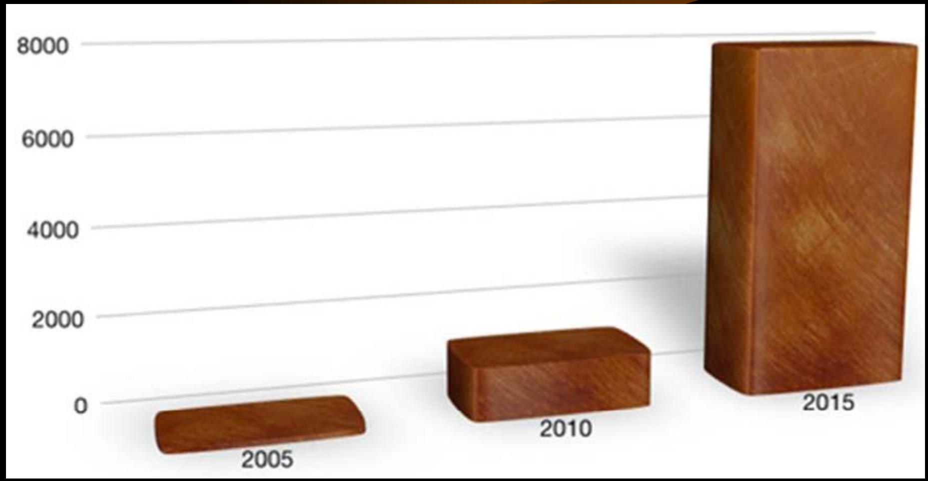
Augmenting Image Semantics Through Web Distances

Clement Leung
Computer Science Department
Hong Kong Baptist University
clement@comp.hku.edu.hk

The Exploding Digital Universe

- Estimated 1.8 zettabytes (10^{21} bytes) created and replicated in 2011
  - The number of bits is approaching the number of stars in the universe
- 1 zetta seconds = 31.71 trillion years = 2300 x age of universe
- The volume of seawater in the Earth's oceans is approximately 1.37 zettalitres
- The world's information is more than doubling every two years - growing faster than **Moore's Law**



A Decade of Digital Universe Growth: Storage in Exabytes

Sources: IDC's Digital Universe Study, sponsored by EMC, June 2011 and EMC Press Release

Dynamic Media Contents

- **Flickr**
  - ✓ > 4 million uploads per day
  - ✓ Total > 4 Billion photos and growing
- **You Tube**
  - ✓ > 200,000 uploads per day
  - ✓ > 100 millions downloads/day
- **Twitter**
  - ✓ Many posting have links to other media sources
- **Facebook**
- 3 million new photos/month
- 1 million photos delivered per second

Big Data – 3Vs and Images

- Big Volume
  - Images can be captured on many devices
  - Many medical images being produced
- Big Velocity
  - Images can be produced at a fraction of a second (unlike text documents)
- Big Variety
  - Images constitute an important data variety

Semantic Image Queries

Varying levels of search complexity

- Canadian sunset
- Woman playing piano
- Female patient having a certain condition
- Boy in red riding on an elephant in London Zoo on a summer evening

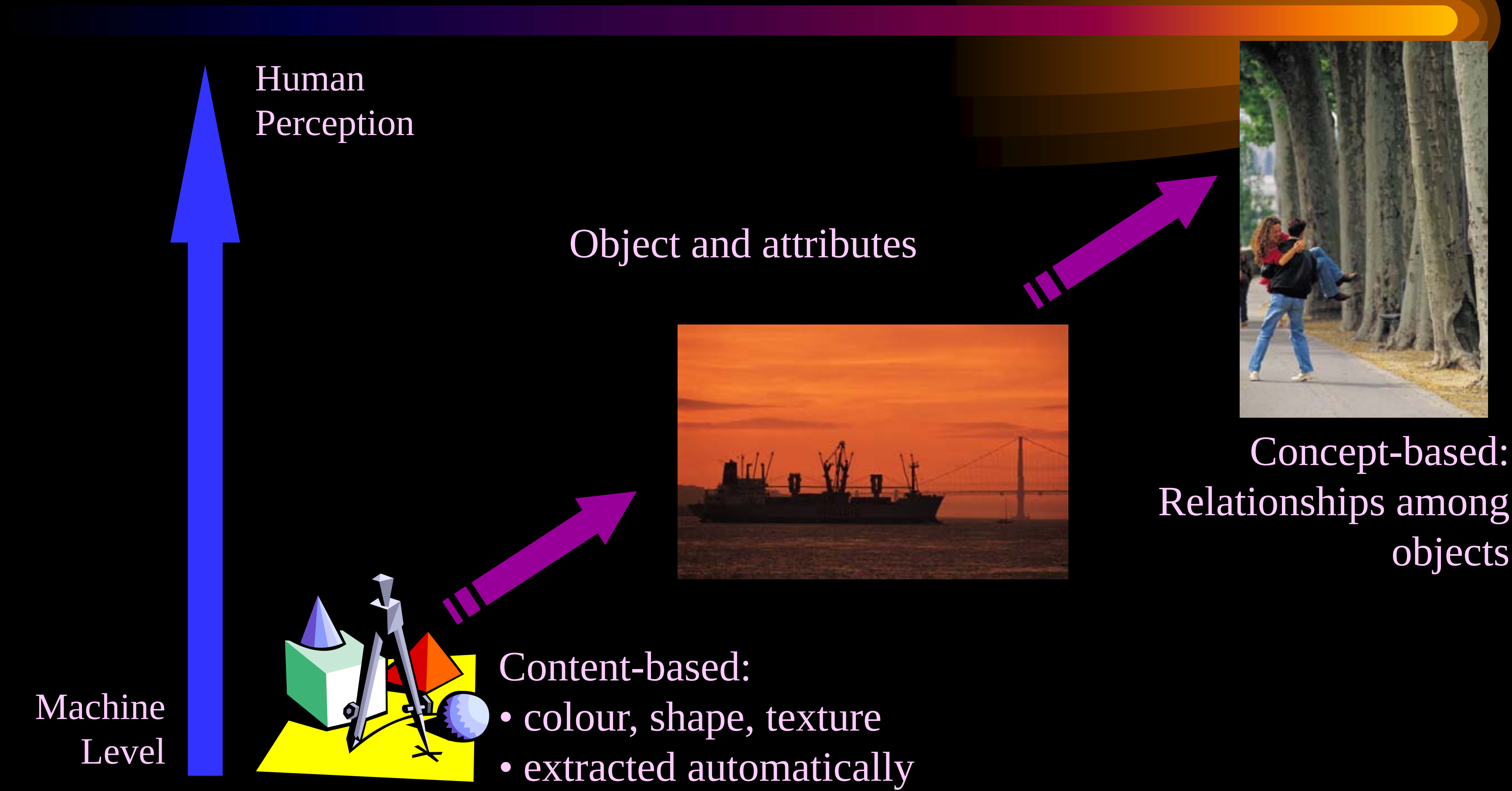
Semantics

- ***Wikipedia***
  - **Semantics** (from Ancient Greek: σημαντικός *sēmantikós*) is the study of meaning
  - It focuses on the relation between *signifiers*, like words, phrases, signs, and symbols, and what they stand for, their denotation
- Image objects and their relationships

Brute Force? Won't Work

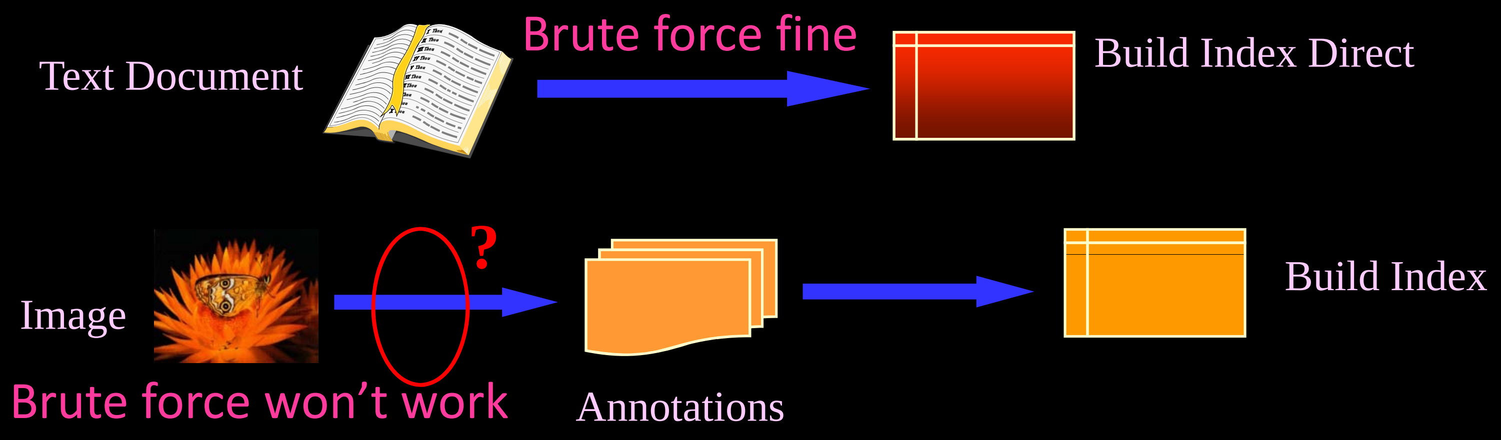
- Difficult for computers to recognise objects in images and videos

