Visual Analysis of Unstructured Data Sets

Michael Granitzer

Know-Center GmbH & Graz University of Technology

http://www.**know-center**.at

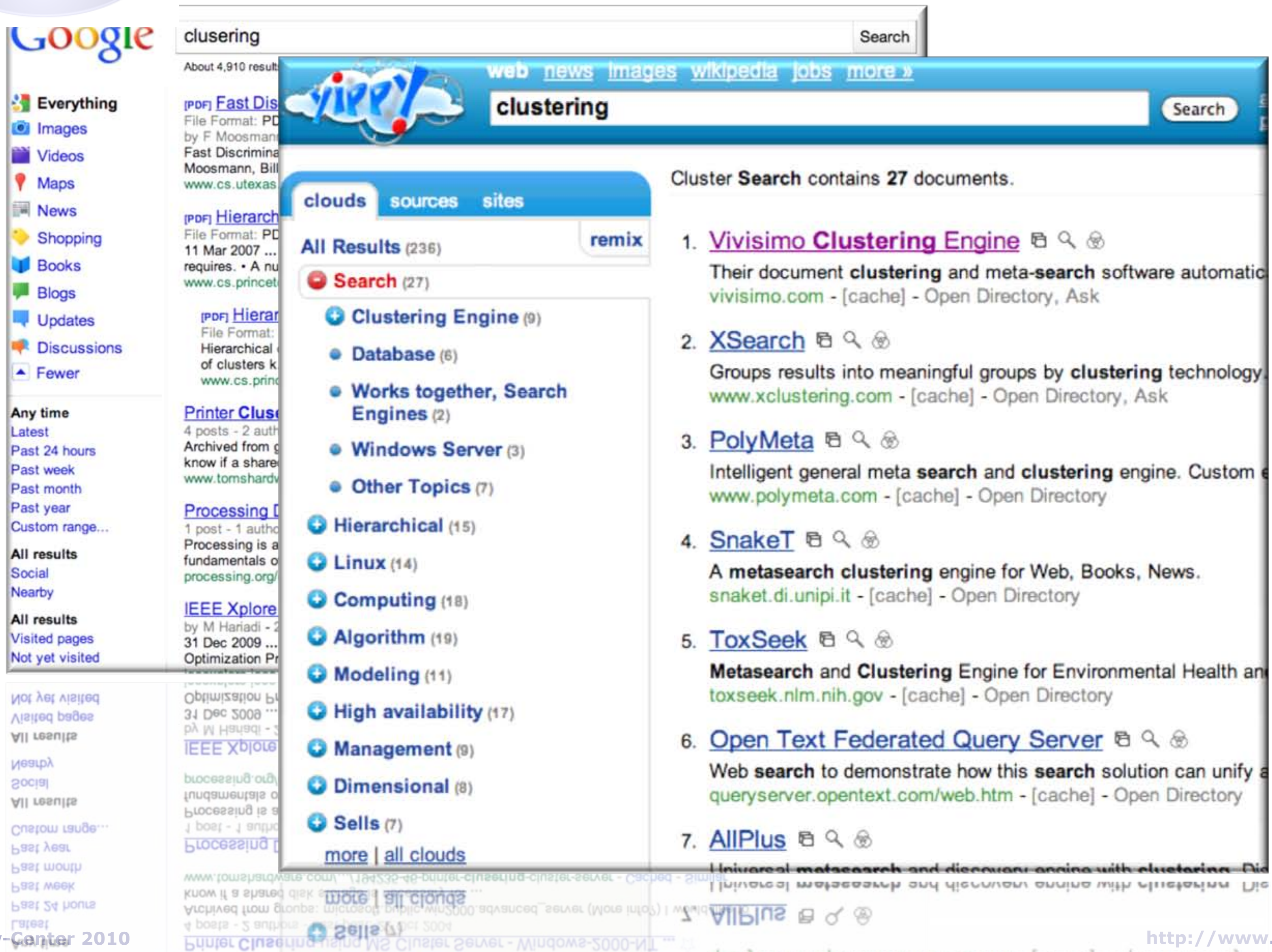
Outline

- Motivation
  - Facetted Retrieval + Scatter/Gather + Some Visual stuff
  - Why visual stuff?
- Clustering Approach (TIR 10)
  - Scalable Top-Down recursive Clustering approach with Model Selection
  - Experiments
- Labelling (SIGIR 2010)
  - Effects of structural relationships: Parent Child and Sibling Relationships
  - Experiments
- Feedback mechanisms (for discussion)
- Experiments
  - Visual Analysis
  - Inex

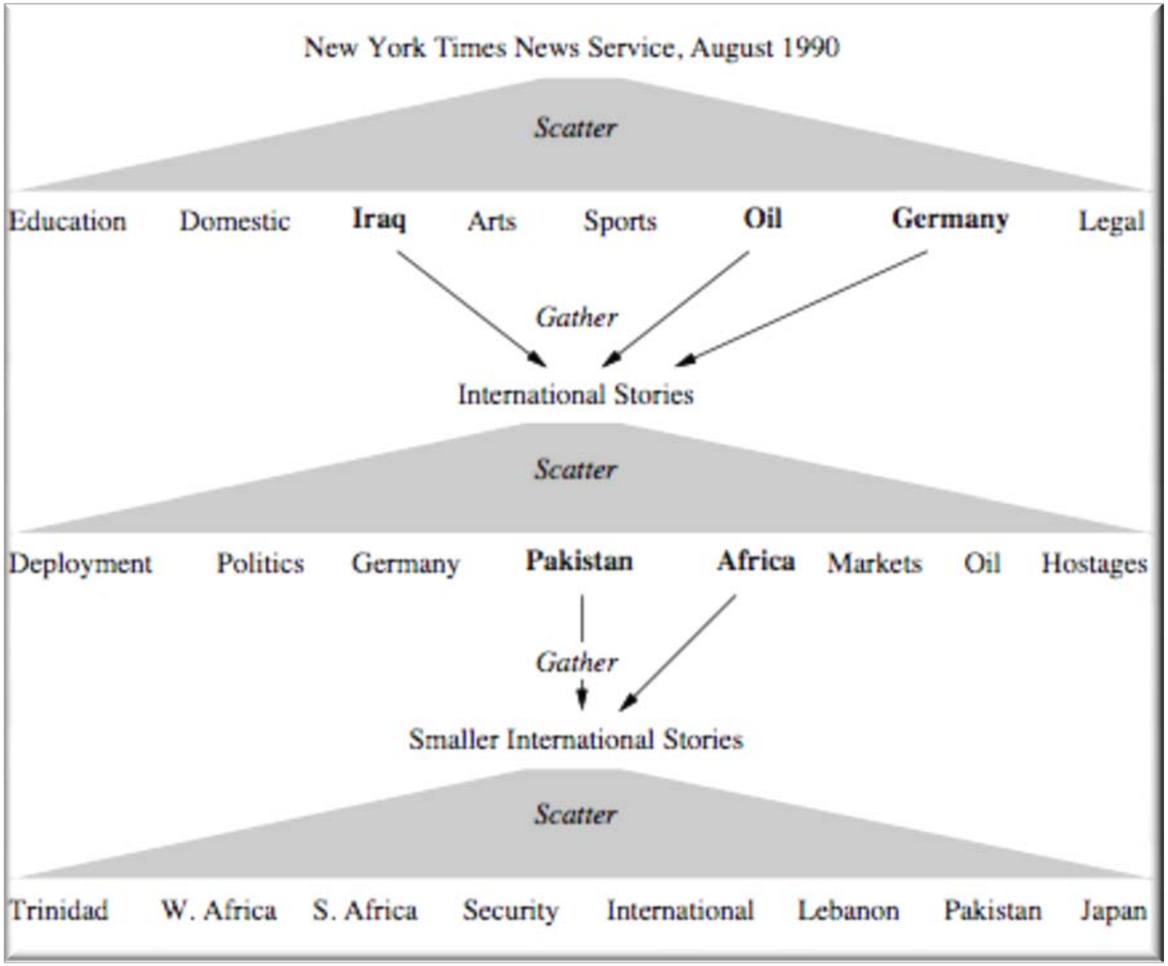
Know-Center ?!?

