analytics is the science of analytical reasoning facilitated by **interactive visual interfaces**

The field started as a result of 9/11

An overwhelming amount of traces and information.

But the scope of visual analytics is much broader and applies to any scenario where large data collections have to be used.

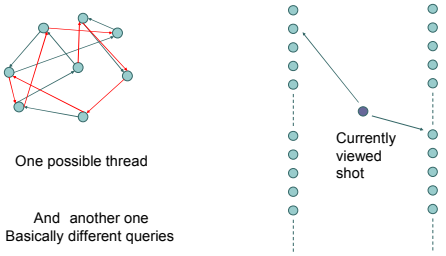
Automatic versus interactive



Threads

- Definition
  - A thread is a linked sequence of shots in a specified order, based upon an aspect of their content.
- Static threads
  - Are pre-computed
- Dynamic threads
  - Are created on the fly

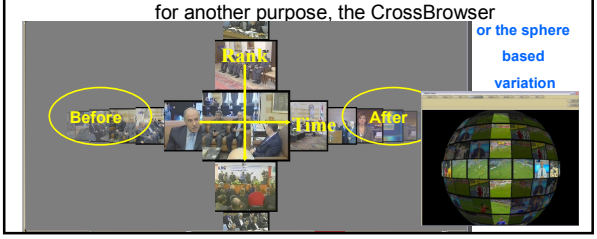
Navigation support: Threads



Snoek, Worring, Koelma, Smeulders
IEEE Trans MM 2007

Beyond the query result

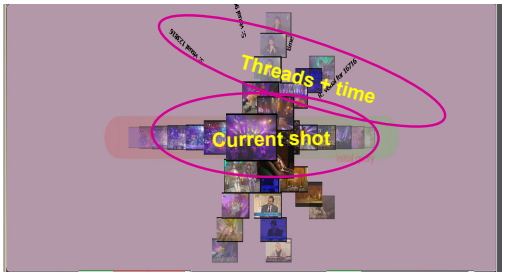
- When I pose a (semantic) query
  - The result is a one dimensional ranked list
  - So one dimension of the display can be used for another purpose, the CrossBrowser



de Rooij, Worring IEEE Trans MM 2010

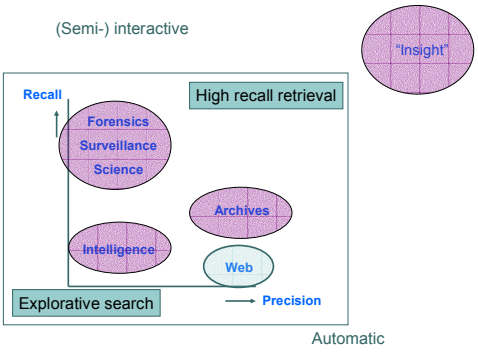
The RotorBrowser

Even more directions to go through



Very good for explorative browsing

Automatic versus interactive



The guidelines we learned

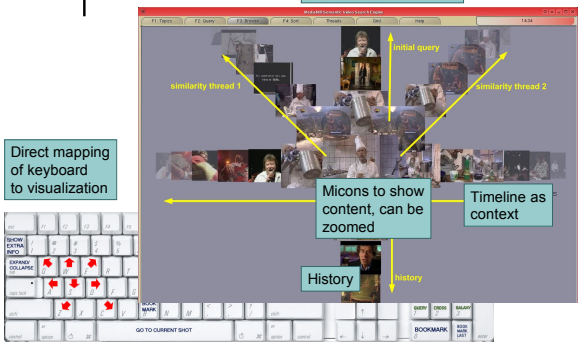
- Guideline 1
  - The user has to be able to efficiently inspect the setting, scene, and all objects and their motions within a shot for relevance and the user should be able to adapt the level of detail given according to the complexity.
- Guideline 2
  - The user has to be able to view the temporal context of video shots in order to determine the relevance of the current shot.
- Guideline 3
  - The user must be able to label multiple shots at the same time, when this is more efficient.
- Guideline 4
  - The system should show unexplored areas within the collection.

Guidelines

- Guideline 5
  - The system should always yield a default navigation path.
- Guideline 6
  - The user has to be able to efficiently inspect the potential relevance of the navigation options in the navigation context.
- Guideline 7
  - The system should aid the user to return to earlier navigation options by visual or spatial means.
- Guideline 8
  - The interface should not switch between different panes.
- Guideline 9
  - The interface must use a clear mapping between navigation and visualization.