Semantic Gap: Content-based vs Concept-based Image Search



Semantic Search Depends on Indexing

- Computer vision
  - Too slow to deliver
- Dedicated intensive manual indexing
  - Costly, laborious and time-consuming
  - $\text{Rate}_{\text{creation}} >> \text{Rate}_{\text{indexing}}$



Indexing Cost & Semantics Tradeoff

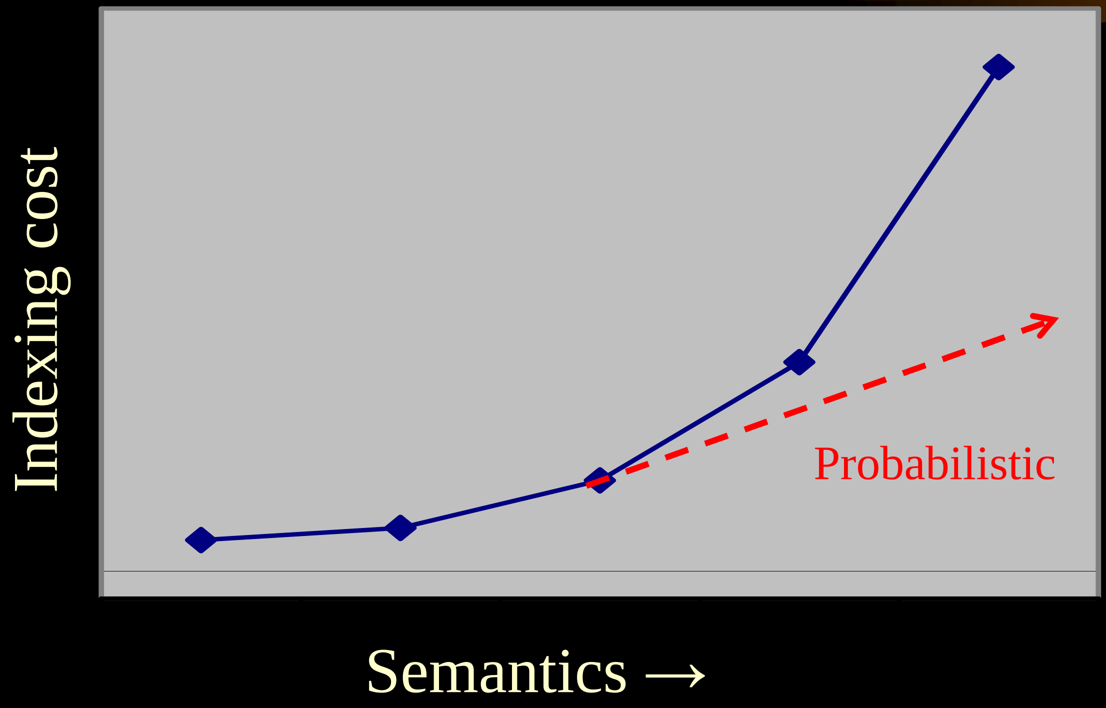


Image distribution in multi-dimensional space

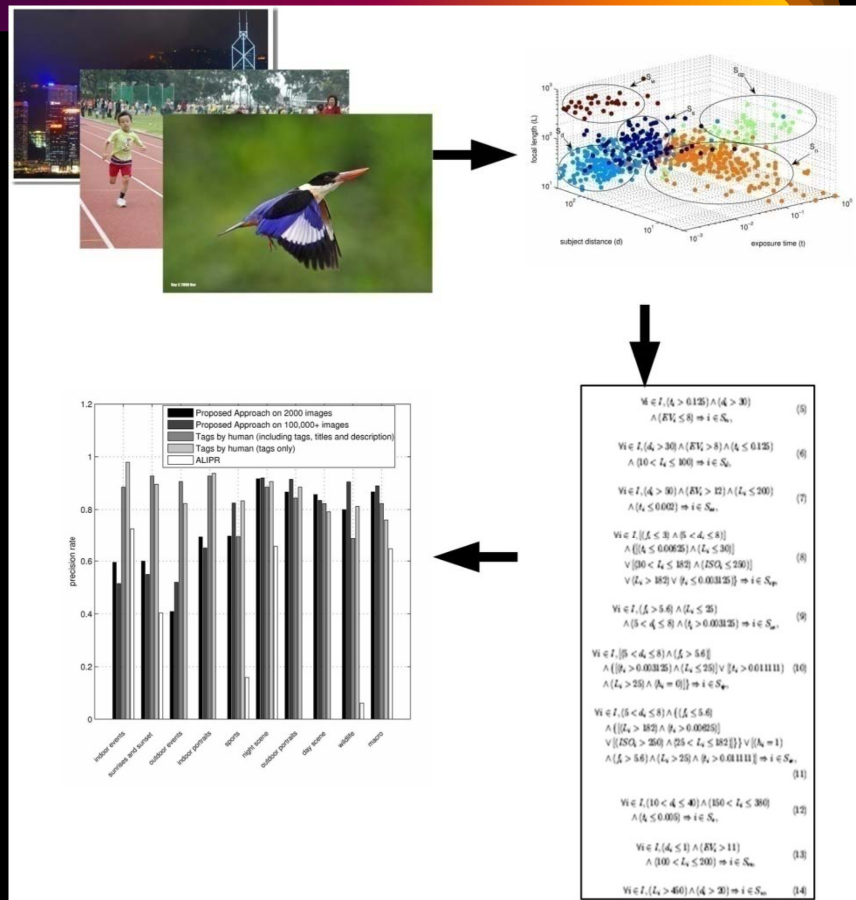
- An image I_i may be characterized by a number of dimensions $d_{i1}, \ldots, d_{ik}$ which correspond to the image acquisition parameters

$$I_i = (d_{i1}, \ldots, d_{ik})$$

- Each dimension has a certain domain D_j, i.e. $d_{ij} \in D_j$
- Each image corresponds to a point in k-dimensional space.

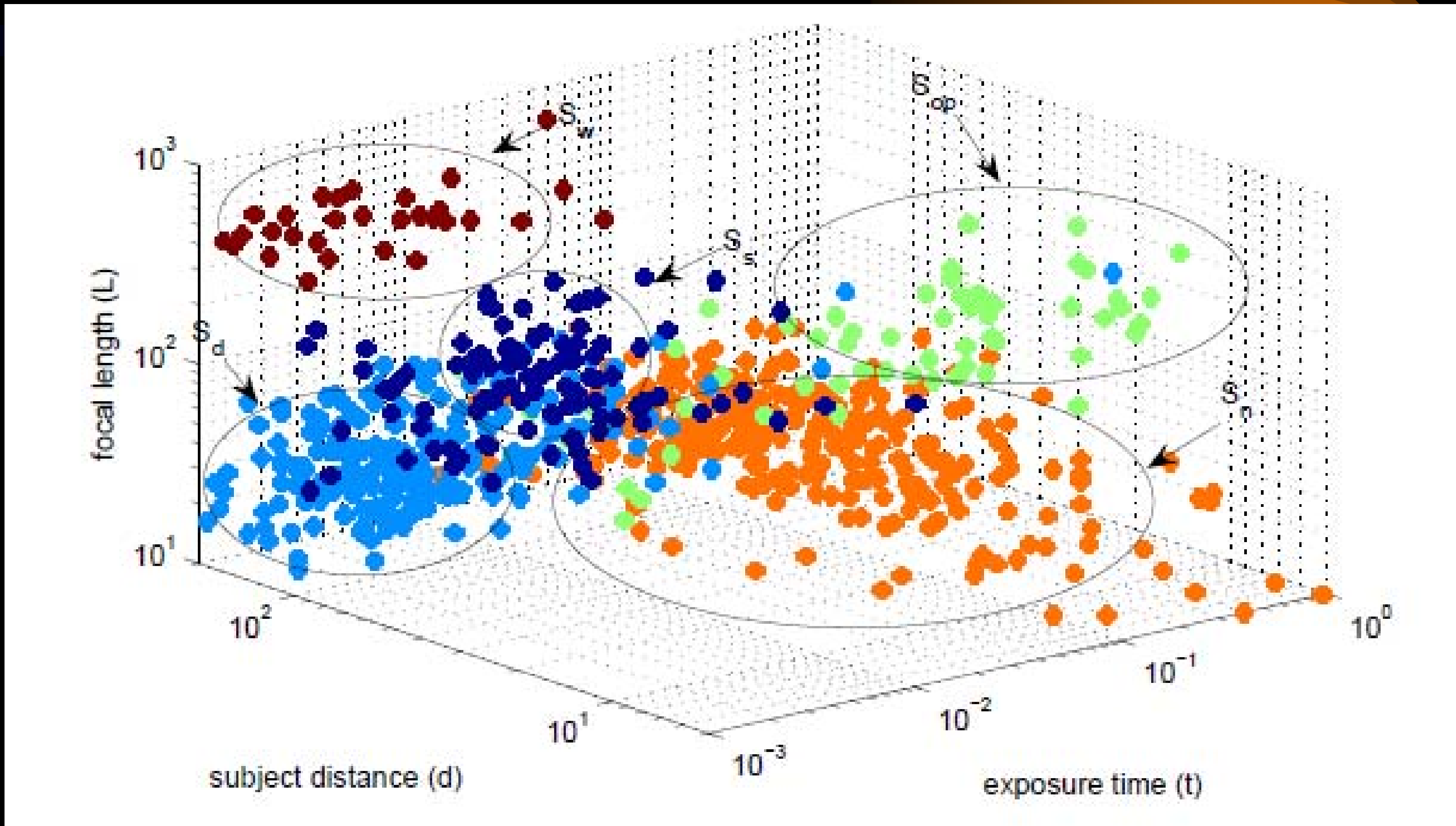
Automatic Semantic Annotation

- Rule-based approach to formulate annotations for images automatically



Source: R. C. F. Wong and C. H. C. Leung. Automatic Semantic Annotation of Real-World Web Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (11) : 1933 - 1944, November 2008.

