媒体大数据搜索与挖掘

程 健 研究员

jcheng@nlpr.ia.ac.cn

中国科学院自动化研究所

模式识别国家重点实验室

2016.12.17

大 纲

- 背景介绍
- 索引与排序
- 近似近邻搜索
- 总结与展望

大 纲

- **背景介绍**
- 索引与排序
- 近似近邻搜索
- 总结与展望

背景介绍

- 网络资源呈爆炸式增长
  - 计算机技术与互联网技术
  - 各种数码设备：MP3、MP4、DC、DV等
- 资源类型
  - 文本：新闻，博客，电子书
  - 图像：网页图片，个人影集，网络图像集（Flickr）
  - 视频：播客，网络视频集（Youtube，优酷），广告
  - 音频：音乐库，语音库

背景介绍

Google

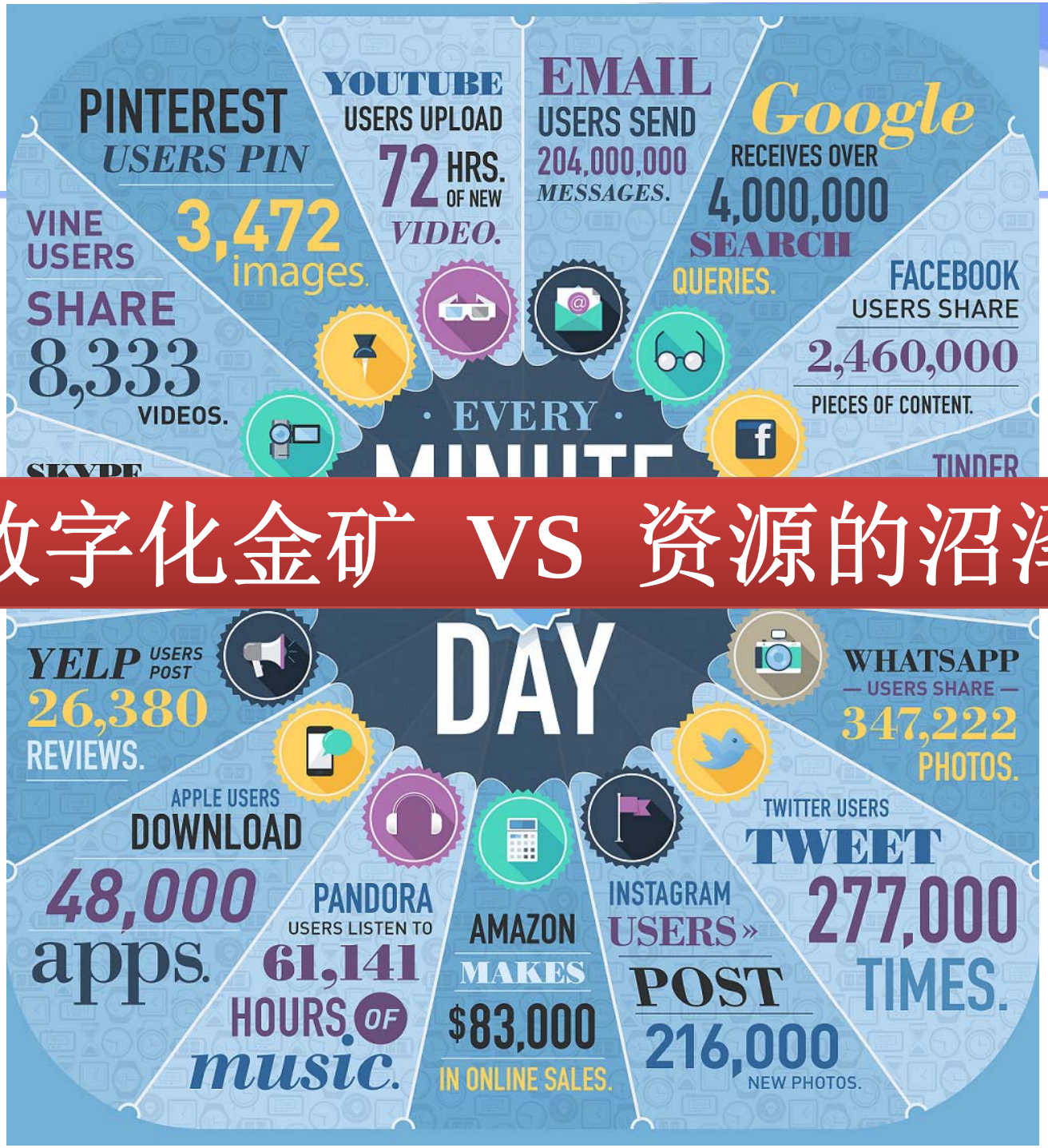
索引网页量超万亿
索引图像量几十亿

flickr

60亿张照片，标注词条超过1亿3千万条

YouTube

日均浏览量超过10亿次，全球平均每位网民每日都浏览了一段视频。



数字化金矿 VS 资源的沼泽

国家战略

US Government Big Data Strategy


REPORT TO THE PRESIDENT
AND CONGRESS
DESIGNING A DIGITAL FUTURE:
FEDERALLY FUNDED RESEARCH
AND DEVELOPMENT IN
NETWORKING AND INFORMATION
TECHNOLOGY