- The Know-Center is Austria's Competence Center for Knowledge-Based Applications and Systems, funded in the COMET program
- Application oriented research
  Bridge the gap between science and industry
- 21 Industry partners, 5 scientific partners (e.g. APA, Bertelsmann, Infonova ...)
  - Area 1: Knowledge Services – Technology enhanced learning, Context Detection
  - Area 2: Knowledge Relationship Discovery – Text Analysis, Visualisation, Retrieval, Plagiarisma Analysis, Social Media (PAN, CLEF, TREC etc.
  - Roman Kern, Elisabeth Lex, (Wolfgang Kienreich, Markus Muhr, VedranSabol, Christin Seifert, Christopher Horn, Mari Zechner, Werner Klieber)
- Applying Basic Research results in different application scenarios
  - Plagiarism Analysis == Media Diffusion Analysis (E.g. „Nike, just Sports")
  - Enterprise Search: not solved
  - Patent Analysis

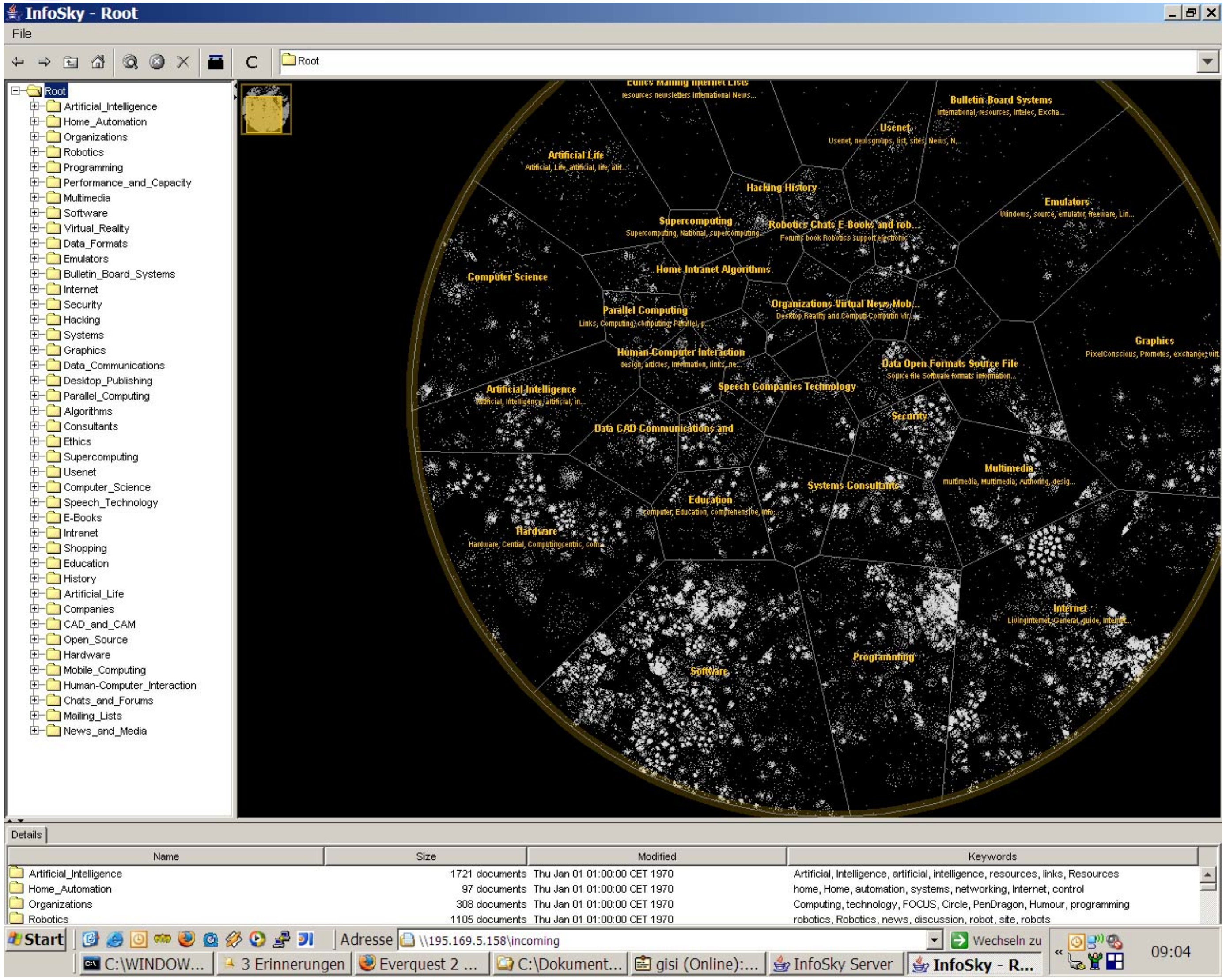
Motivation
Facetted Retrieval



Motivation
Scatter/Gather [Cutting et. al. 1992]



Motivation
InfoSky: Visual Exploration [Andrews et. al. 2002]



Motivation
Visualization, why?

- Exploit the capacity of the visual cortex to immediately recognies certain circumstances
- Example: PreattentiveProcessing

  A resricted set of visual properties can be recognized immediately

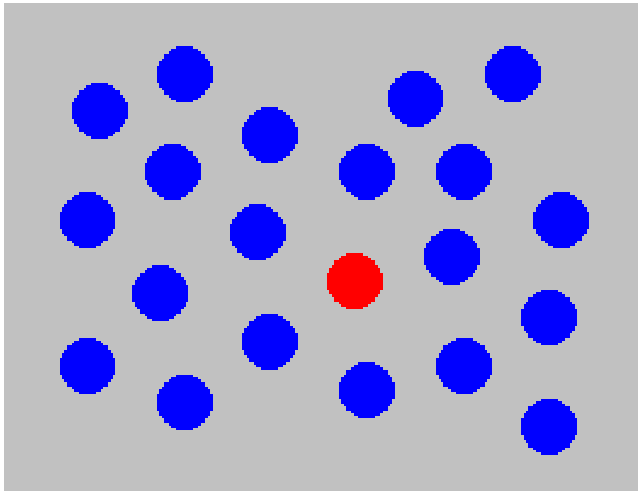
  Criteria 1: Processing time below <200 - 250ms (within the blink of an eye = 200ms)

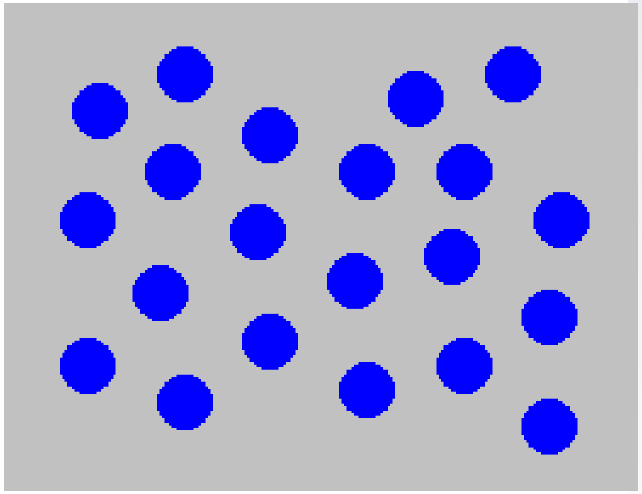
  Criteria 2: fixed time period independent of the number of noise

Where is the red circle?