Image distribution in 3-Dimensional space



Scene Classification (Tree Pruning)

From metadata…

- Aperture
- Exposure time
- Focal length
- Flash activation
- Subject distance

…to scenes of Images

Categories	Scenes	Symbols
Landscape	Day scenes	(S_d)
	Night scenes	(S_n)
	Sunrises and sunsets	(S_{ss})
Portraits	Indoor events	(S_{ie})
	Indoor portraits	(S_{ip})
	Outdoor events	(S_{oe})
	Outdoor portraits	(S_{op})
	Sports	(S_s)
Nature	Macro	(S_m)
	Wildlife	(S_w)

Rules for ASA

- night scene is having
  - the exposure time exceeding 0.125
  - the subject distance exceeding 30
  - the exposure value not greater than 8

$$\forall i \in I, (t_i > 0.125) \wedge (d_i > 30) \wedge (EV_i \leq 8) \Rightarrow i \in S_n,$$

- day scene is having
  - the subject distance exceeding 30
  - the exposure value greater than 8
  - the exposure time less than 0.125
  - The focal length in between 10 and 100

$$\forall i \in I, (d_i > 30) \wedge (EV_i > 8) \wedge (t_i \leq 0.125) \wedge (10 < L_i \leq 100) \Rightarrow i \in S_d,$$

Rule induction using C4.5 decision trees

Rule 1) $\forall i \in I, (t_i > 0.125) \wedge (d_i > 30) \wedge (EV_i \leq 8) \Rightarrow i \in S_n$

Rule 2) $\forall i \in I, (d_i > 30) \wedge (EV_i > 8) \wedge (t_i \leq 0.125) \Rightarrow i \in S_d$

Rule 3) $\forall i \in I, (f_i > 20) \wedge (d_i > 50) \wedge (EV_i > 11) \Rightarrow i \in S_{ss}$

Rule 4)
$$\forall i \in I, [(f_i \leq 5.6) \wedge (5 < d_i \leq 8)] \wedge \{[(t_i \leq 0.00625) \wedge (L_i \leq 30)] \vee [(30 < L_i \leq 182) \wedge (ISO_i \leq 250)] \vee (L_i > 182) \vee (t_i \leq 0.003125)\} \Rightarrow i \in S_{op}$$

Rule 5)
$$\forall i \in I, (f_i > 5.6) \wedge (L_i \leq 25) \wedge (5 < d_i \leq 8) \wedge (t_i > 0.003125) \Rightarrow i \in S_{oe}$$

Rule 6)
$$\forall i \in I, (f_i > 5.6) \wedge (0.003125 < t_i \leq 0.011111) \wedge (5 < d_i \leq 8) \wedge (L_i > 25) \Rightarrow i \in S_{ip}$$

Rule 7)
$$\forall i \in I, (5 < d_i \leq 8) \wedge \{(f_i \leq 5.6) \wedge \{[(L_i \leq 30) \wedge (t_i > 0.00625)] \vee [(ISO_i > 250) \wedge (30 < L_i \leq 182)]\}\} \vee [(h_i = 1) \wedge (f_i > 5.6) \wedge (L_i > 25) \wedge (t_i < 0.011111)] \Rightarrow i \in S_{ie}$$

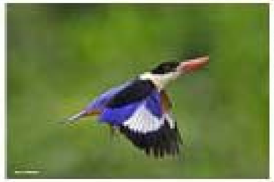
Rule 8)
$$\forall i \in I, (d_i > 10) \wedge (150 < L_i \leq 400) \wedge (t_i \leq 0.005) \Rightarrow i \in S_s$$

Rule 9) $\forall i \in I, (d_i \leq 5) \wedge (EV_i > 9) \Rightarrow i \in S_m$

Rule 10) $\forall i \in I, (L_i > 450) \wedge (d_i > 20) \Rightarrow i \in S_w$

Symbols: aperture (*f*), exposure time (*t*), subject distance (*d*), focal length (*L*) and flash activation (*h*)

Sample Annotations

	Landscape, night scenes, Victoria Harbor, Hong Kong, summer, night, sea, building		Portrait, indoor events, people, Cambridge, United Kingdom, spring, afternoon
	Nature, macro, animal, Taroko Natioanl Park, Japan, summer, morning leaf		Portrait, outdoor events, people, Cotton Tree Drive Marriage, Registry, Hong Kong, autumn, afternoon
	Portrait, sports, people, Yio Chu Kang Stadium, Singapore, summer, afternoon, motion		Nature, wildlife, animal, Orlando Wetlands Park, Florida, United States, autumn, afternoon, feather, motion
	Landscape, day scenes, Chaopraya, Bangkok, Thailand, spring, morning, sea, building, sky		Portrait, indoor events, people, The Mesa Arts Center, Mesa, Arizona, United States, summer, night, motion
	Landscape, sunrise and sunset, SaiKung, Hong Kong winter, evening sea, sky, wood		Nature, wildlife, animal, Wetland Park, Hong Kong, autumn, afternoon, sea, feather
	Portrait, outdoor events, people, Yunlin County Stadium, Taiwan, winter, afternoon		Portrait, sports, people, Wulihe Stadium, Shenyang, China, summer, afternoon motion

Top Matches to Semantic Queries

(a) night scenes (b) outdoor portraits (c) day scenes (d) wildlife (e) sports

Automatic categorization of different scenes of images obtained from Flickr

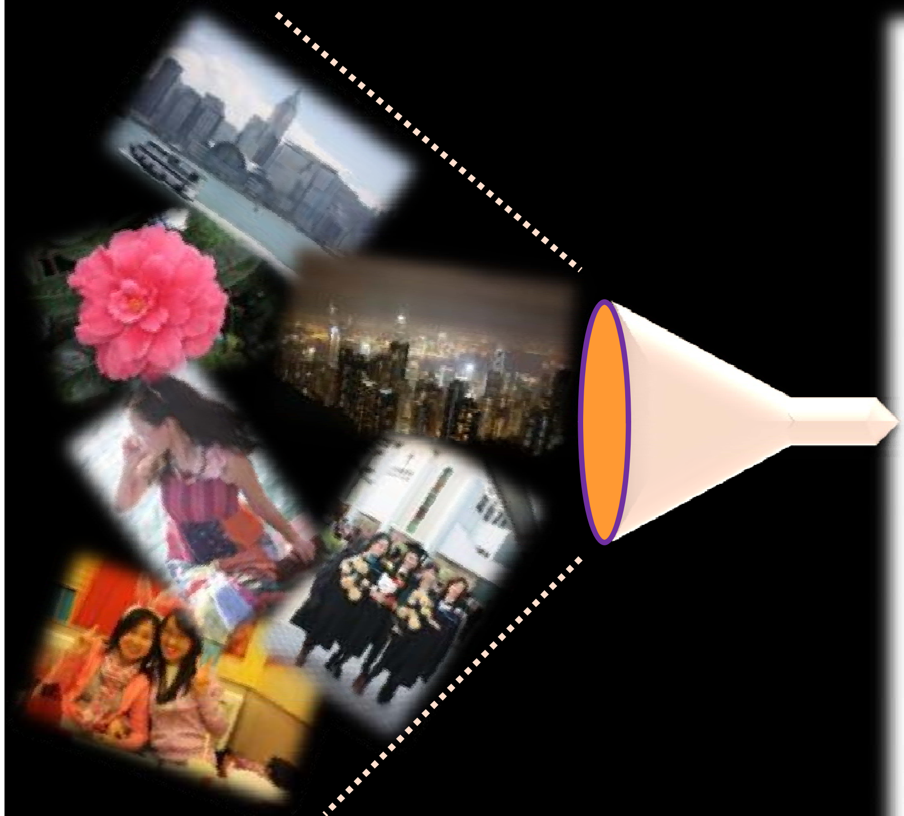


Figure 2. Each colum shows the different categories of "Micro", "Night Scenes", "Indoor Activities", "Day Scenes", "Portraits" and "Outdoor Activities"