- White House has acknowledged the importance of Big Data

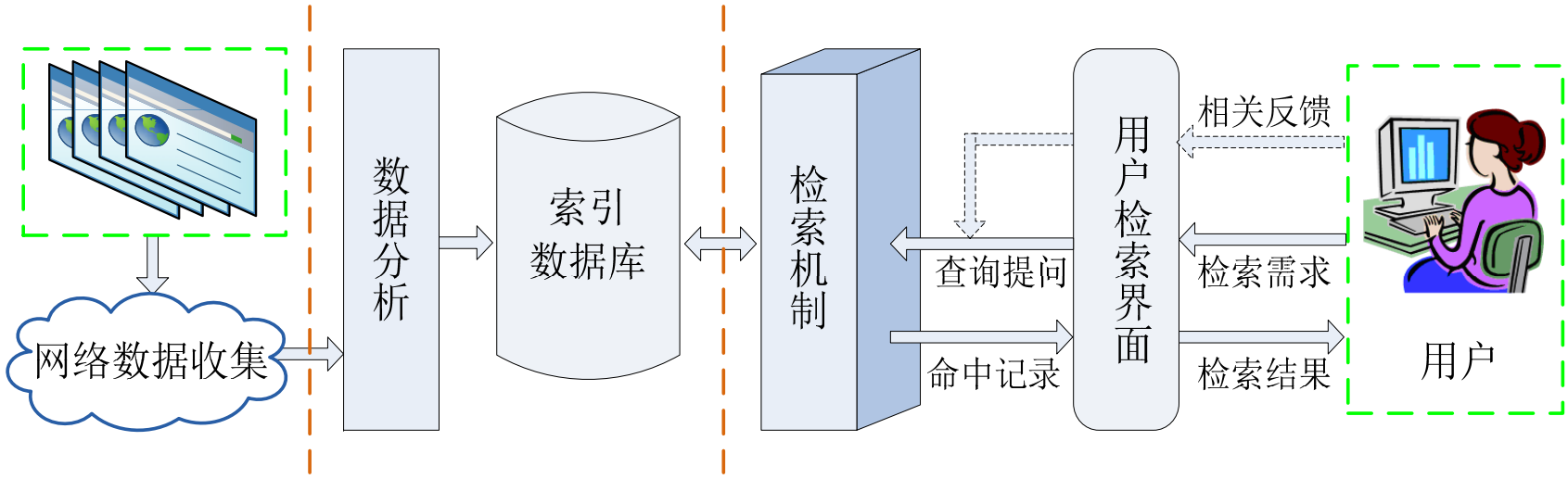
- **以大数据为代表的第三次工业革命已经来临：2012年3月，奥巴马政府将大数据定位为国家战略，并将数据定义为“未来的新石油”**
- **过去一年中，美国总统的”每日情报简报”共有1477个条目使用了网络大数据分析**
- **大数据是未来国家竞争力的利器。错失良机，将会拉大我国与发达国家之间的差距**

背景介绍

- 基于文本的检索
  - 依赖于手工标注
  - 检索结果相对准确
  - 耗时费力，且带有主观性
- 基于内容的检索
  - 依赖于数据对象的基本特征
    - 文本：向量形式
    - 图像：视觉特征
    - 视频：视觉+音频+字幕
  - 能够自动的实现索引
  - “语义鸿沟”问题：如何让计算机理解数据