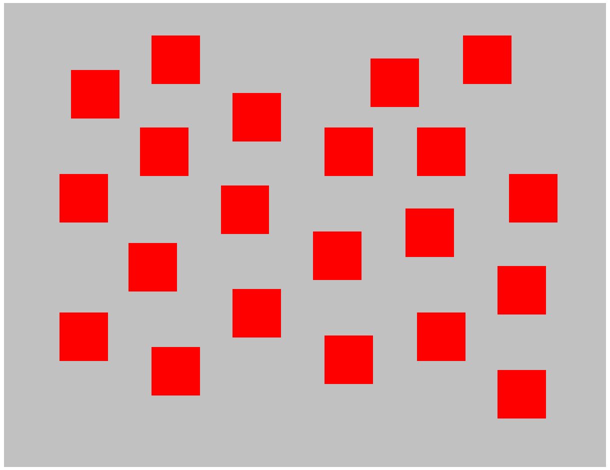
Motivation

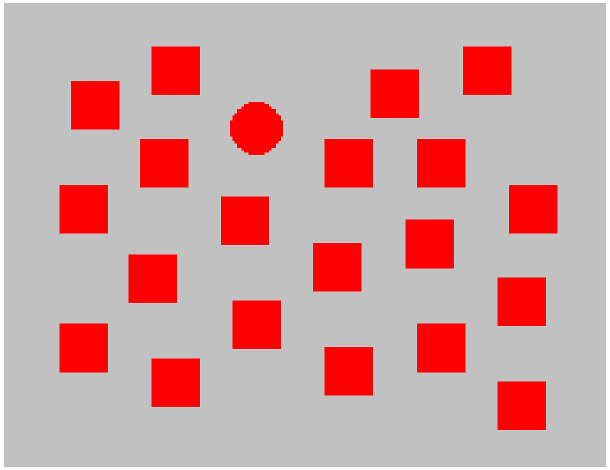
Visualization, why?



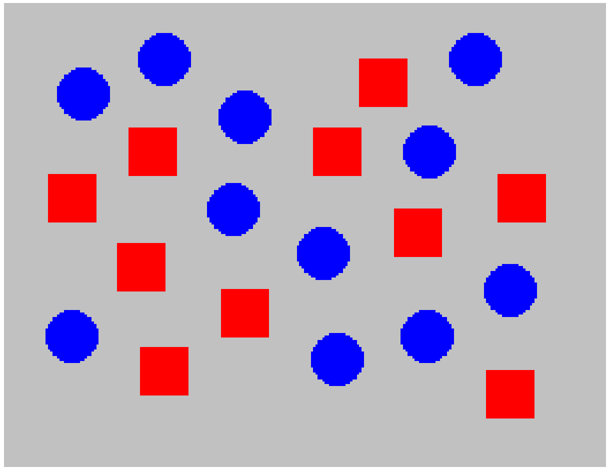


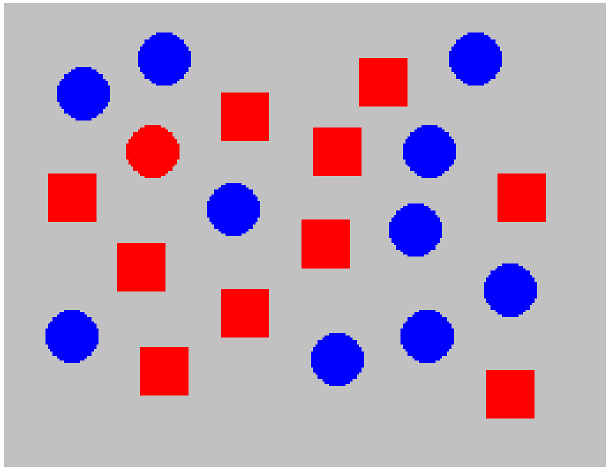
Motivation
Visualization, why?





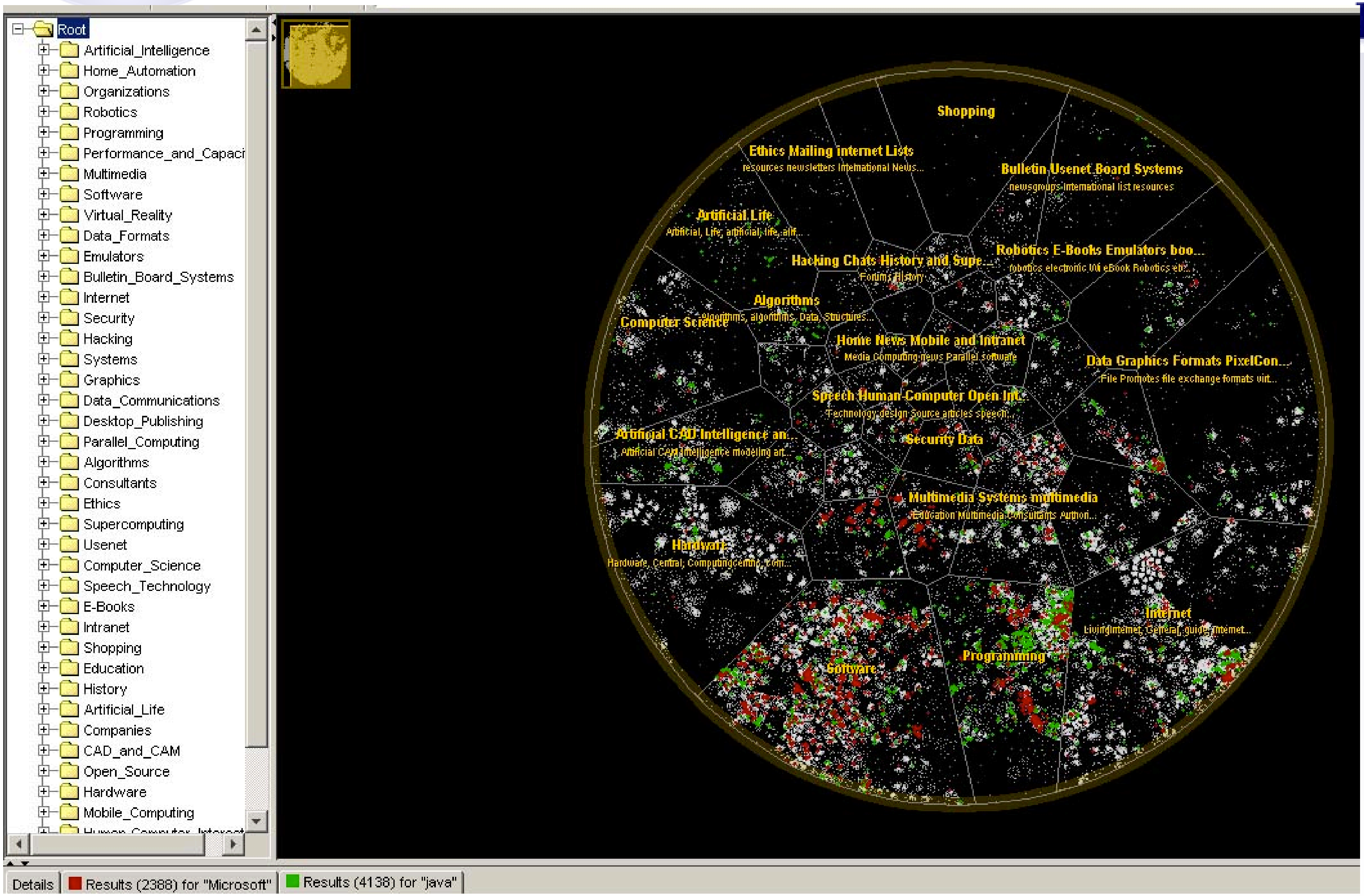
Motivation
Visualization, why?