Comparison with Human Tags

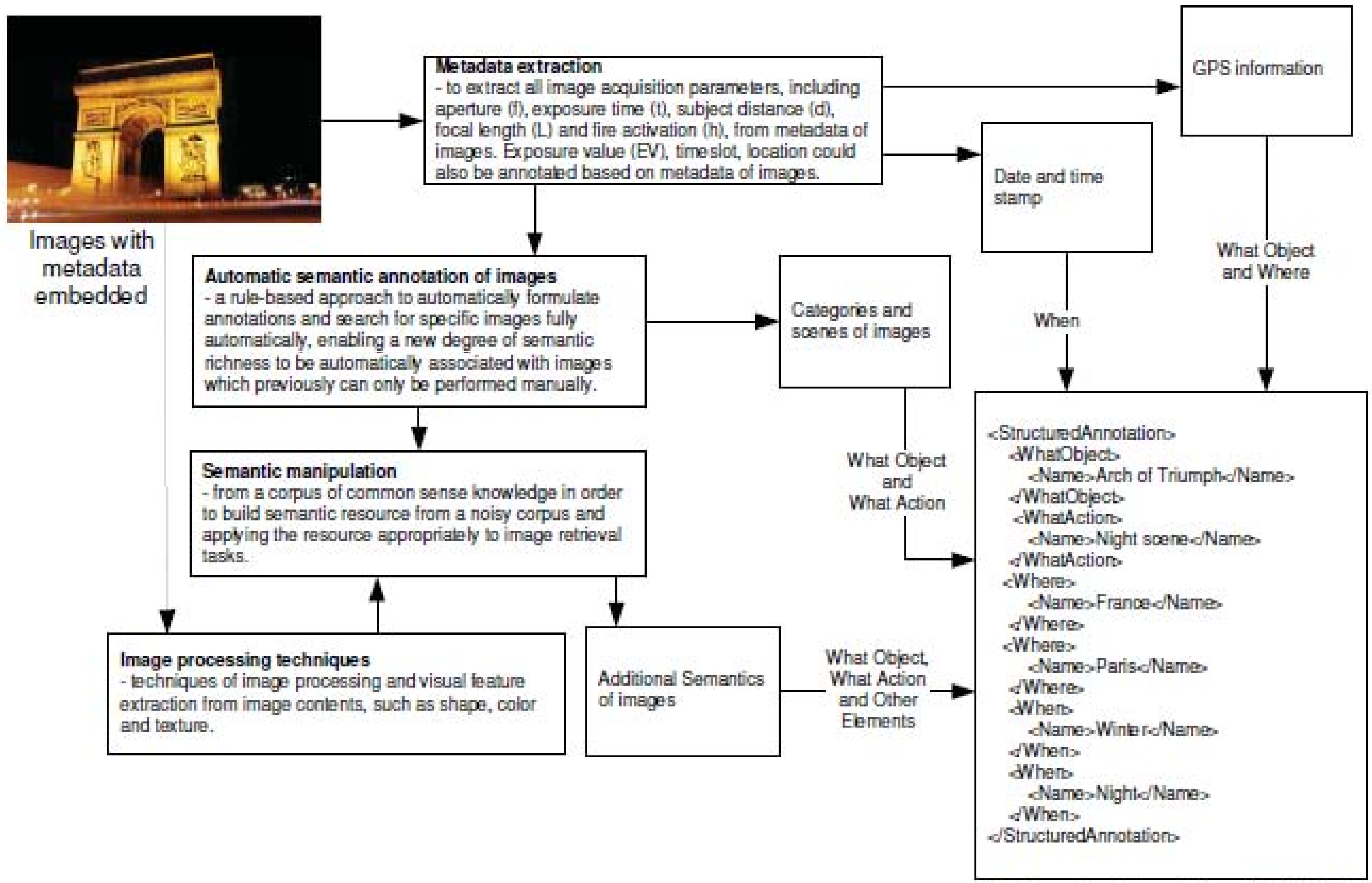
Scenes	Proposed approach	Human tags (including tags, titles and description)
S_m	89.00%	82.29%
S_{ss}	55.00%	92.71%
S_s	82.50%	69.79%
S_n	92.00%	88.54%
S_{op}	91.50%	84.38%
S_w	90.50%	68.75%
S_d	83.50%	82.29%
S_{ie}	51.50%	88.54%
S_{oe}	52.00%	90.63%
S_{ip}	65.00%	92.71%

Relationship with MPEG-7

- Landscape, night scenes, sea, building
  - MPEG-7 WhatObject
- Victoria Harbour, Hong Kong
  - MPEG-7 Where
- Summer, night
  - MPEG-7 When

- Nature, wildlife, animal, feather, motion
  - MPEG-7 WhatObject
- Orlando Wetlands Park, Florida, United States
  - MPEG-7 Where
- Autumn, afternoon
  - MPEG-7 When

Automatic Generation of MPEG-7 Semantic Descriptions



Automatic Generation of MPEG-7 Semantic Descriptions

Deeper Semantics

- Objective factual description
- Interpretation of the objects in the picture, and is based on prior knowledge
- Familiarity with the subject matter

Augmenting Image Semantics Using Collaborative Intelligence

- Collective Intelligence for Advanced Multimedia Semantics
  - Perception & interpretation of semantic multimedia content
    - Depends on user knowledge & experience
    - Degree of content richness can be distinguished into 3 levels:

      - **Primary Level** — Objective description of content
      - **Secondary Level** — Basic level of interpretation/summary of some/all entities
      - **Tertiary Level** — Knowledge-based elements

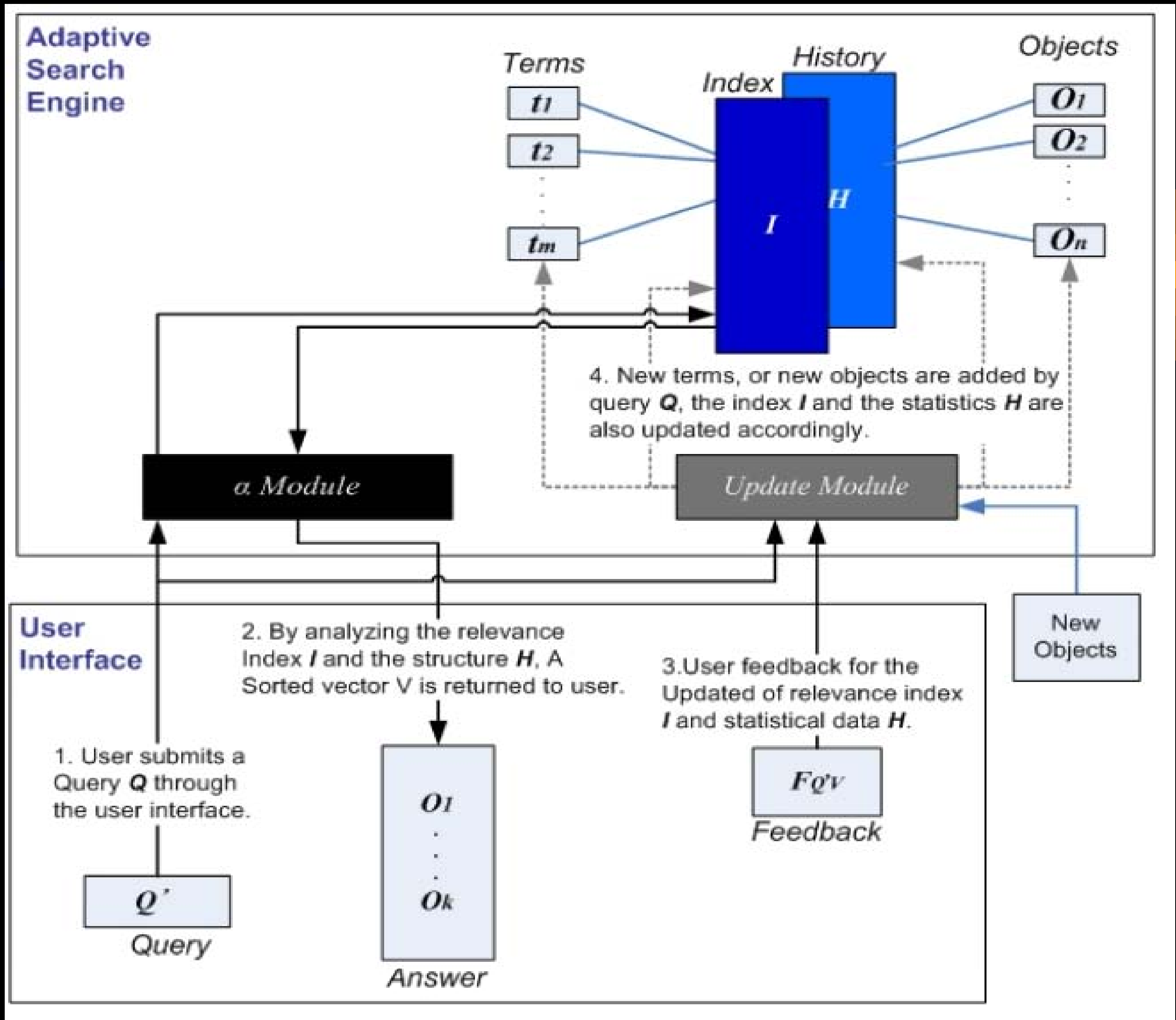
    - Information encoded in a multimedia object can be potentially unlimited