“语义鸿沟”问题的图例说明

网络数据检索的基本流程

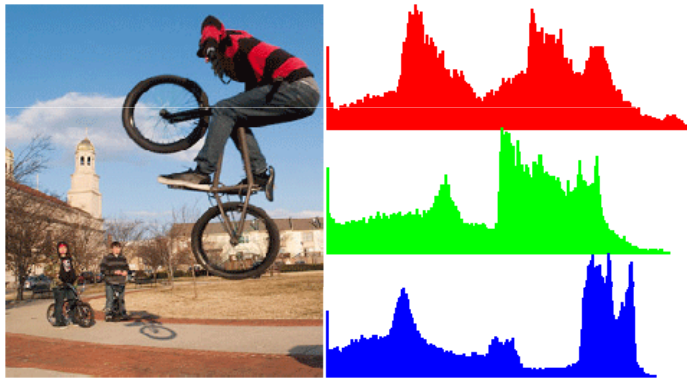


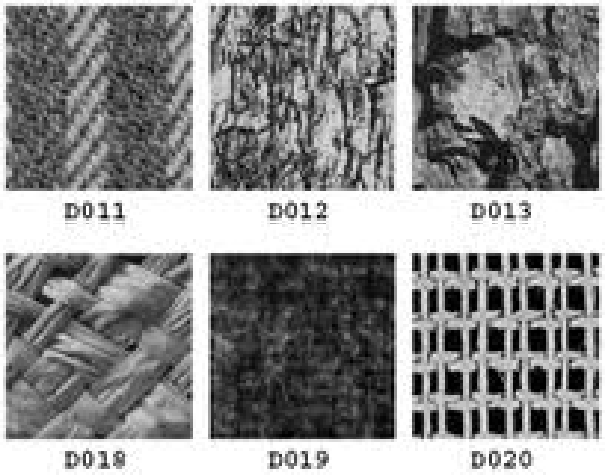
大 纲

- 背景介绍
- **索引与排序**
- 近似近邻搜索
- 总结与展望

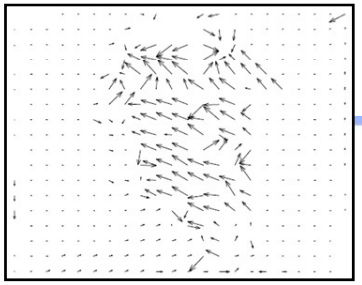
特征表示

- 低层视觉特征：颜色、纹理、形状



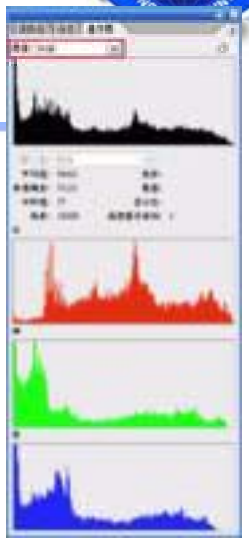


特征表示

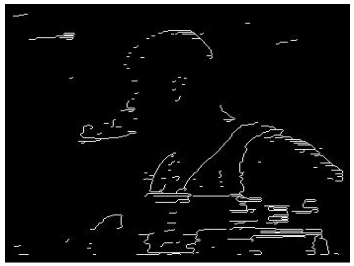


运动

颜色



语音



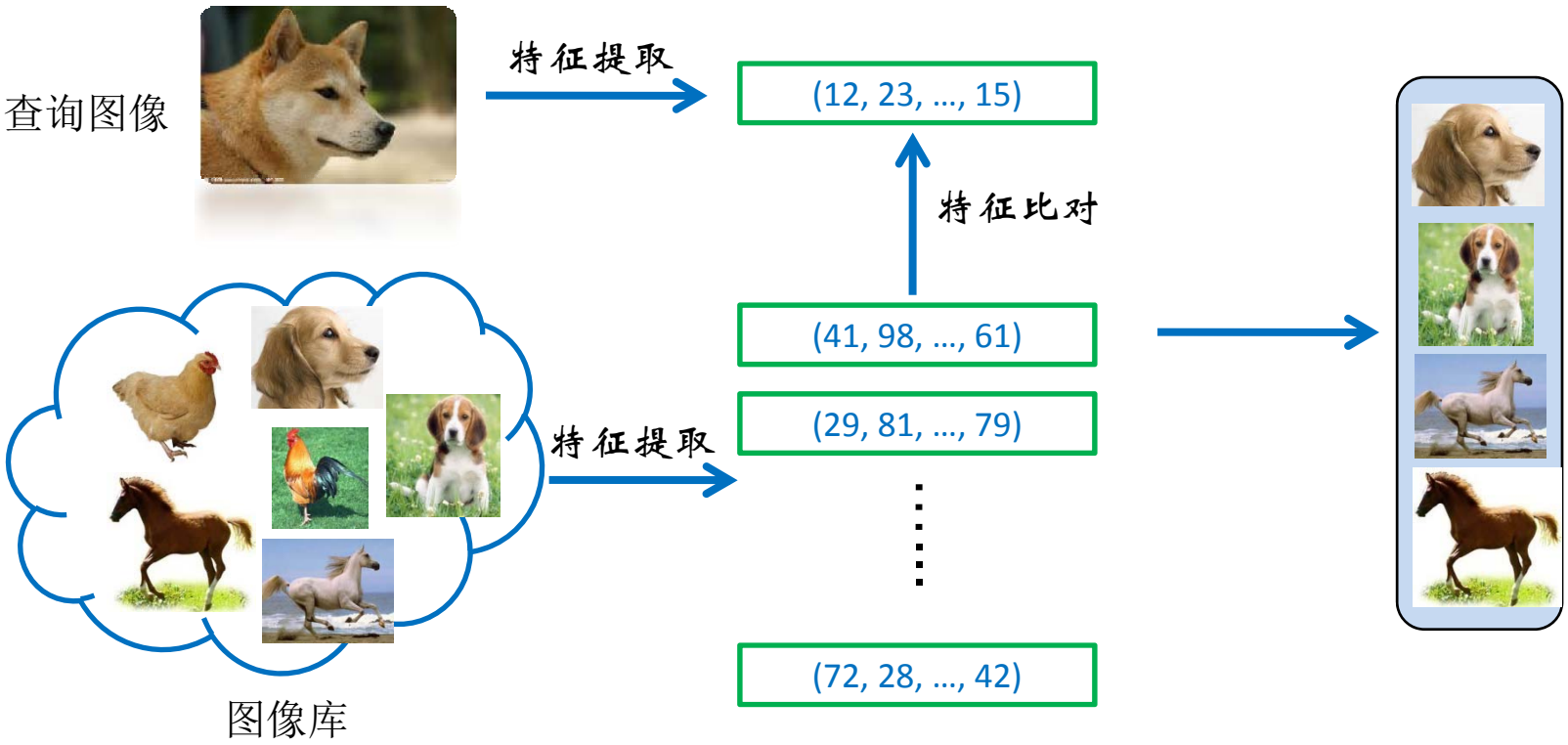
边缘

ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ

1234567890-=

字符

索引与排序



快速索引

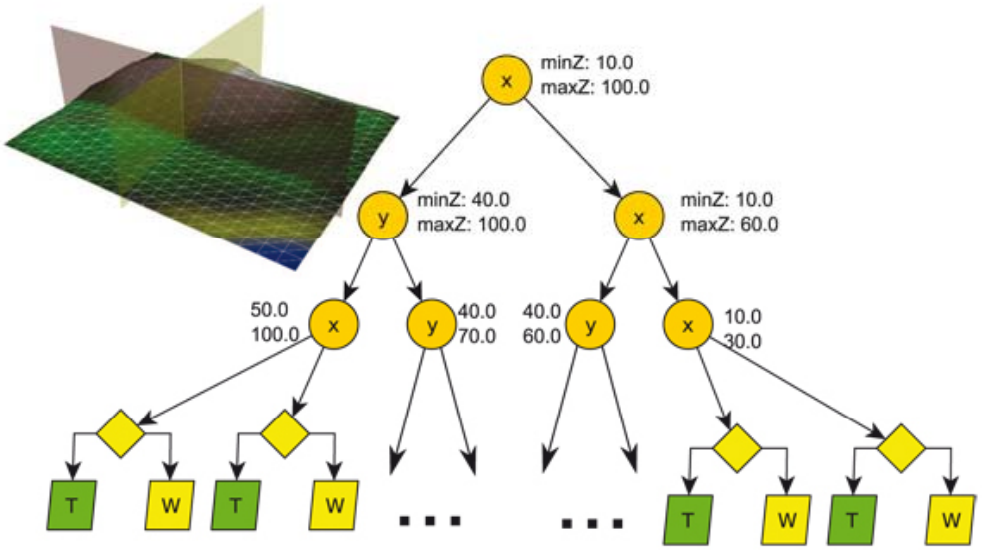
TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Document analysis