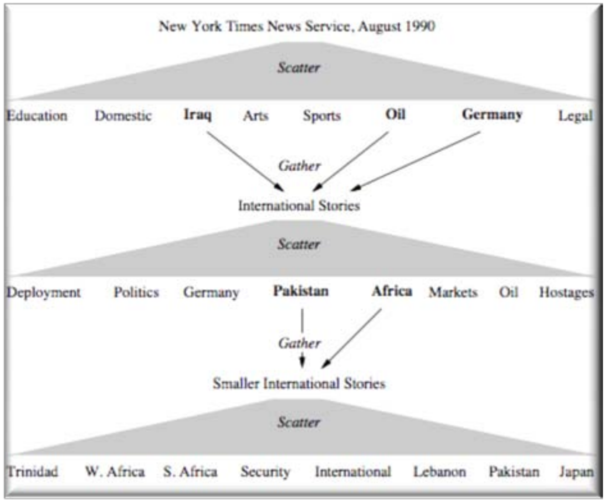


Text isabstract and hardlypre-attentive in contrast to images

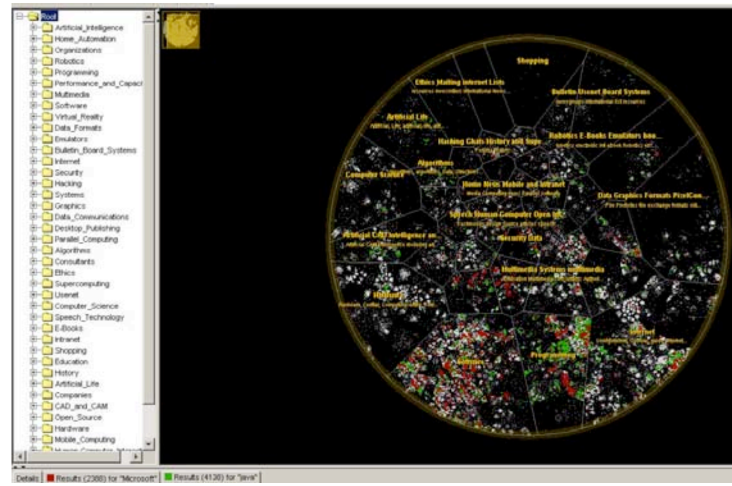
Motivation
InfoSky: Visual Exploration [Andrews et. al. 2002]



Motivation
InfoSky + Scatter/Gather



+



- Automatic creation of the cluster hierarchy while retaining InfoSky‘s analysis capabilities

Questions

- What is an efficient hierarchical clustering algorithm therefore?
- How to combine statistical data set properties with visual requirments?

Approach & Contribution

- Preprocessing
  - Nothingnewhere.....
- Clustering
  - Combinewell-knowntechniques (Growingk-means, Model Selection....)
- Projection
  - Clustering + Force DirectedPlacement: $O(n^3) \rightarrow O(n*\log(n))$
- Labelling
  - Label qualitydepends on thehierarchystructure
  - Ad-hocsolution, yet no well foundedtheoreticalapproach
  - Clustering + Force DirectedPlacement: $O(n^3) \rightarrow O(n*\log(n))$
- Metric Feedback: Just fordiscussions...

RecursiveTop-DownHierarchicalClustering

- Hierarchical, top-down, polythetic, documentclusteringapproach
- Dynamicclusterstructure on eachlevel of thehierarchysupportingsplitting and merging of clusters.
- Constraints on themaximum and minimumnumber of elements per hierarchylevel
- Resultingreducedcomputationalcosts of the layout algorithm
- Scalable to datasetsconsisting of millions of documentswith a reasonabletrade-offbetweenruntime and accuracy

Top-Down, scalableclusteringalgorithmforcreating a topicalhierarchy

Clustering Overview

Divide and conquer: decompose into tasks starting at the root node

For every task

- Step 1: Preprocess documents to be clustered
  - Bag-of-Words, BM 25, cosine inner product
- Step 2: Cluster documents using a flat clustering algorithm
- Step 3: Split and merge clusters till constraints are met
- Step 4: Recursion: Evaluate the stopping criterion for dividing into further sub-tasks
- Step 5: Cluster Labeling
- Step 6: Project clusters into a 2 dimensional space

Clustering
Step 2: Clustering Algorithm (1/4)

Given a set of documents X, find a set of K groups of similar documents (clusters)

- Utilize existing clustering methods
  - HAC, DBScan or Chameleon > O(n^2)
  - GNG, BIRCH fast and storage efficient, but order dependent

- Growing k-means
  - Online Competitive Learning with Winner-takes it all approach
  - trade-off between runtime and accuracy [Zhao and Karypis 02]
  - Allows for efficient model selection (determine k)

Clustering
Step 2: Clustering Algorithm (2/4)

Algorithm 1 Growing Spherical K-Means

input:
$\mathcal{X} = \{x_1, \ldots, x_N\}$ with $x_i \in \Re^d$, K, l, η, ν
output:
$\mathcal{C} = \{c_1, \ldots, c_K\}$, $\mathcal{Y} = \{y_1, \ldots, y_N\}$ $\forall$ $y_n \in \{1, \ldots, K\}$
steps:

initialize centroids c_1 and c_2 by a seeding mechanism
for $m = 2$ to K **do**

Init and loopformaximumk-clusters

for $n = 1$ to N **do**
$y_p = y_n$
$y_n = \arg\max_{1 \le k \le m} x_n^T c_k$
$c_{y_n} = c_{y_n} + \eta x_n$
$c_{y_p} = c_{y_p} - \nu x_n$

Update clusterhypothesis

if $||c_{y_n}|| - 1.0 > l$ **then**
$c_{y_n} = \frac{c_{y_n}}{||c_{y_n}||}$

Runtimeimprovement of centroid update

for $n = 1$ to N **do**
$y_n = \arg\max_{1 \le k \le m} x_n^T c_k$
$s_k = s_k + \max_{1 \le k \le m} x_n^T c_k$