Collaborative Evolutionary Indexing: Indexing Through Usage

- More than 2 billion people are connected to the Internet in 2011
- 20 sec of their time is more than 10 million man hours
- Empire State Building took 7 million man hours to build
- 4 Empire State Buildings built every minute

Collaborative Evolutionary Indexing: Indexing Through Usage

- Users spend a lot of time on searching and viewing information
  - Exploits visual judgement and perceptive intelligence of web users
- An evolutionary indexing paradigm
  - Capture, analyze, interpret user behaviour and response
  - Support semantic visual information search through selection scoring & incremental indexing
  - Allows semantic concepts to be gradually discovered and migrated through an index hierarchy
  - Rich semantics
  - Robust and fault-tolerant



Source: C. H. C. Leung, W. S. Chan, J. Liu, A. Milani, Y. Li "Intelligent Social Media Indexing and Sharing Using an Adaptive Indexing Search Engine" *ACM Transactions on Intelligent Systems and Technology*, 2012 Vol.3, No.3

Score Updating Algorithms

- User input search query $Q(T_1, T_2, \ldots, T_m)$
  - Search result: *n* multimedia objects $O_1, O_2, \ldots, O_n$
- Increase the score when
  - User select O_x in the query result list
  - Receive +ve feedback from user
  - Relevant index scores increased by Δ_x
  - Possible promotion of index term *T* to the next higher level
- Decrease the score when
  - User don't select any O_x in the query result list
  - -ve feedback from user
  - Relevant index scores decreased by Δ_y
  - Possible dropping *T* to the next lower level
- Adding a new term when
  - Term is present in the query but not in the hierarchy, and user selects object

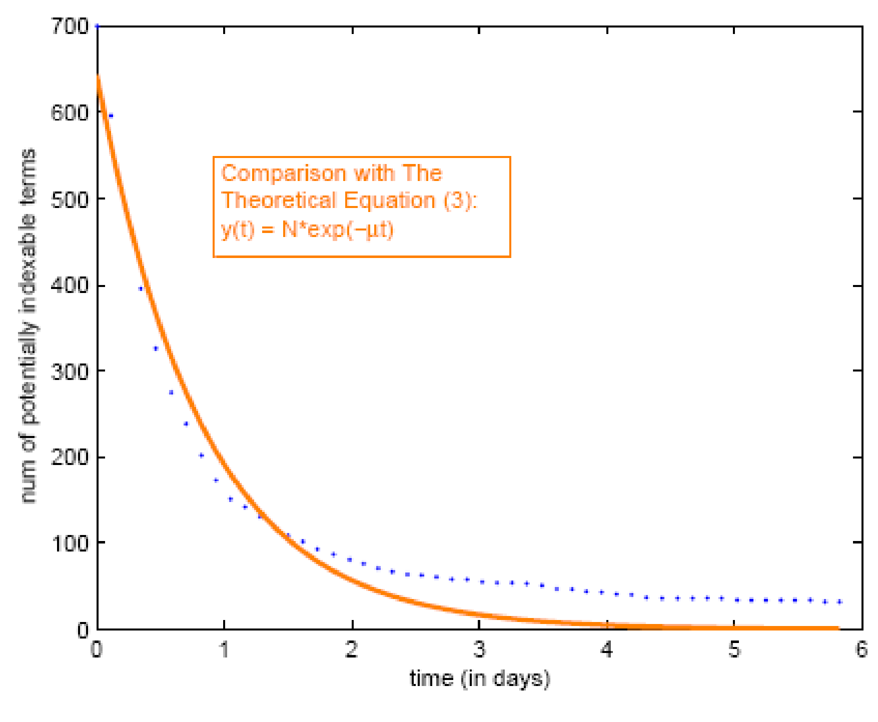
Collaborative Evolutionary Indexing

- Ranking Approach
  - Naïve Strategy
    - Returns the best ***k*** result objects ordered by index scores in decreasing order
    - Would lead to **local maxima problem**
  - Randomized Strategy with Genetic Algorithms (GA)
    - Returns ***k*** result objects by random extractions
    - Discover ‘hidden’ objects by randomness of GA
    - Performance quality by ***Elitism***

Experiments and Evaluations

To examine decay behaviour of number of potentially indexable terms *Nt*

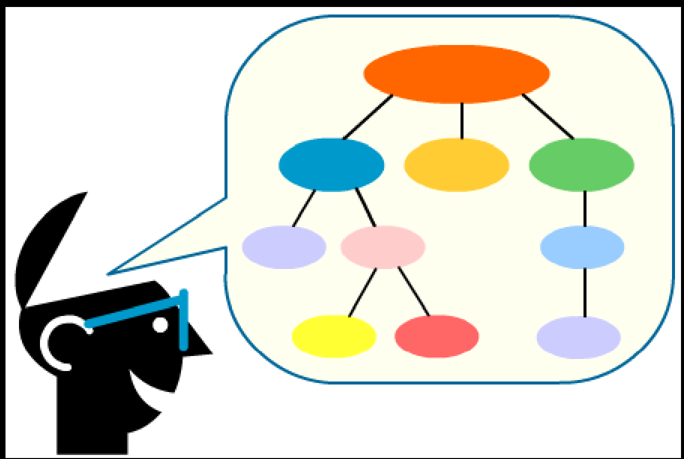
- Perform 50,000 queries with variables:
  - 100 data objects
  - 3 initial indexed terms
  - 2 search terms per query
  - 10 results per answer
- Conclusion:
  - ***Nt* 'decays' exponentially** over time
  - As t → infinity, the collection tends to be fully indexed
  - Our approach is robust with regard to no. of data objects

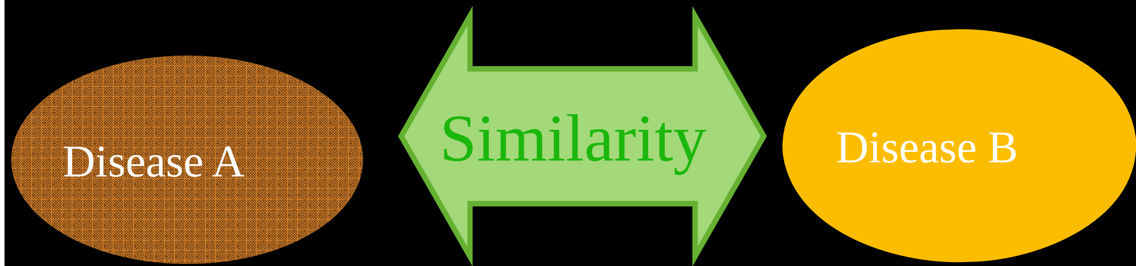


(a) t_{max}=10 and N_0=700

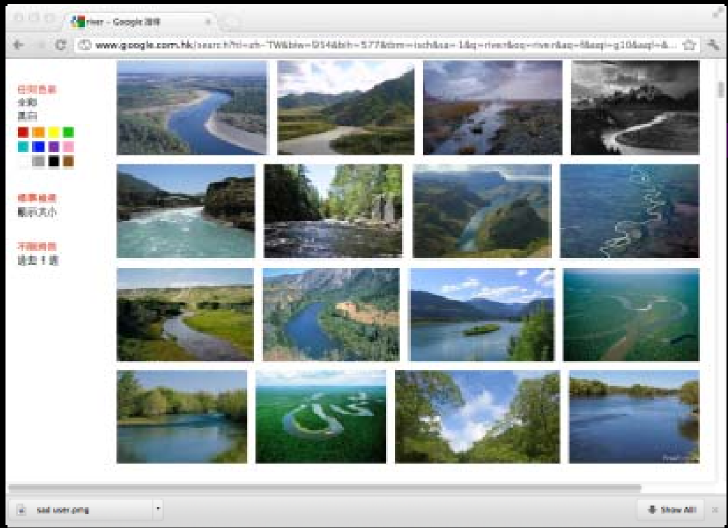
Augmenting Semantics Using Distances

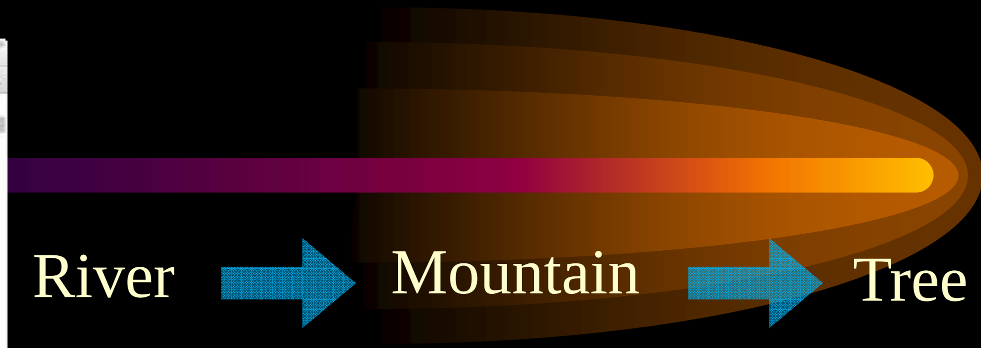
Use similarity through ontologies to define a distance between concepts (e.g. Computer-Aided Diagnosis)