Tree-based



Curse of dimensionality

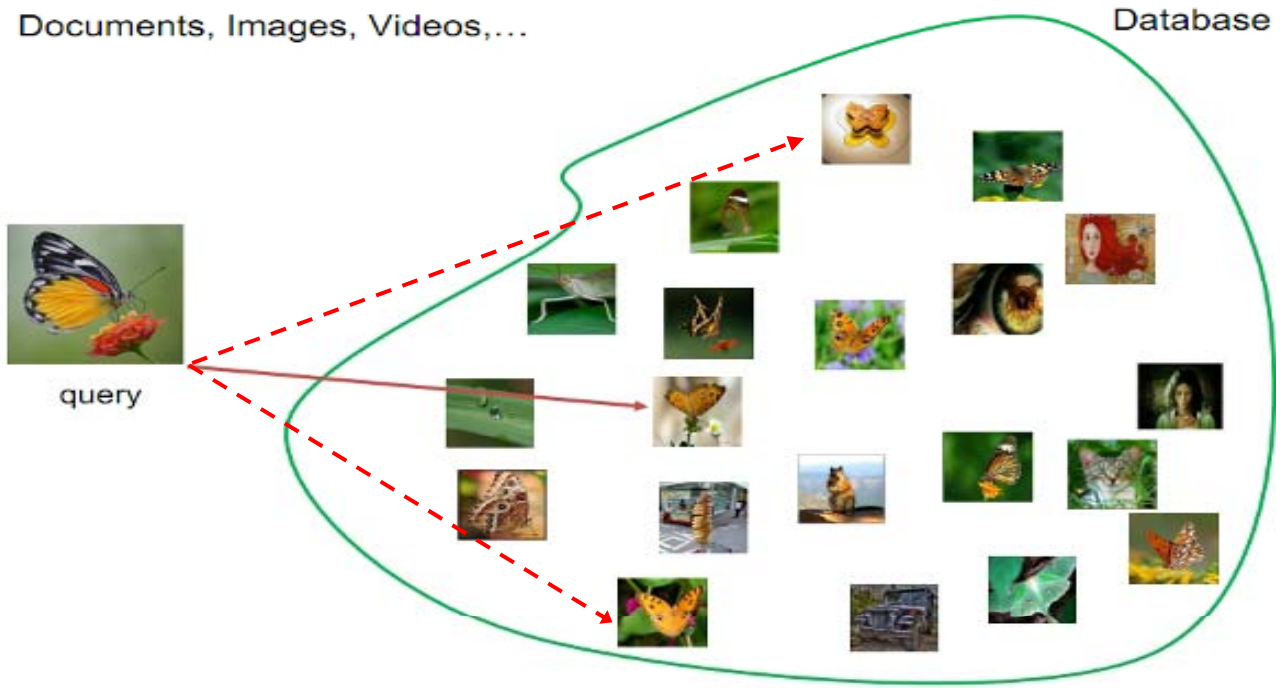
Pictures from web

大 纲

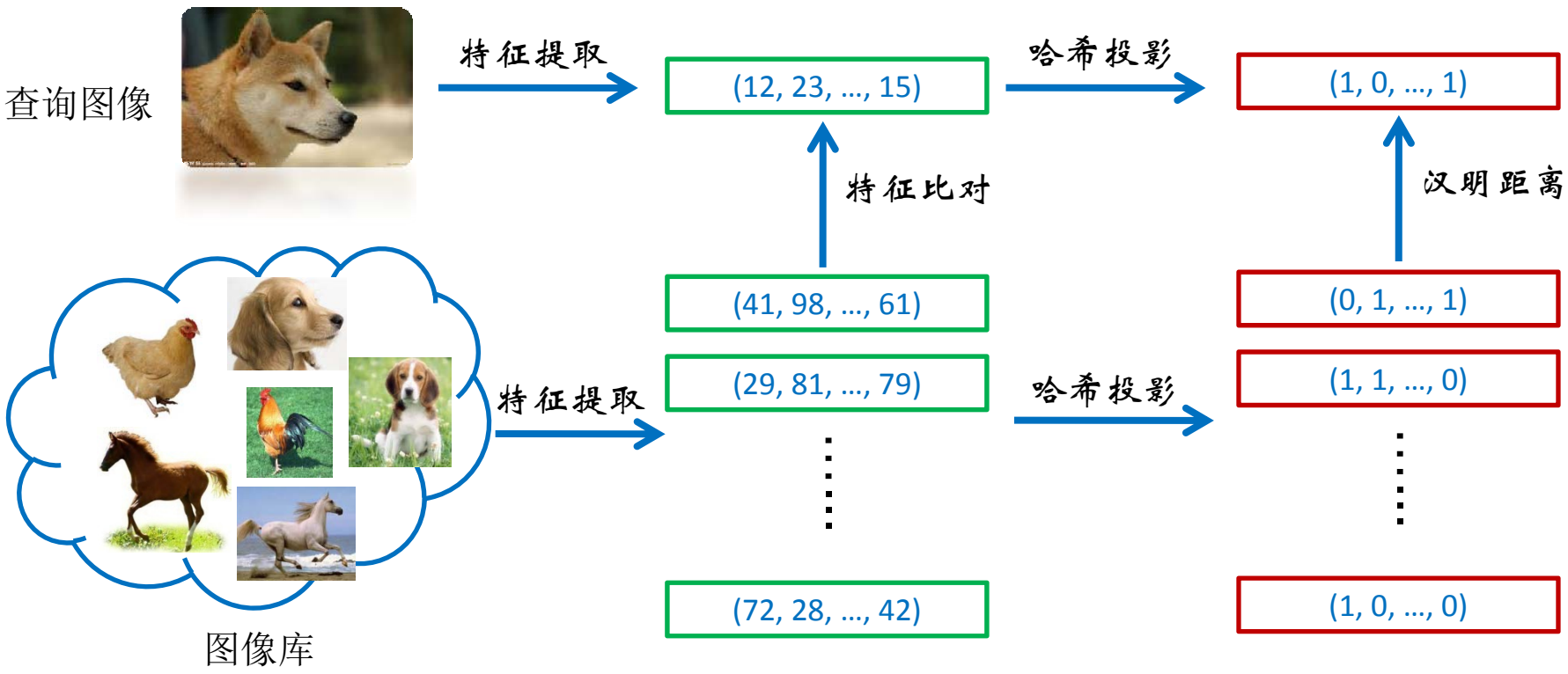
- 背景介绍
- 索引与排序
- **近似近邻搜索**
- 总结与展望

近似近邻搜索

- 近似近邻：检索精度和检索时间的平衡
- 对大多数应用，近似近邻足够满足需求

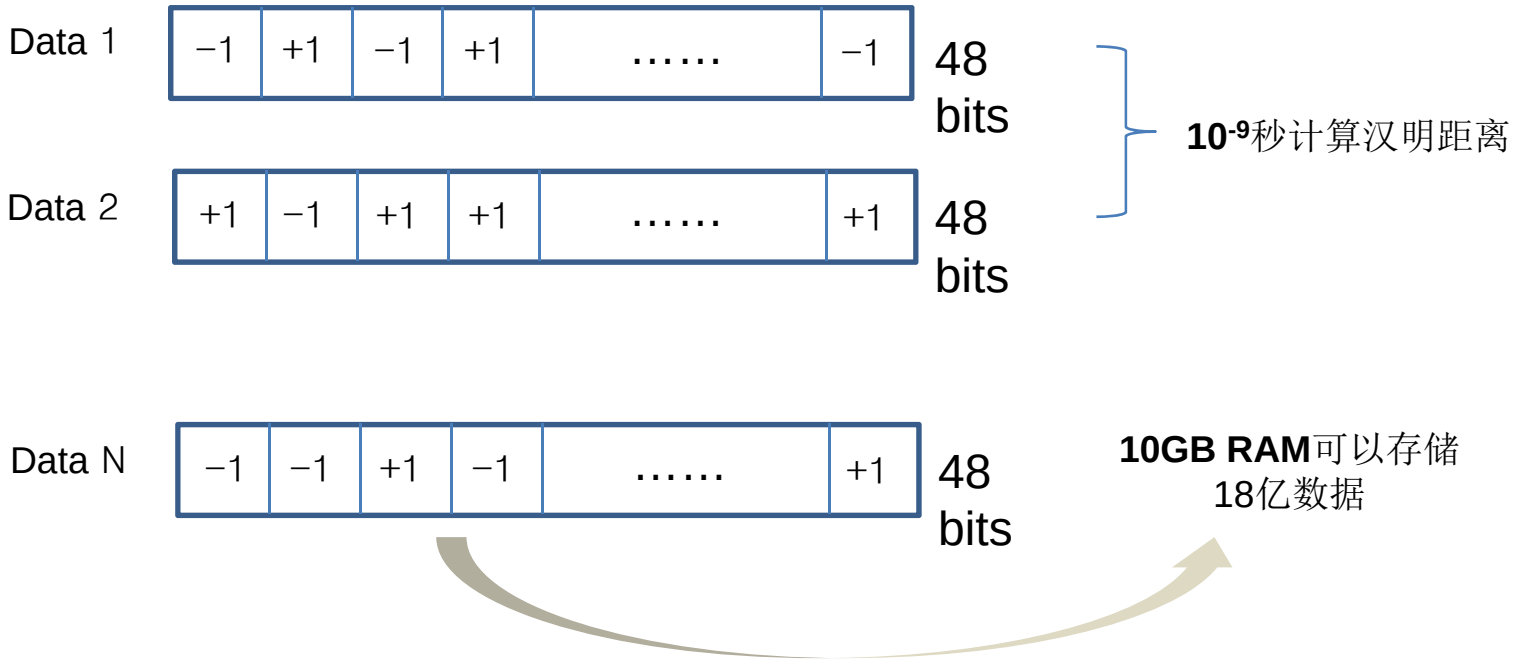


基于哈希的表示


查询图像
特征提取
(12, 23, ..., 15)
哈希投影
(1, 0, ..., 1)
特征比对
汉明距离
(41, 98, ..., 61)
(0, 1, ..., 1)
特征提取
(29, 81, ..., 79)
哈希投影
(1, 1, ..., 0)
(72, 28, ..., 42)
(1, 0, ..., 0)
图像库