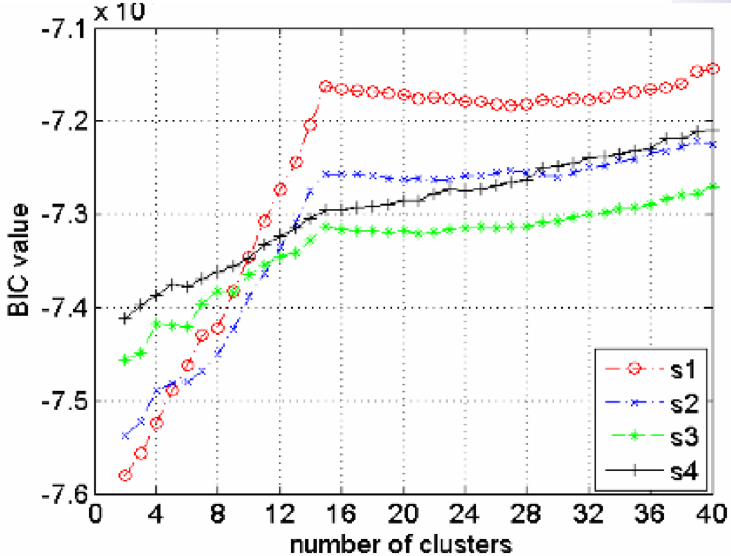
Assigndocuments and averagesimilarity

if $m < K$ **then**
$c_i = \arg\min_{1 \le k \le m} S(c_k)$
$x_j = \arg\min_{x \in \mathcal{X}_i} x^T c_i$ with $\mathcal{X}_i = \{x_n | y_n = i\}$
$c_t = \frac{c_i - x_j}{2}$, $\mathcal{C} = \mathcal{C} \cup \{c_t\}$

Createm-thcentroid

Clustering
Step 2: Clustering Algorithm (3/4)

- Model Selection methods
  - Obtain fitness criterion for different number of clusters (Bayesian Information Criterion (BIC), Stability based approaches)
  - Monotonical increasing/decreasing
  - Overtraining on the data
  - Determine the „best cluster number“ using knee-point detection [Zhao et. al. 2008]



- Efficient calculation for the growing k-means by simply calculating the fitness criterion for each new centroid

Clustering
Step 2: Clustering Algorithm (4/4)

Heuristics

- Efficient update rules [Zhong 2005]

  Move a fraction of the distance between sample and centroid

  $$c_{y_n} = \frac{c_{y_n} + \eta(x_n - c_{y_n})}{||c_{y_n} + \eta(x_n - c_{y_n})||}$$

  Simply update the angle and ignore non unit length

  Track norm changes and rescale after norm exceeds numerical boundaries

  $$\begin{aligned} c_{y_n} &= c_{y_n} + \eta x_n \\ c_{y_p} &= c_{y_p} - \nu x_n \\ \textbf{if } ||c_{y_n}|| - 1.0 > l \textbf{ then} \\ c_{y_n} &= \frac{c_{y_n}}{||c_{y_n}||} \end{aligned}$$

- Decreasing learning rate with the size of the cluster for balancing

$$\eta = 1/|\sqrt{\mathcal{X}_{k(x)}}|$$

Clustering
Step 3: Split and Merge

Split and Merge Clusters to fulfill the following constraints

- # Cluster at one level
  - Merge the most similar cluster if #cluster > maximum number of clusters
  - Split the least coherent or biggest cluster if #cluster < minimum number of clusters
- # documents in a cluster
  - Below the Maximum number of documents for a cluster ➔ clusterokforbrowsing
  - More than 1.5 times the upper limit to ensure meaningfull clustering at next hierarchical level

If all clusters fullfill this constraint, cluster recursively (Step 4)

Clustering Experiments INEX Clustering

- Initiativ for Evaluation of XML Retrieval
- XML Mining Track – Cluster the English Wikipedia
  - Small data set 54k documents
  - Large data set 2.6 Million Documents
  - Preprocessed document vectors (uni and bi-grams)
- Ground truth provided by YAGO ontology, but no hierarchical structure
- Document assigned to each cluster on the path to facilitat multi cluster assignment as it is the case in Wikipedia

Clustering Experiments
INEX Clustering

- 10,467 Clusters for the small data set
  - 4 Minutes to compute on a 16GB Quad Core including I/O

MacroPurity	BIC	Stability
73k Categories	0.4959	0.4945
12k Categories	0.5473	0.5303

- 133,704 Clusters on the large data set
  - Runtime 2 hours
  - 348 k Categories: Macro Purity of 0.4457
  - 12k Categories: Macro Purity of 0.5359
- Clusters appear to be reasonable, but good evaluation strategy remains an open issue
  - High level clusters are more important
  - Accurate ground truth reflecting good browsing strategies

Clustering
Step 5: Labeling - Overview

- Labeling via Jensen Shannon Divergence
  - How to achieve good labelingqualityforbrowsing?
  - Doeslevel of thehierarchy has an impact on thelabelquality?
- Intuition
  - Take structuralrelationshipsintoaccount to improvelabelquality
  - Siblings - labelsshouldhelp to separate neighborclusters
  - Hierarchies - labelsshouldbecomemoregenericthehigherthecluseriswithinthehierarchy
- Open Issueshere
  - Most state-of-the-artlabelingapproaches do notexploitstructuralrelationships
  - No standardized test dataset
  - No evaluationforbrowsingpurpose

Clustering Labeling - Approach

- Extendexistingwell-knownlabelingtechniquesbystructuralrelationships
  - Maximum termweightbasedmethods
  - Referencecollectionbasedmethods

- Types of structuralrelationships
  - Siblingrelationships
  - Parent-childrelationship

Assumption: All labelingalgorithmsarebased on a bag of wordmodel. Extension possiblewithbi-grams, tri-grams etc.

Clustering
Labeling – Maximum Term Weight Labeling

- Pick the top k terms according to a weighting scheme by summing over all cluster documents
  - Local weights
  - Global weights (IDF, BM25)
  - Named in the evaluation: $MTWL_{raw}$

$$L_j \leftarrow best_k\Big(\sum_{d_i \in \mathbf{D}_{c_j \to *}} idf_{global} \cdot tfWeight(d_i)\Big)$$

Clustering
Labeling – Reference Collection based

- Compare the distribution of terms within a cluster with a reference collection
  - χ^2 Popescul and Ungar [2000]
  - Information Gain Geraci et al. [2007]
  - Jensen-Shannon Divergence Carmel et al. [2006]
  - Named in the evaluation: *JSD*

$$L_j \leftarrow best_k\Big(JSD(\mathbf{D}_{ref}, \mathbf{D}_{c_j \to *})\Big)$$

Clustering
Labeling – Inverse Cluster Weight Labelling (ICWL)

How to exploit the sibling relationship?

- Follow the approach of the CFC classification algorithm Guan et al. [2009]
- *Intuition*: If one term occurs often in one sibling cluster only, this term should be preferred over terms occurring in all sibling clusters
- Integrate sibling weighting into the maximum term weight labeling
- Named in the evaluation: $ICWL_{raw}$

$$icf_{j,k} = \exp\left(\frac{\#(t_k, \mathbf{D}_{c_j \to *})}{|\mathbf{D}_{c_j \to *}|}\right) \log\left(\frac{\#(c_p)}{\#(t_k, c_p)} + 1\right)$$

Clustering
Labeling – Hierarchical Labelling

How to exploit the parent-child relationship?