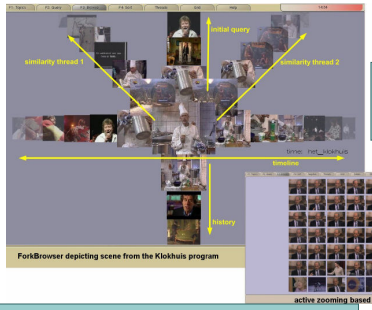
ForkBrowser

Navigation possibilities



Direct mapping of keyboard to visualization

Active Zooming



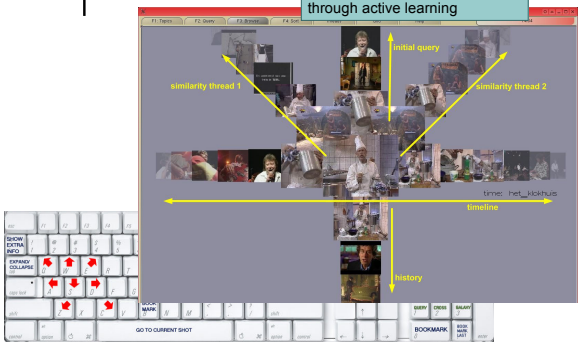
Quick inspection of a large part of a thread

1. Inspecting navigation possibilities
2. Label many at the same time

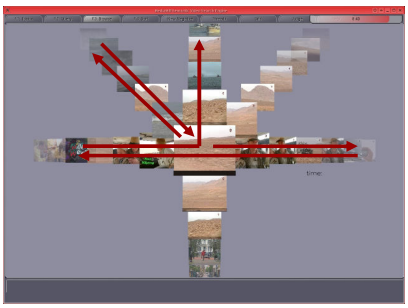
The only screen switch in the system, but different task

ForkBrowser

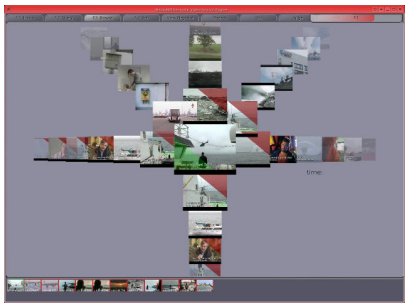
Navigation possibilities obtained through active learning



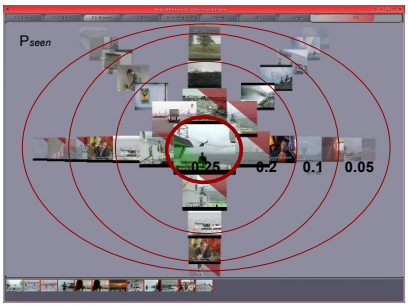
Browsing through results



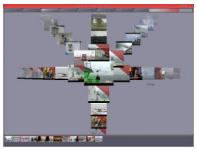
User marks relevant results



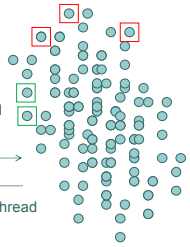
Negative results marked automatically



Update navigation path



Uncategorized images



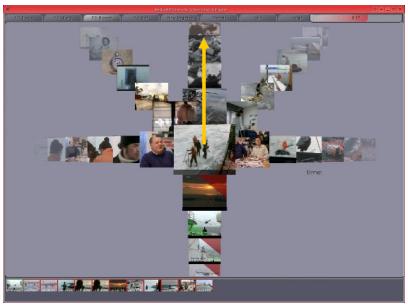
Learn and apply model for whole dataset

New ranking of current thread

After selecting positive and negative elements

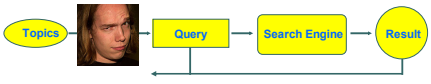
and a time-span with no results selected

Refined default navigation path



Interactive Benchmarking

- TRECVid
  - Search topics defined by NIST
  - Participants have fixed amount of time to search through the collection



Visual evaluation of the result by NIST

Experimental setup

- Dataset: TRECVID 2008
  - 200 hours of video
  - 35766 individual shots
- Three experiments
  - Study of potential benefit of threads
  - Determining the potential of relevance feedback in this dataset
  - Trecvid participation

Example tasks in TRECVID

Task 1: Find pictures of "Job Cohen"

Only 3 clusters: full frontal face, 2 people, and table of people.