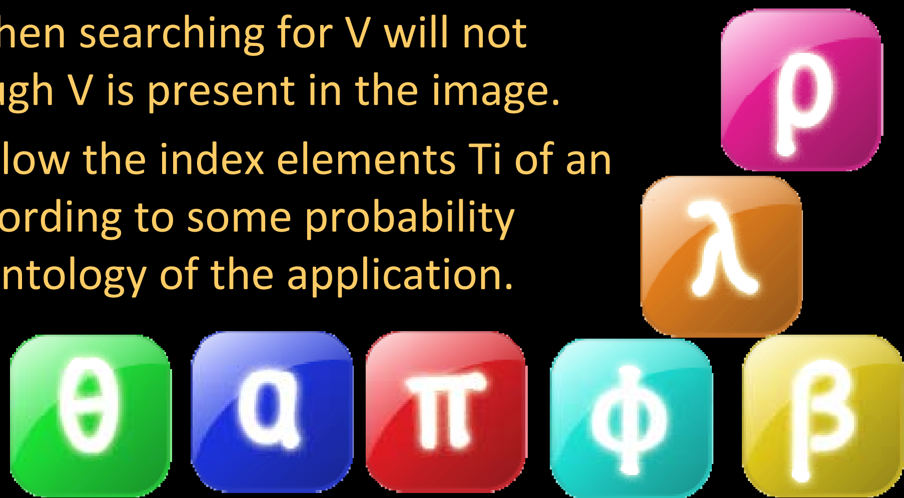


Index Expansion





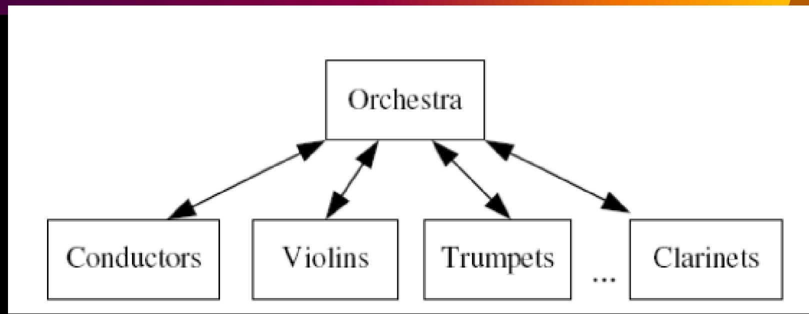
- The presence of particular objects in an image often implies the presence of other objects.
- If term U → V, and if only U is indexed, then searching for V will not return the image in the result, even though V is present in the image.
- The application of such inferences will allow the index elements Ti of an image to be automatically expanded according to some probability which will be related to the underlying ontology of the application.

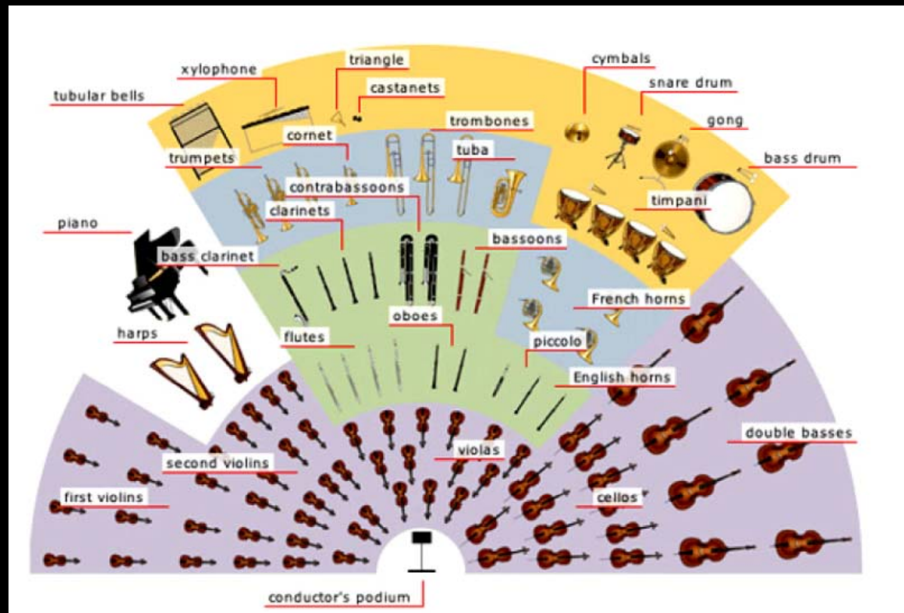


Ontology-based expansion

Aggregation hierarchical expansion

- This relates to the aggregation hierarchy of sub-objects that constitute an object.
- In this example, an orchestra expands to conductors, violins, trumpets, clarinets etc

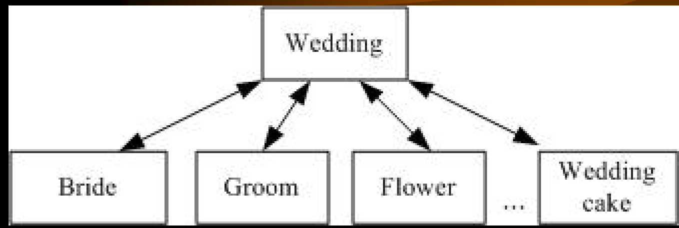




Ontology-based expansion

Co-occurrence expansion

- The relevant weighting may be expressed as a conditional probability given the presence of other objects.
- In this example, it is expected that certain semantic objects (e.g. bride, groom, flower) tend to occur together.

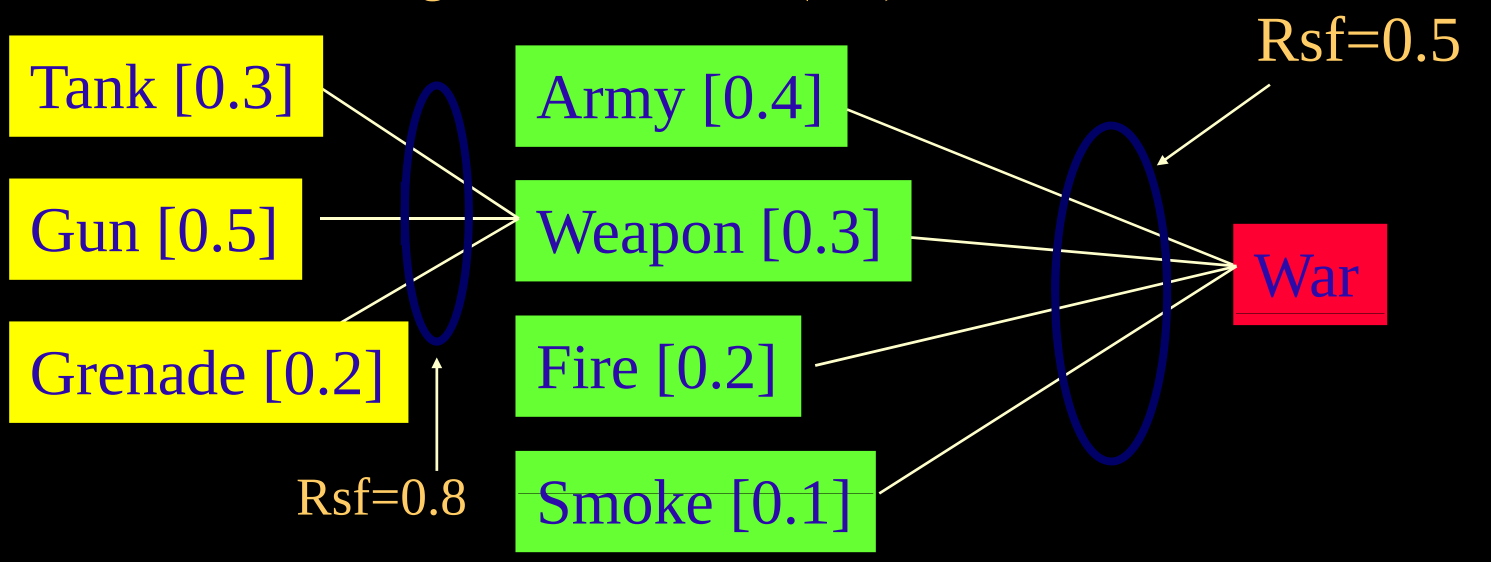


Index Weights

Index Weight = $R \times \Sigma$ (Constituent Weight)

Weight = Significance/prominence value of object within a concept

R = Rule significance factor (Rsf)



Index Weight Propagation

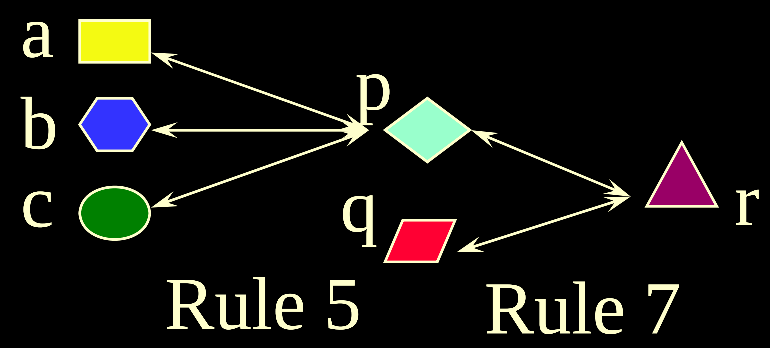
One-level: $W'_k = R_k \times \Sigma_i W_i$

Two-level: $W''_m = R_m \times \Sigma_k [R_k \times \Sigma_i W_i]$

Multi-level: $W_n^{(p)} = R_n \times \Sigma_l \{\ldots\Sigma_k [R_k \times \Sigma_i W_i]\}$

for given path of length p

Bi-directional Expansion of Rules



Forward (Rule 5): η= 4/3 = 1.33
Backward (Rule 5): η= 4/1 = 4
Forward (Rules 5 & 7): η= 6/4= 1.5
Backward (Rule 5 & 7): η= 6/1=6

Index Expansion Efficiency $\eta = \omega / \xi$

where ω is the number of output index items,
ξ is the number of explicitly input index items.