- *Intuition*: Integrate the path length (distance between cluster to label and document a term occurs in) into the label calculation and promote terms occurring in a higher number of child clusters.
- Hierarchical labeling extends all introduced labeling approaches
- Added prefix *hier* in the evaluation

$$L_j \leftarrow best_k\Big(\sum_{c_i \in \mathbf{C}_{c_j \rightarrow *}} \frac{1}{l(j,i)} * cf_{l(j,i)} \cdot v_{j,i}\Big)$$

Clustering Labeling - Evaluation

- Open Directory Project (ODP)
  - Top categories: arts, business, games, health, home, news,society, sports
  - Ignored soft links, ignoredsingleletter categories
- Wikipedia
  - Top categories: arts, computing, health, sports
  - Restricted to 10 sub-categories and 80 articles (drawnrandomly)
  - Ignoredinternalcategories, ignored “authorsbyyear”categories, limitednumber of documents per category, ignoredcycles
- Oshumed
  - Meshtreehierarchy
  - Documentsonly at leafcategories
- European Patents
  - Years 1991-2000
  - IPC classificationhierarchy

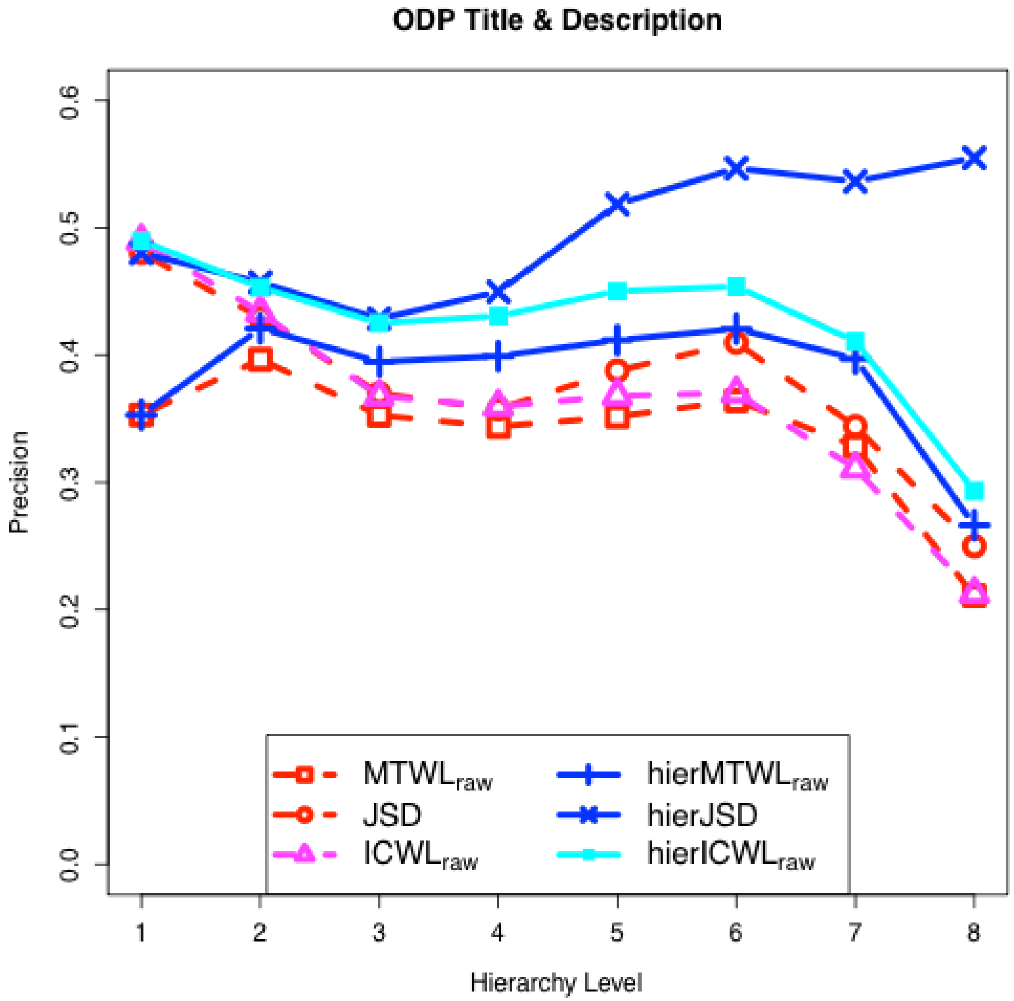
Clustering
Labeling - Evaluation

	Categories	Documents
ODP	150,000	800,000
Wikipedia	50,000	400,000
Oshumed	7,724	348,564
Patents	60,000	265,409