Type B - specific visual example.

Task 2: Find boats or ships

Much variety: top-down, closeup, inside boats, with horizon.

Type C - generic visually diverse w/ concept "boat-ship".

Task 3: Find closeups of hands

Extremely high variety of results, both closeup and further away.

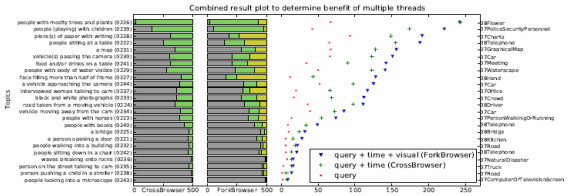
Type C - generic visually diverse w/ concept "hands".

Task 4: B/W pictures of airplanes

In sky and on ground, mostly from war documentaries, bad quality, many angles.

Type D - generic visually diverse, no direct concept match.

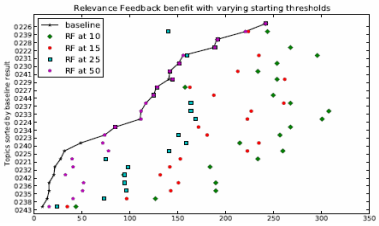
Evaluation



Grey: Query
Time: Green
Visual similar: Yellow

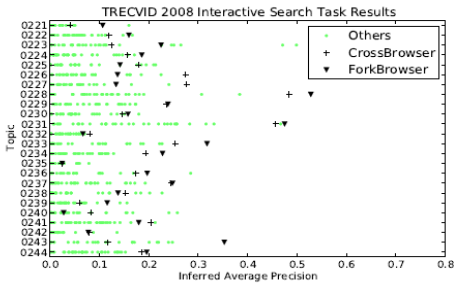
But first using simulated users
-Judge shot based on ground truth
-Follow thread with most results if possible
-Otherwise follow query thread

Evaluation

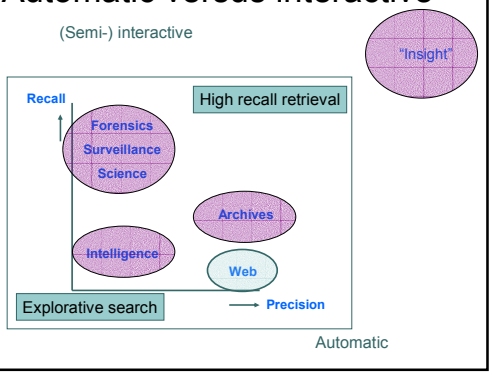


Activating RF at X indicating that X times a non-relevant item was found (500 User Interaction Steps).

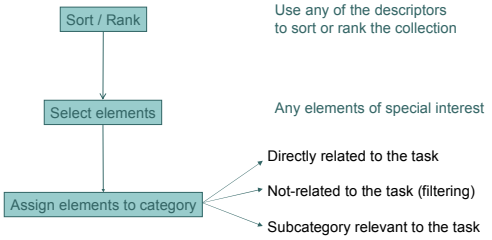
And @ TRECVID



Automatic versus interactive



High recall search processes

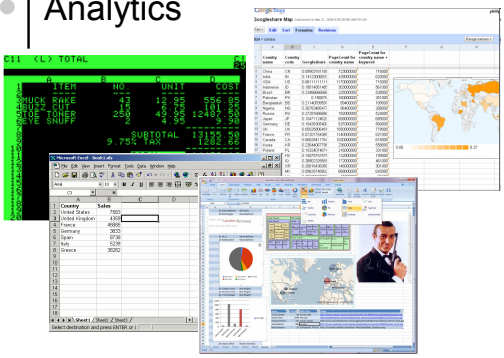


An Analytics Approach

- What is the best known Analytic tool?

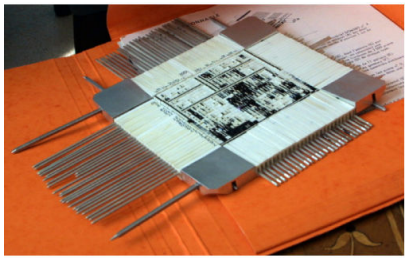
Yes Microsoft Excel

Analytics



Fischer et.al, TVCG 2010.