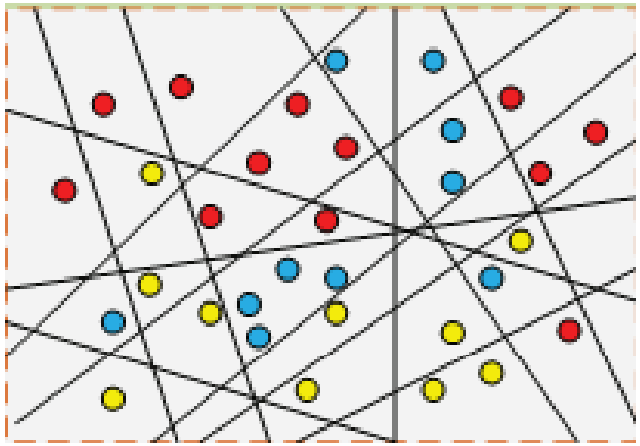
基于哈希的表示

- 时间上高效：基于XOR操作的快速计算
- 存储上高效：基于位存储的紧致表达



基于哈希的表示

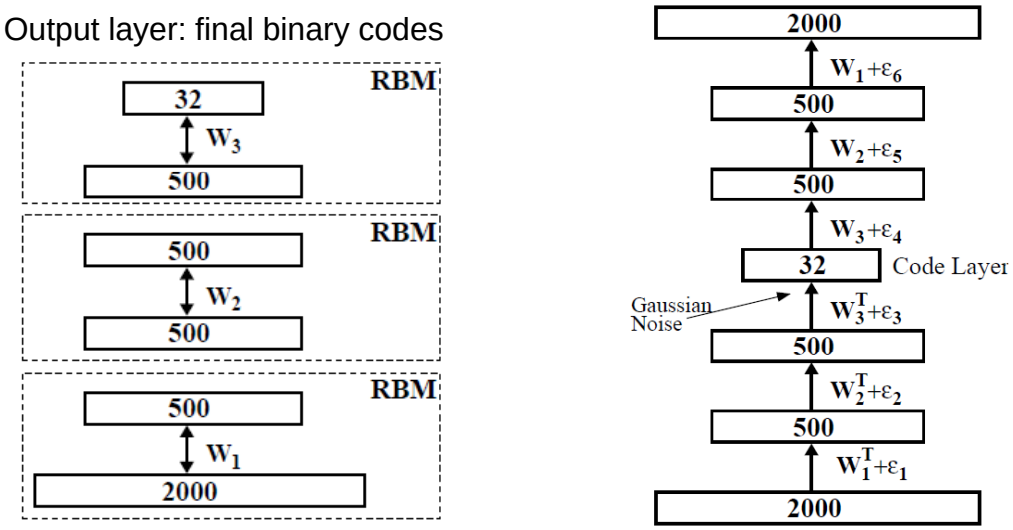
- Locality Sensitive Hashing (LSH)
  - ✓ Date-independent, unsupervised
  - ✓ Map the data points with random projections



P.Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. STOC'98

基于哈希的表示

- Restricted Boltzmann Machines (RBMs)
  - ✓ Date-dependent, unsupervised/supervised
  - ✓ Learn the binary codes layer by layer with deep networks