Wikipedia Link Vector Model

- Wikipedia is the world largest collaboratively edited source of encyclopedic knowledge.
- The Wikipedia Link Vector Model (WLVM) uses Wikipedia to provide structured world knowledge about the terms of interest
- Uses the hyperlink structure of Wikipedia rather than its category hierarchy or textual content.

Wikipedia Distance

- The Wikipedia Link Vector Model (WLVM) uses Wikipedia to provide structured associative and contextual knowledge about the terms of interest by using the hyperlink structure of Wikipedia
- WLVM makes use of the total number of links to the target article over the total number of articl
- If t is the total number of articles within Wikipedia, then the weighted value w for the link $x \rightarrow y$ is

$$\rho_{Wk}(x, y) = |x \rightarrow y| \times log(\sum_{z=1}^{t} \frac{t}{|z \rightarrow y|})$$

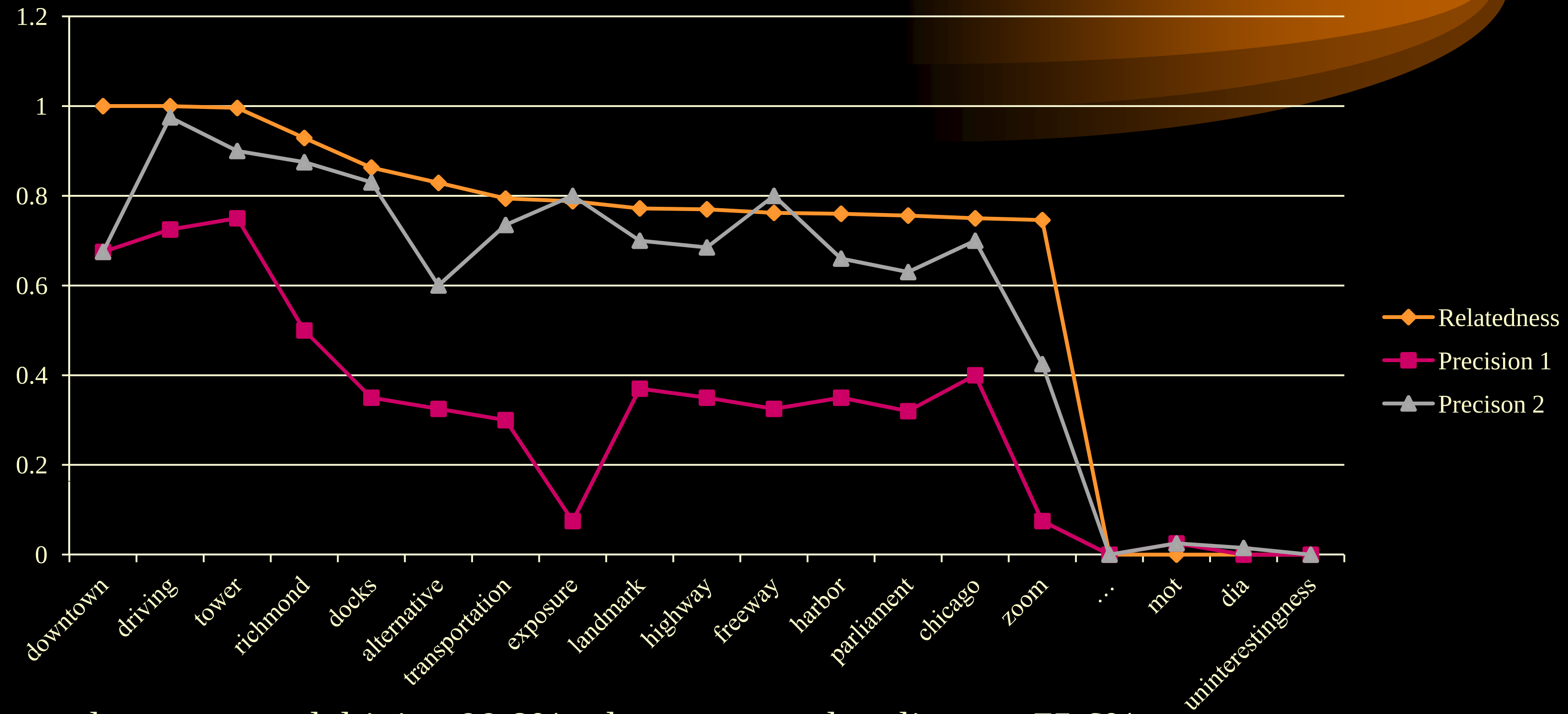
where $x$and y denote the search terms.

Wikipedia Similarity

Example: similarity between Israel and Jerusalem, the nation and its capital city.

- the number of times the term Israel is used to link to it: e.g. 95% of links are to the nation, 2% to the football team, 1% to the ancient kingdom, and a mere 0.1% to the Ohio township.

Relatedness between concept “downtown” and other concepts computed from the WLVM distance



Relatedness: downtown and driving 99.8% ; downtown and parliament 75.6%

Searching images using “downtown”: precision 67.5%,

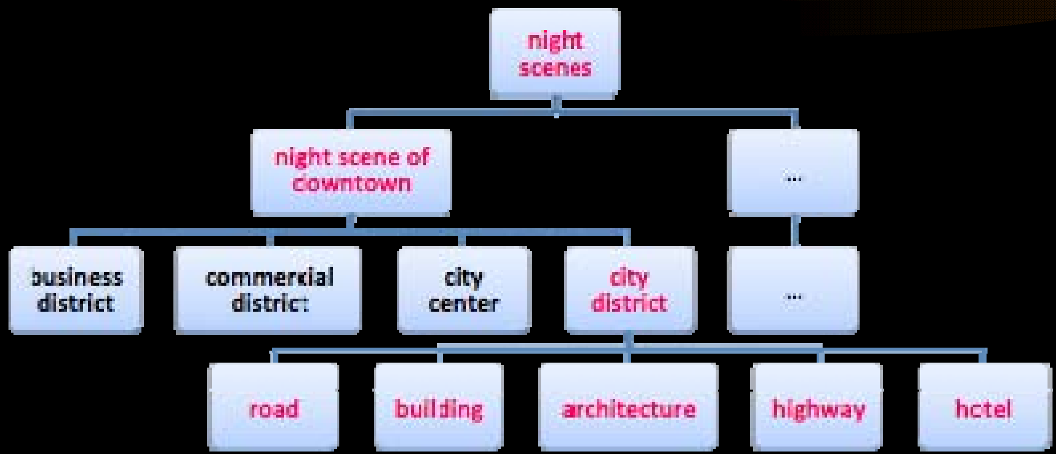
using “driving” instead of “downtown” yields downtown images with greater precision 72.5%.

using “downtown + driving”, dramatically raises the precision to 97.5%.