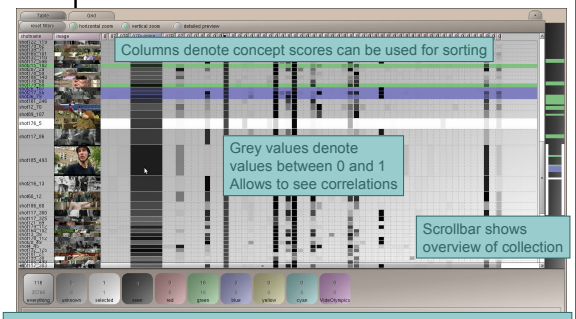
And even earlier



Bertin: Domino device

de Rooij, van Wijk, Worring, CG&A 2010
de Rooij, Worring. IEEE Transactions on Multimedia (in press)

MediaTable

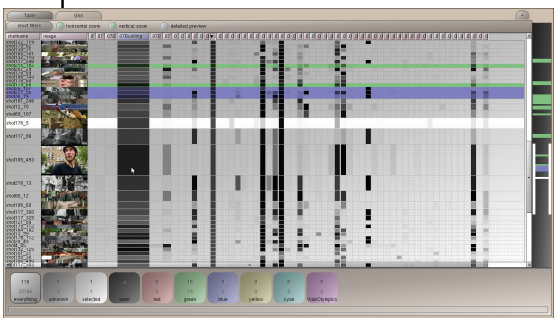


Colors denote categories and buckets are used to collect elements of (sub-) category

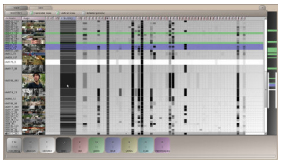
Three interactive strategies

- Fully interactive
  - User is interactively performing the sort/select/categorize process
- Manual relevance feedback
  - In addition to the above the user can perform relevance feedback on any of the categories
- Unobtrusive relevance feedback
  - In addition to the above the system automatically indicates new potentially relevant elements

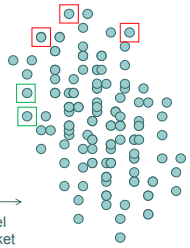
Fully interactive



On demand suggestions



Uncategorized images

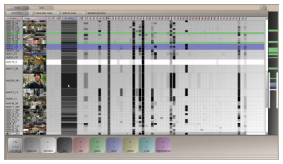


After categorizing some elements

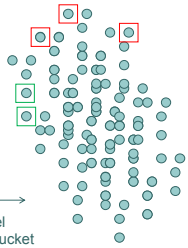
Learn and apply model for **user** selected bucket

Category suggestions

Unobtrusive assassistance



Uncategorized images

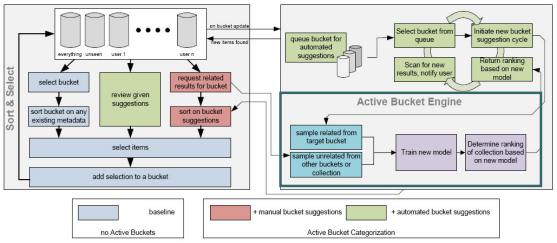


Continously observe what happens

Learn and apply model for **system** selected bucket

Category suggestions

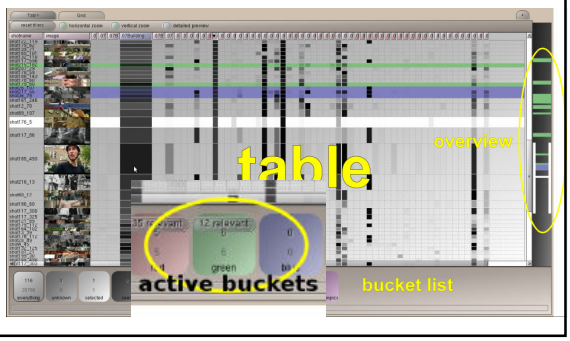
Active Buckets



Sampling and learning

- Two strategies
  - Nearest Mean: fast and simple
    - Compute mean of elements in bucket
    - Rank collection with respect to distance to the mean
  - Support Vector Machines: more expensive, more accurate
    - Dynamically learn new model from examples in the bucket and randomly selected (presumed) negatives