R.Salakhutdinov and G.Hinton. Semantic hashing. SIGIR 2007 workshop

基于哈希的表示

- Spectral Hashing (SH)
  - ✓ Date-dependent, unsupervised
  - ✓ Design compact binary codes based on spectral graph partitioning

$$
\begin{aligned}
&minimize: \sum_{ij} W_{ij}\|y_i - y_j\|^2 \\
&subject\ to: y_i \in \{-1,1\}^k \\
&\sum_i y_i = 0 \\
&\frac{1}{n}\sum_i y_i y_i^T = I
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
&minimize: trace(Y^T(D-W)Y) \\
&subject\ to: Y(i,j) \in \{-1,1\} \\
&Y^T 1 = 0 \\
&Y^T Y = I
\end{aligned}
$$

Y. Weiss, A.B. Torralba, and R. Fergus. Spectral hashing. NIPS 2008

基于哈希的表示

- Semi-Supervised Hashing (SSH)
  - ✓ Date-dependent, semi-supervised
  - ✓ Combines the empirical loss over the labeled data with other desirable constraints over both labeled and unlabeled data.

$$\max \; J(W) = \mathrm{tr}\,(W^T X_l S X_l^T W) + \eta * \mathrm{tr}\,(W^T X X^T W)$$

  - ✓ A sequential learning scheme (SPLH) is also developed for hash function leaning.

J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for large scale search. TPAMI 2012.

基于哈希的表示

- Extensions of traditional hashing methods:
  - Hyper-graph Spectral Hashing (HSH), Anchor Graph Hashing (AGH)
  - Optimized Kernel Hashing (OKH), Supervised Kernel Hashing (SKH)
  - Composite Hashing with Multiple Information Sources (CHMS), Kernel Hashing with Multiple Features (MFKH)

Y. Zhuang, Y. Liu, F. Wu, Y. Zhang, and J. Shao. “Hyper-graph spectral hashing for similarity search of social image” In ACM MM 2011.
W. Liu, J. Wang, S. Kumar, and S.-F. Chang. “Hashing with graphs” In ICML 2011.
J. He, W. Liu, and S.-F. Chang. “Scalable similarity search with optimized kernel hashing” In KDD 2010.
W. Liu, J. Wang, R. Ji, Y. Jiang, and S.-F. Chang. “Supervised hashing with kernels” In CVPR 2012.
D. Zhang, F. Wang, and L. Si. “Composite Hashing with multiple information sources” In SIGIR 2011.
X.liu, J.HE, D.Liu, B.Lang. “Compact kernel hashing with multiple features.” In ACM MM 2012.

在线学习哈希方法

在实际检索系统中数据是流式的、不断更新的

在线更新

大规模

对于超大规模数据，很难将其载入内存进行哈希函数学习

在线学习哈希方法

全部数据 = 之前数据 + 新数据

- **综合所有新老数据，重新学习**
- **超高的时间和计算代价**

在线学习: Passive-Aggressive (PA) 算法 (JMLR'06)

$$W^t = \arg\min_W \frac{1}{2}\|W - W^{t-1}\|_F^2 + C\xi$$

$$\text{s.t. } L(W; \langle (x_i^t, x_j^t), s_{ij} \rangle) \leq \xi \text{ and } \xi \geq 0$$

Online Hashing (IJCAI'13)

Long-Kai Huang, Qiang Yang, Wei-shi Zheng. Online Hashing. IJCAI 2013

在线学习哈希方法

- 矩阵素描(sketching or coreset)

  ✓一个比原始矩阵小的多的矩阵

  ✓保留原始矩阵的大多数特性

$$\forall x,\ \|x\| = 1 \qquad |\|P^T x\|^2 - \|Q^T x\|^2| \leq \varepsilon \|P\|_F^2$$

在线学习哈希方法

Frequent Directions (FD)

Algorithm 1 Frequent Directions (Liberty [21])

Input: Data matrix $P \in \mathbb{R}^{d \times n}$, sketch matrix $Q \in \mathbb{R}^{d \times l}$.
Output: Sketch matrix Q.

```
if Q not exists then
    Q ⟵ all zeros d × l matrix
end if
for each column P_i in P, do
    Insert P_i into a zero valued column of Q
    if Q has no zero valued columns then
        [U, S, V] = SVD(Q)
        \\ C = US    [just for notation]
        Set δ = s²_{l/2}    [the squared (l/2)^th entry of S]
        Set Ŝ = sqrt(max(S² − I_l δ, 0))
        Q = UŜ
    end if
end for
```

Lemma 1. (Liberty [21]) Apply Algorithm 1 to matrix P to obtain a sketch Q with prescribed l, then

$$\forall x,\ \|x\| = 1 \qquad 0 \leq \|P^T x\|^2 - \|Q^T x\|^2 \leq \frac{2}{l}\|P\|_F^2$$

or

$$0 \leq \|PP^T - QQ^T\|_2 \leq \frac{2}{l}\|P\|_F^2$$

Edo Liberty, "Simple and Deterministic Matrix Sketching". SIGKDD 2013 (Best paper award)

在线学习哈希方法

广泛意义上的PCA Hashing:

$$\max_{W \in \mathbb{R}^{d \times r}} \quad tr(W^T (X - \mu)(X - \mu)^T W)$$

$$\text{s.t.} \quad W^T W = I_r$$

其中 μ 是全体数据的均值向量.

能否为矩阵 $X - \mu$ 在线维护一个素描矩阵 Y，以至于

$$YY^T \approx (X - \mu)(X - \mu)^T$$

在线学习哈希方法

- 均值漂移问题:
  - 在流式问题中，数据在不断更新，因此数据的均值 μ 也会不断变化.
- 给每个数据块加一个虚拟的样本：
  - 对于流式数据 $X_t = [D_1, D_2, \ldots, D_t]$，重新设计数据矩阵 E_t:

$$E_t = [\mathcal{D}_1 - \overline{\mathcal{D}_1}, \quad \mathcal{D}_2 - \overline{\mathcal{D}_2}, \sqrt{\frac{n_1 m_2}{n_1 + m_2}}(\overline{\mathcal{D}_2} - \mu_1), \cdots,$$
$$\mathcal{D}_i - \overline{\mathcal{D}_i}, \sqrt{\frac{n_{i-1} m_i}{n_{i-1} + m_i}}(\overline{\mathcal{D}_i} - \mu_{i-1}), \cdots,$$
$$\mathcal{D}_t - \overline{\mathcal{D}_t}, \sqrt{\frac{n_{t-1} m_t}{n_{t-1} + m_t}}(\overline{\mathcal{D}_t} - \mu_{t-1})]$$

在线学习哈希方法

- 基于以上的设计，在任意适合 t, 可以证明：

$$E_t E_t^T = cov(X_t)$$

- 通过为新设计的矩阵 E_t 在线维护素描矩阵 Y, 我们可以基于这个新的小矩阵 Y学习哈希函数.