Preprocessing: Tokenized openNLP, Stemmed, Stop-word removal

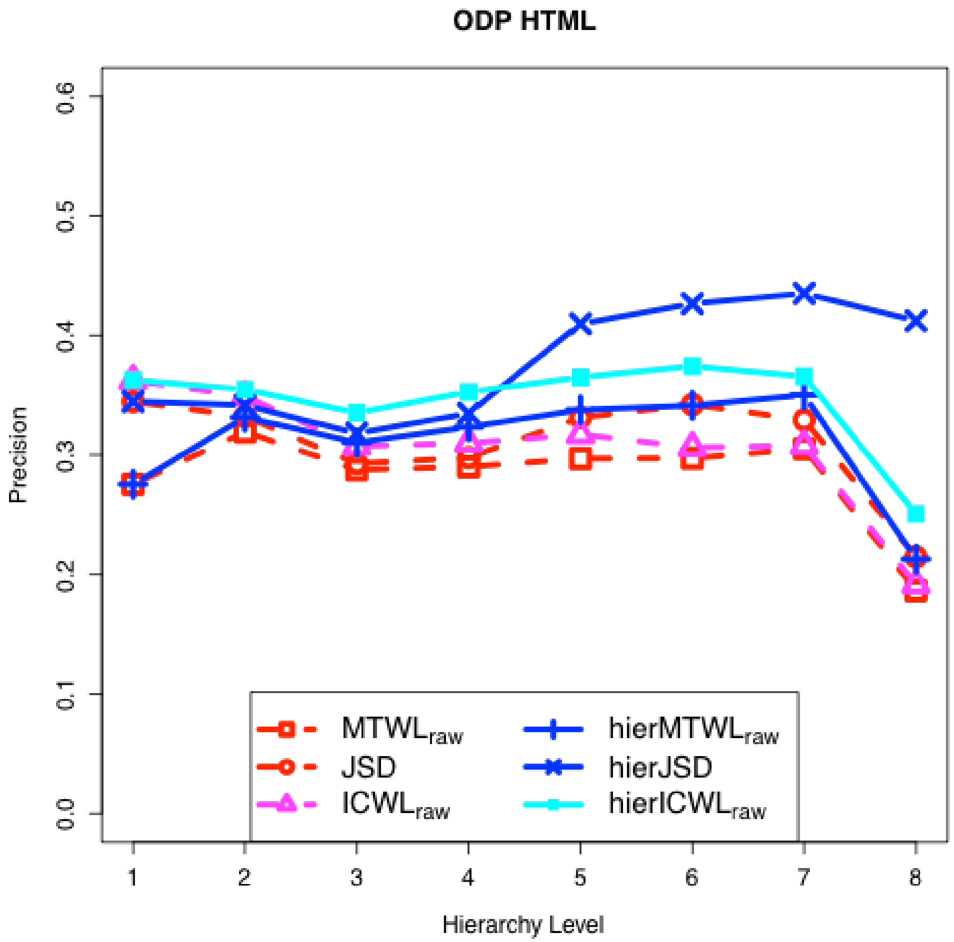
Clustering Labeling - Evaluation

- Precision over hierarchies with different depths



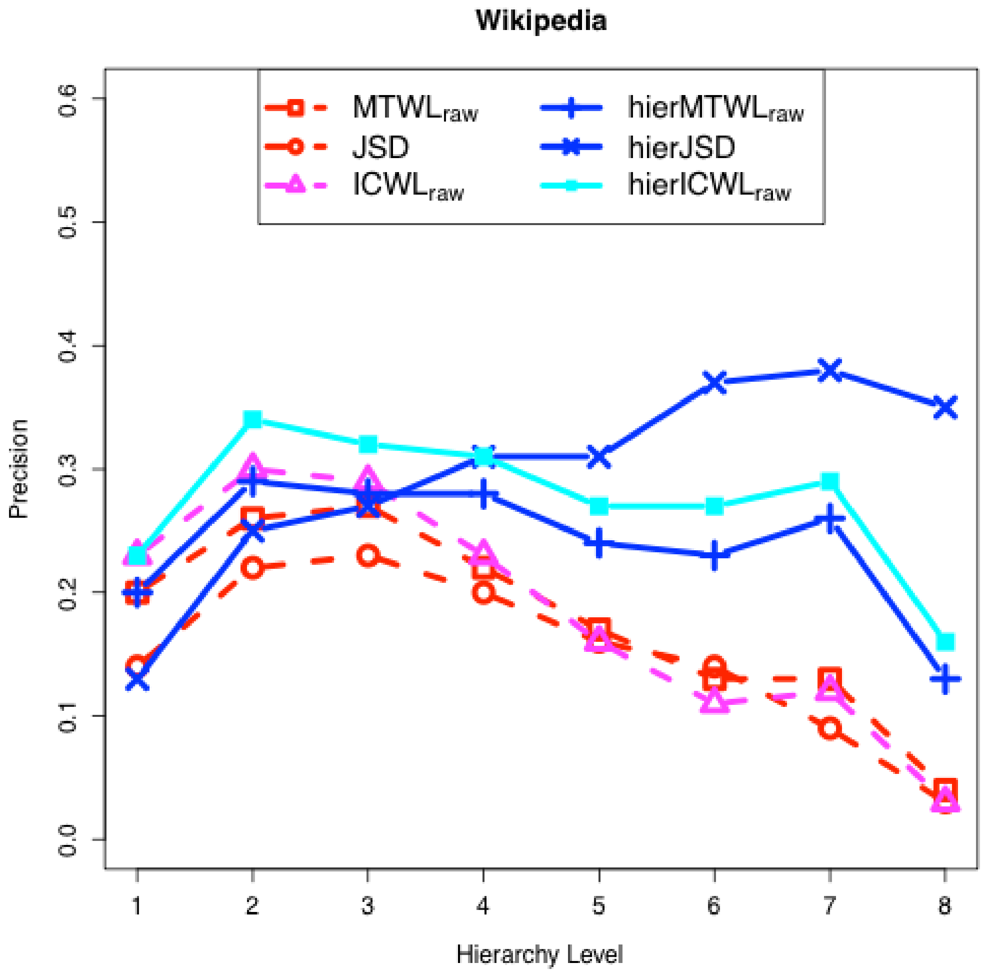
Clustering Labeling - Evaluation

- Precision over hierarchies with different depths



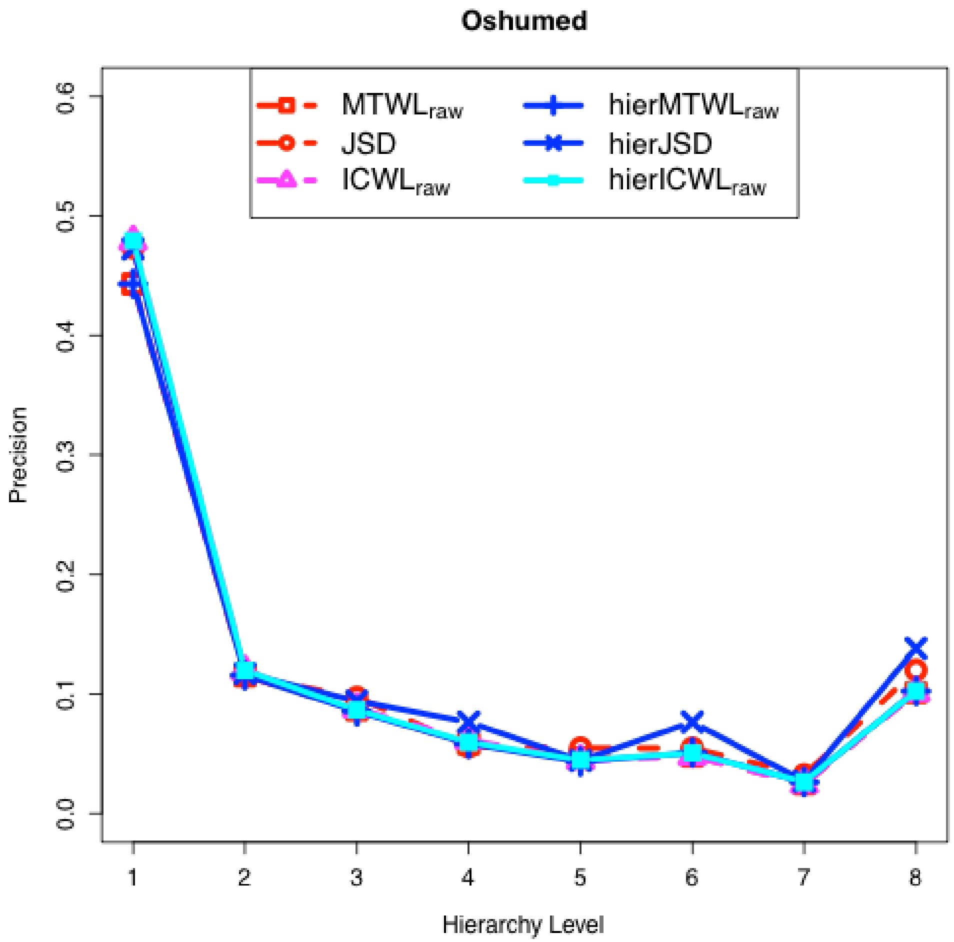
Clustering Labeling - Evaluation

- Precision over hierarchies with different depths



Clustering Labeling - Evaluation

- Precision over hierarchies with different depths



Clustering
Labeling – Evaluation - Summary

- Sibling Relations

  No impact on ODP and Ohsumed

  Slightimprovementsovertherespective MTWL methodsfortheWikipediadataset

- Summary Partent Child Relations

	$MTWL_{raw}$	JSD	$ICWL_{raw}$
ODP - T&D	0.06	0.15	0.08
ODP - HTML	0.04	0.09	0.05
Wikipedia	0.08	0.19	0.12
Oshumed	0.00	0.01	0.00

Table: Average relative difference of the precision for all hierarchy levels greater than 2 for all datasets between the different methods either with and without exploitation of hierarchical information.