Active Buckets interface



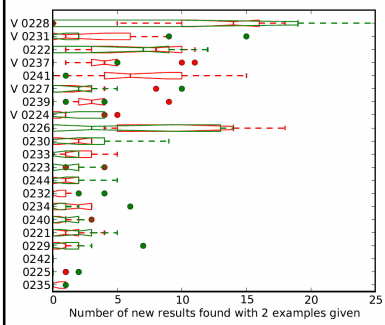
Experimental setup

- Dataset: TRECVID 2008
  - 200 hours of video
  - 35766 individual shots
- Two experiments
  - Determining the potential of relevance feedback in this dataset
  - User experiment comparing the three strategies

Experiment 1: Potential

- Protocol
  - 10 times random selection of N positive elements
  - Learning using the two strategies
    - Nearest Mean Classifier
      - approx 0.5 seconds per bucket
    - Support Vector Machines
      - approx 2-6 seconds per bucket
- Measure
  - Number of new correct examples in the top 50 ranked results

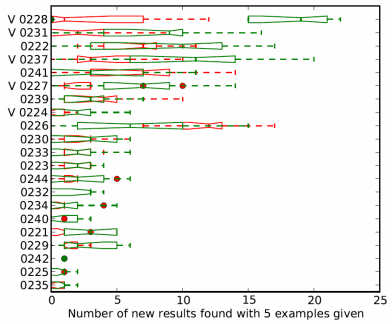
Learning experiment: N=2



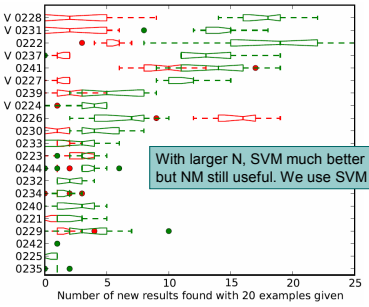
Nearest Mean

SVM

Learning experiment N=5



Learning experiment N=20



With larger N, SVM much better performance, but NM still useful. We use SVM

User experiment

- 21 student users in three groups
  - Baseline
    - Interactive retrieval only
  - Passive buckets
    - Bucket expansion on demand only
  - Active buckets
    - Unobtrusive bucket expansion
- 5 minutes per topic

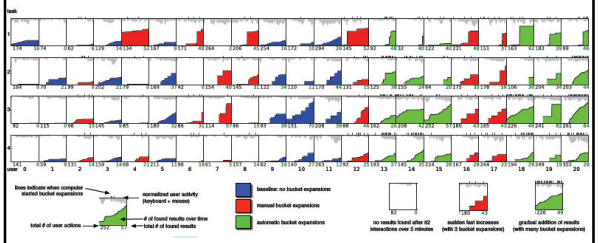
Results: elements found

Task	Baseline	Manual	Automatic
1	13 ± 7	36 ± 19 *	46 ± 8 * °
2	24 ± 14	23 ± 19	32 ± 8 * °
3	27 ± 24	31 ± 18	53 ± 18 * °
4	13 ± 8	19 ± 7 *	28 ± 5 * °
Average	20 ± 16	27 ± 18 *	40 ± 15 * °
Participants	8	7	6

• significant at the p=0.01 level compared to baseline
o significant at the p=0.01 level compared to manual

Task 1: specific, high visual similarity
Task 2: generic visually diverse, concept available
Task 3: generic visually diverse, concept available
Task 4: generic visually diverse, no concept available

Results: individual users



Discussion: task 1

- Specific, visually similar
  - Active bucket users able to select available results in one go
  - When a few results are found active buckets provide the rest of the cluster

Task 1: Find pictures of "Job Cohen"

Only 3 clusters: full frontal face, 2 people, and table of people.

Type B - specific visual example.

Discussion: task 2, 3

- Generic, concept available
  - Presence of appropriate concept makes fully interactive retrieval easier
  - Task 2
    - Active buckets more results at once
    - In total same amount of results
  - Task 3
    - Manually triggered more at once, but active more results in total

Task 2: Find boats or ships

Much variety: top-down, closeup, inside boats, with horizon.

Type C - generic visually diverse w/ concept "boat-ship".

Task 3: Find closeups of hands

Extremely high variety of results, both closeup and further away.

Type C - generic visually diverse w/ concept "hands".