Algorithm 2 Zero Mean Sketching

Input: Streaming data chunk $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_k$,
All zeros matrix Y of size $d \times l$.

1: Sketch $\mathcal{D}_1 - \overline{\mathcal{D}_1}$ into Y with Algorithm 1
2: $n \leftarrow m_1$ and $\mu \leftarrow \overline{\mathcal{D}_1}$
3: **for** $i = 2 : k$, **do**
4: Sketch $[\mathcal{D}_i - \overline{\mathcal{D}_i}, \sqrt{\frac{nm_i}{n+m_i}}(\overline{\mathcal{D}_i} - \mu)]$ into Y
5: $\mu \leftarrow \frac{n\mu}{n+m_i} + \frac{m_i \overline{\mathcal{D}_i}}{n+m_i}$ [update the data mean]
6: $n \leftarrow n + m_i$ [update the data size]
7: **end for**

对比实验

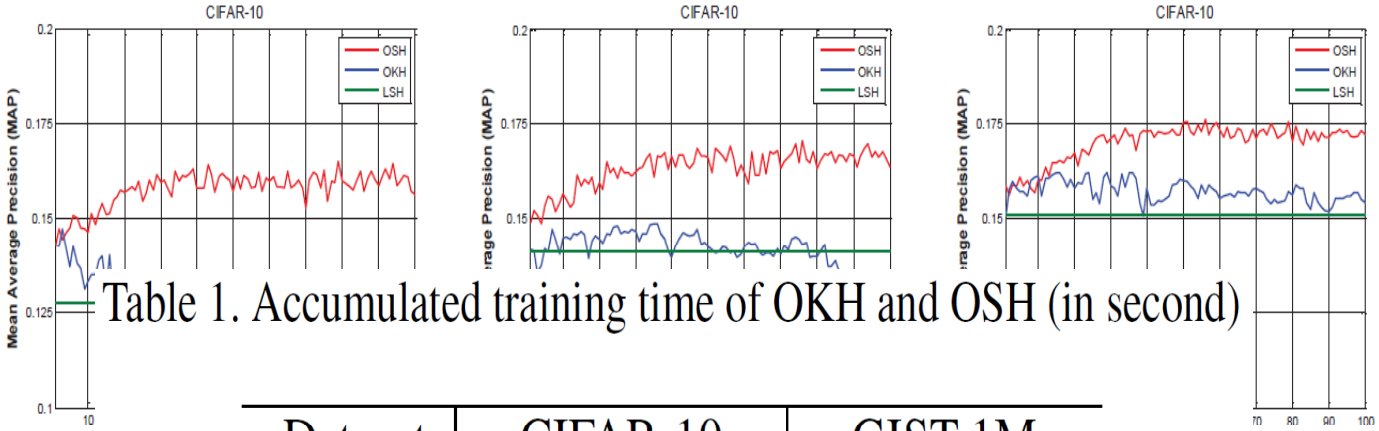


Table 1. Accumulated training time of OKH and OSH (in second)