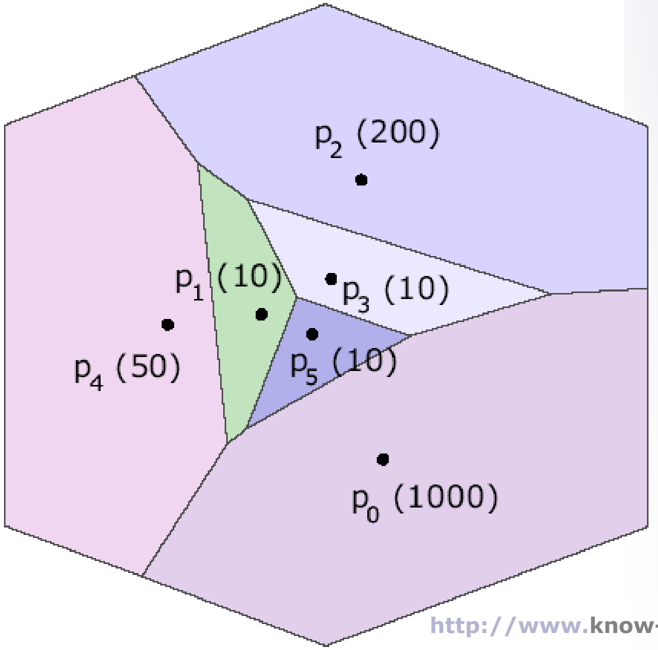
Clustering
Labeling – Evaluation - Conclusio

Interpretation of theResults

- Flatlabelingapproachessupportthebrowsing of leafnodesratherthanthebrowsing of high levelnodes → a resultquitecontradictory to theusersneed
- Usingsiblinginformationincreaseslabelingaccuracy in somedatasets
- Integratinghierarchicalinformationproducesbetterlabelingresultsfor all datasets
- Labelingaccuracyisstronglydomaindependent

Clustering
Step 6: Projection

- Projection [Andrews et. Al. 2004]
  - Force directed placement $O(n^3)$
  - Recursive application on cluster hierarchy using document and cluster centroids as points to layout
  - Due to the constraints we achieve a runtime of roughly O(n*log(n))
  - Voronoiinscription of rectangular Layout



Clustering
Step 7: Metric Feedback

Not implemented/analysed yet

- High dimensional distances➔ Low Dimensional Distances
- User movespoints on theplain
- New Low Dimensional Distances➔ Update high dimensional similartiy
- MetricLearning: Therearesomeapproaches

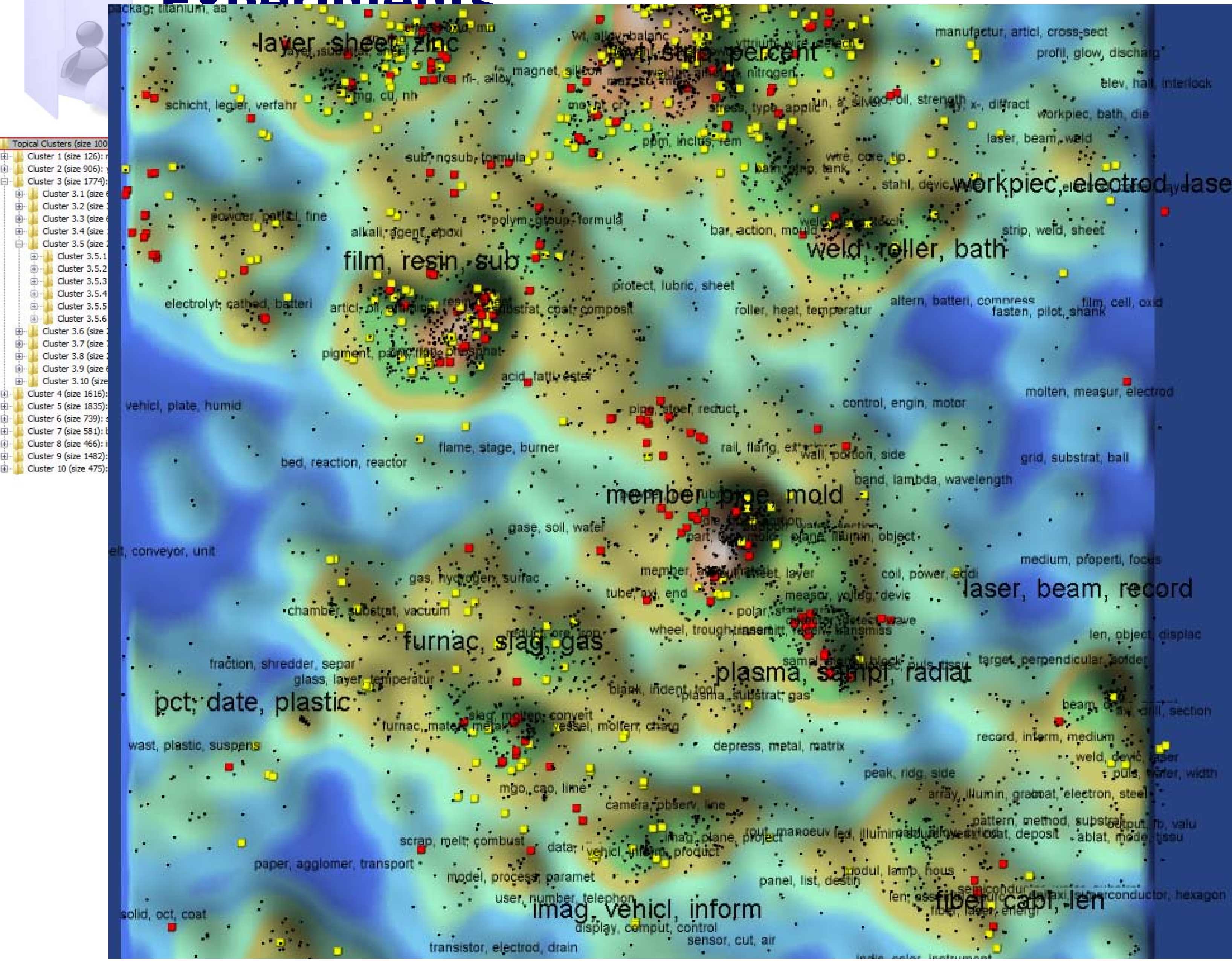
Donald Metzler and Hugo Zaragoza. Semi-parametric and non-parametrictermweightingforinformationretrieval. In Proceedings of the 2nd International Conference on theTheoryof Information Retrieval (ICTIR 2009), 2009.

Marco Ernandes, Giovanni Angelini, Marco Gori, Leonardo Rigutini, and Franco Scarselli. Adaptive context-basedterm (re)weightinganexperiment on single-wordquestionanswering. Frontiers in ArtificialIntelligence and Applications; Vol. 141, page 1, 2006.

Shai S. Shwartz, Yoram Singer, and Andrew Y. Ng. Online and batchlearning of pseudo-metrics. In ICML ’04: Proceedings of thetwenty-first international conference on Machinelearning, pages 94+, New York, NY, USA, 2004. ACM

Granitzer M., Adaptive Term WeightingthroughStochasticOptimization, CICLING 2010, Springer

Experiments



Experiments
Clustering based Visualisation

- Not for search, but for analysis of unstructered text documents
- Preliminary user evaluation
  - Combination of visualisation and standard components helpful for explorative tasks [Andrews et. Al. 2002]
  - Improved interaction and navigation paradigms to support explorative search tasks
  - Patent analysis tasks improved in real world use case
  - Suitable for high recall search tasks
- Detailed evaluation still missing
- Similarity biases results

?? Could the user be utilized to learn similarity metrics via such visualisations??

Summary & Conclusio

- Support explorative search and analysis tasks, not standard retrieval
- Top-down, recursive algorithm with different model selection strategy to scale
  - K-Means based approaches simply work well, invest in features in stead of algorithms
- Labeling exploiting hierarchical relationships improves labeling accuracy
  - External resources + hierarchical relationships + !bag-of-words= ??

Evaluation forBrowsingbehaviourhard to conduct: Missing measures&datasets; no comparison to literature

Thansk for your attention

Questions?

Michael Granitzer
Scientific Director
Know-Center Graz
Inffeldgasse 21a
8020 Graz

+43 316 873 9263
mgrani@know-center.at
www.know-center.at