Discussion: task 4

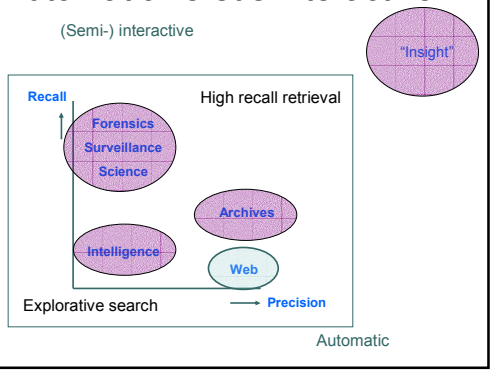
- Generic, no concept available
  - Having to select a combination of concepts is more difficult for users
  - User used passive buckets when not appropriate
  - Active buckets of clear benefit here

Task 4: B/W pictures of airplanes

In sky and on ground, mostly from war documentaries, bad quality, many angles.

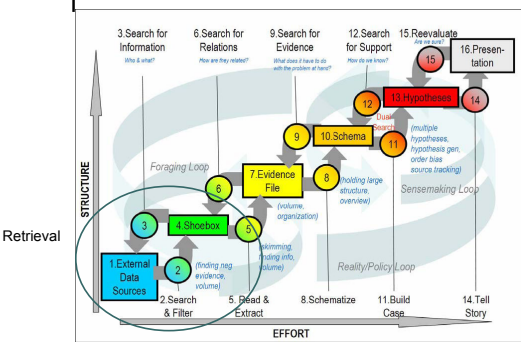
Type D - generic visually diverse, no direct concept match.

Automatic versus interactive



Piroli and Card 1995

The sense making process



Characteristics of insight

- *Complex*
  - Insight is complex, involving all or large amounts of the given data in a synergistic way, not simply individual data values.
- *Deep*
  - Insight builds up over time, accumulating and building on itself to create depth.
  - Insight often generates further questions and, hence, further insight.
- *Qualitative*
  - Insight is not exact, can be uncertain and subjective, and can have multiple levels of resolution.

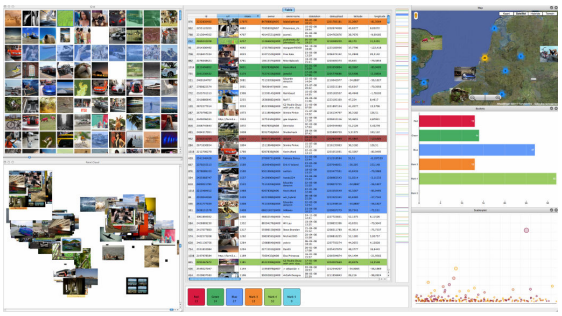
Characteristics of insight

- *Unexpected*
  - Insight is often unpredictable, serendipitous, and creative.
- *Relevant.*
  - Insight is deeply embedded in the data domain, connecting the data to existing domain knowledge and giving it relevant meaning.
  - It goes beyond dry data analysis, to relevant domain impact.

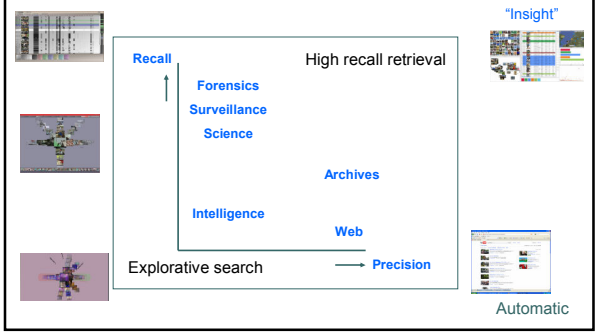
What is needed?

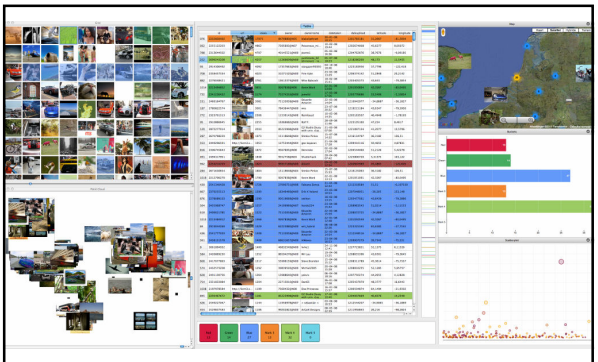
- Different ways
  - to explore the data
  - to view the data
- Categories
  - denoting current understanding
- Coordinated views
  - When a change is made in one view it should be reflected in the whole visualization

Towards insight



Conclusion





m.worring@uva.nl