Wordnet, NGD, CYC, Flickr Distances

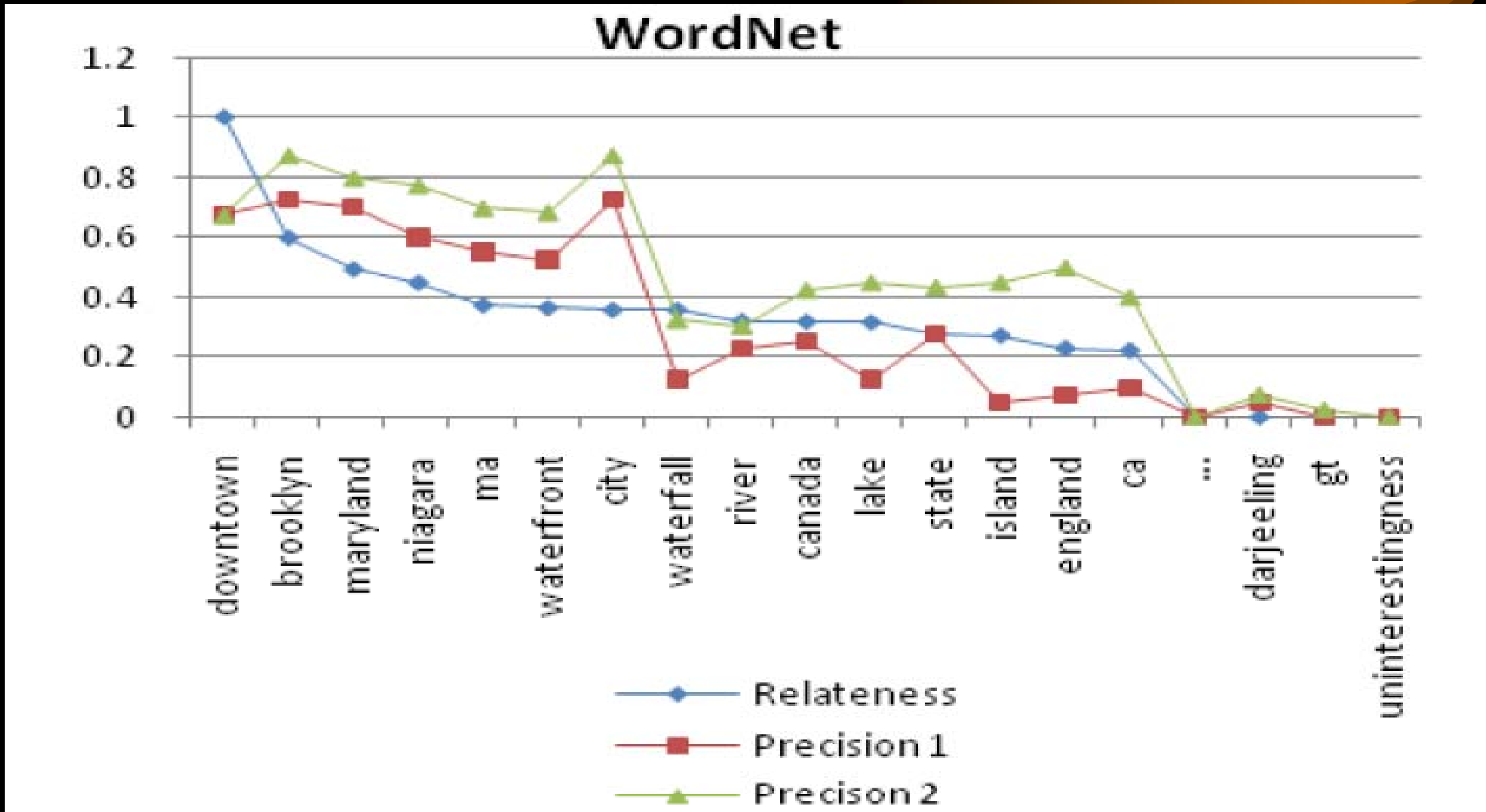
- $\rho_{Wn}(c_1, c_2) = \frac{2 \times log(lso(c_1, c_2))}{log\, p(c_1) + log p(c_2)}$
- $\rho_g(x, y) = \frac{\max\{log\, f(x), log(y)\} - log\, f(x,y)}{log\, N - \min\{log\, f(x), log\, f(y)\}}$
- $\rho_c(x, y) = \frac{2 * N_3}{N(x) + N(y) + 2 * N_3}$
- $\mathrm{FD}(c_1, c_2) = \sqrt{\frac{\sum_{i=1}^{K} \sum_{j=1}^{k} D_{JS}(P_{z_i} c_1 | P_{z_j} c_2)}{K^2}}$

WordNet Distance



- "downtown" can be expanded to "business district", "commercial district", "city center" and "city district", while "city district" can be expanded to "road", "building", "architecture", "highway" and "hotel".

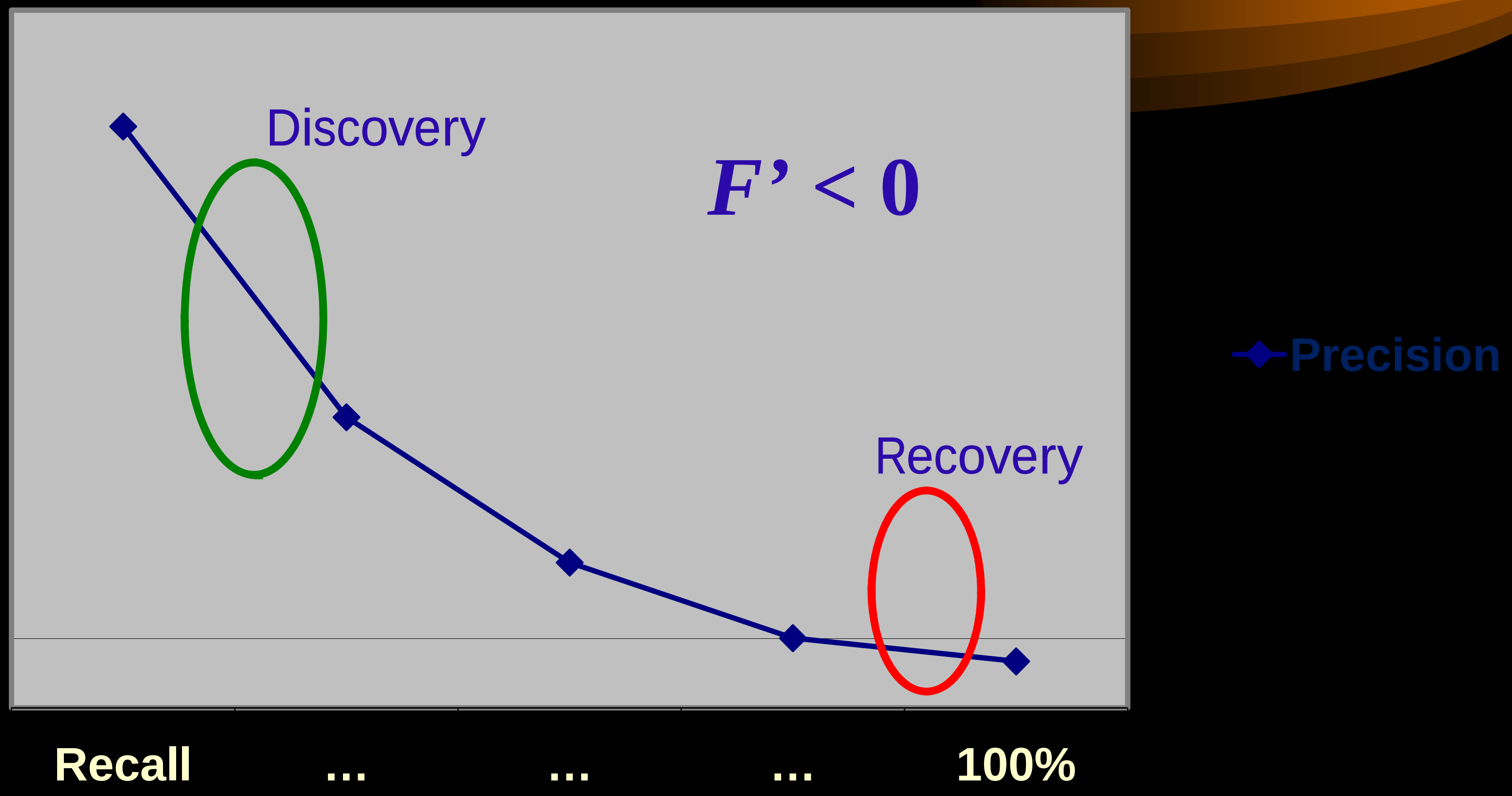
WordNet Distance



Relatedness between ‘downtown’ and other concepts

Top	WLVM	NGD	WD
1	downtown (100.0%)	downtown (100.0%)	downtown (100.0%)
2	driving (100.0%)	highway (98.2%)	brooklyn (59.7%)
3	tower (99.6%)	river (95.9%)	maryland (49.2%)
4	richmond (92.9%)	street (95.7%)	niagara (44.7%)
5	docks (86.3%)	park (95.2%)	ma (37.2%)
6	alternative (82.9%)	museum (93.1%)	waterfront (36.3%)
7	transportation (79.4%)	creek (93.0%)	city (35.6%)
8	exposure (78.8%)	transportation (92.8%)	waterfall (35.6%)
9	landmark (77.2%)	rain (92.8%)	river (31.9%)
10	highway (77%)	construction (92.2%)	canada (31.7%)
11	freeway (76.2%)	harbor (91.5%)	lake (31.5%)
12	harbor (76.0%)	one (91.3%)	state (27.5%)
13	parliament (75.6%)	bridge (91.1%)	island (27.1%)
14	chicago (75.0%)	road (91.1%)	england (22.8%)
15	zoom (74.6%)	arc (90.9%)	ca (22.1%)
...	...	...	...
415	mot (0%)	aigina (28.8%)	darjeeling (0%)
416	dia (0%)	sydney (25.9%)	gt (0%)
417	uninterestingness (0%)	uninterestingness (0%)	uninterestingness (0%)
Stdev	0.1379	0.1490	0.0884

Recall or Precision?



Expansion Criteria

- For high precision – expansion using intersection of the high ranking sets
- For high recall – may expand using unions of the relevant sets, or exclude certain distances

Challenge

How to develop an algorithm to automatically correlate low-level features to high-level concepts and run it fast enough to combat the big velocity problem?

Thanks