Dataset	CIFAR-10	GIST-1M
OKH	31.96	490.73
OSH	8.23	180.74

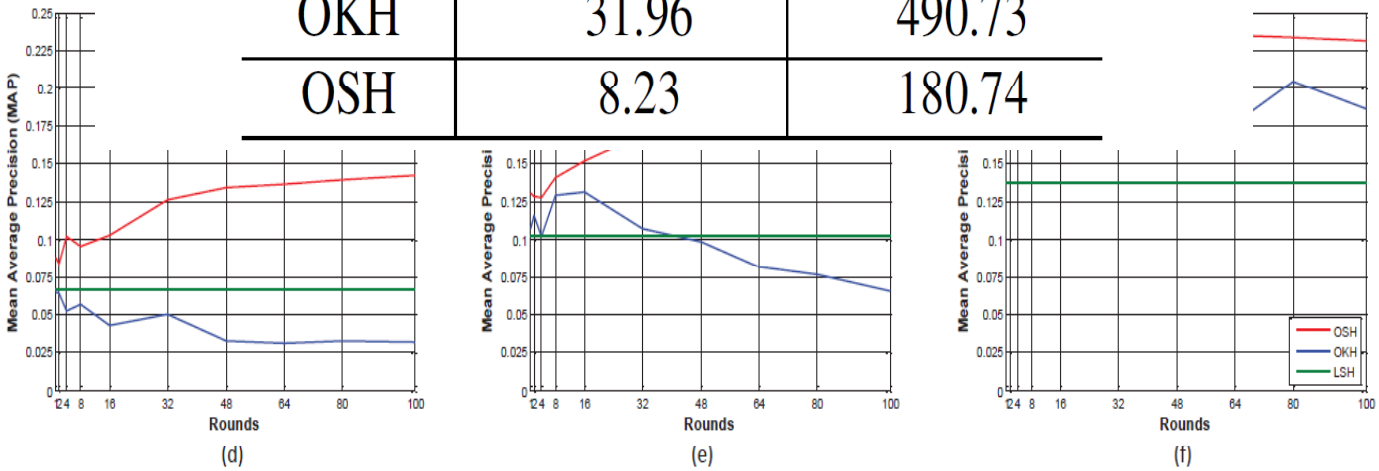
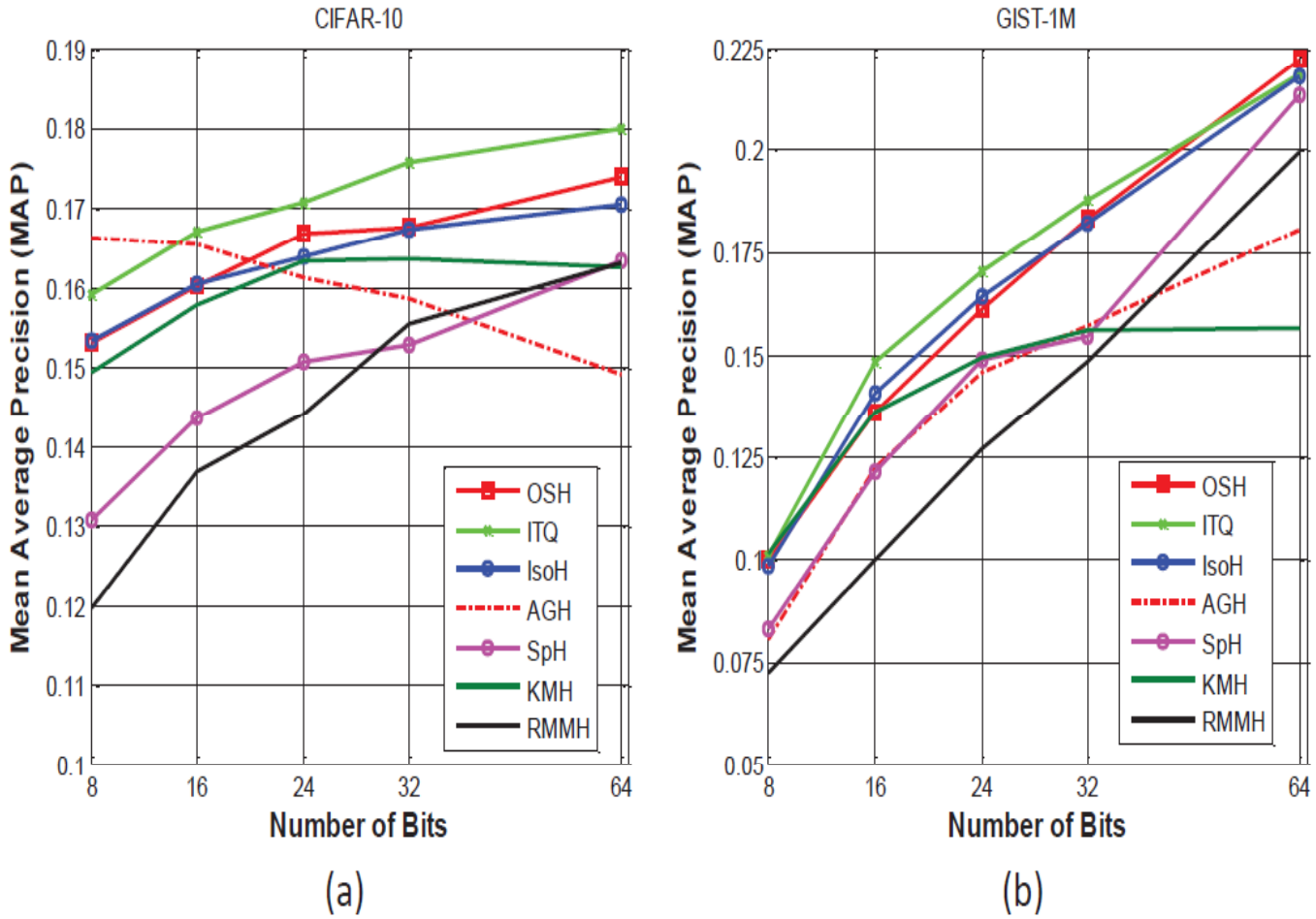


Figure 3. (a)(b)(c) Mean average precision (MAP) on CIFAR-10 dataset at each round with 16, 32, 64 bits. (d)(e)(f) MAP on GIST-1M dataset at each round with 16, 32, 64 bits. (Best viewed in color)

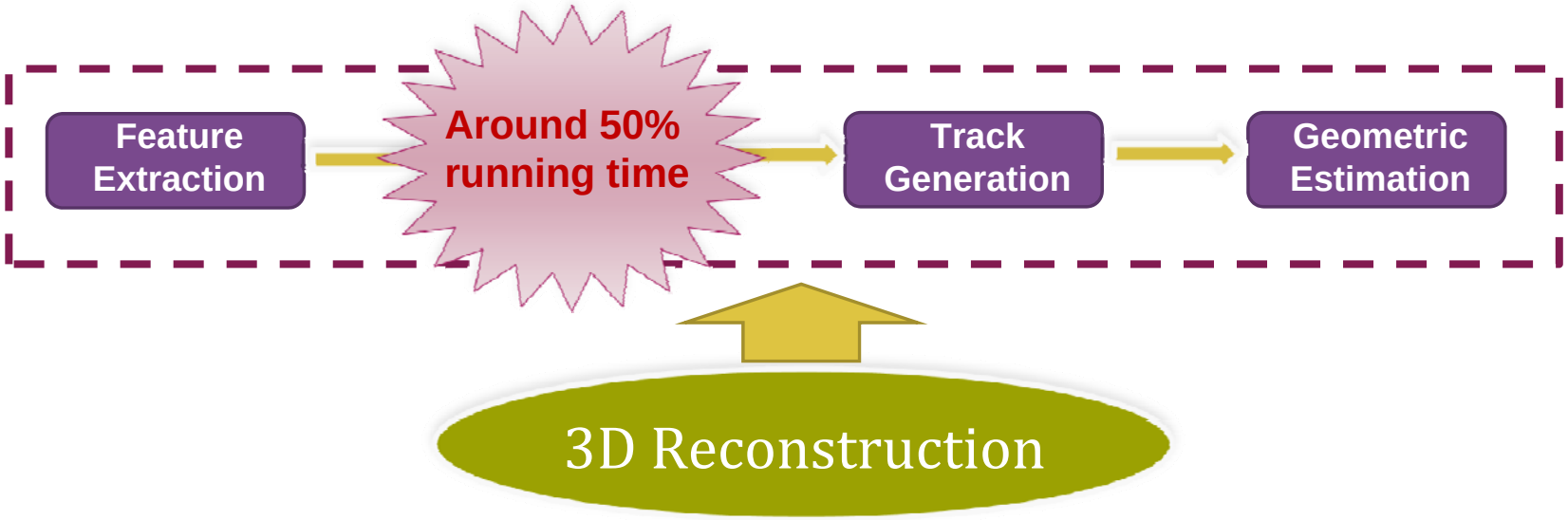
对比实验



Cong Leng et al., “Online Sketching Hashing”. CVPR 2015

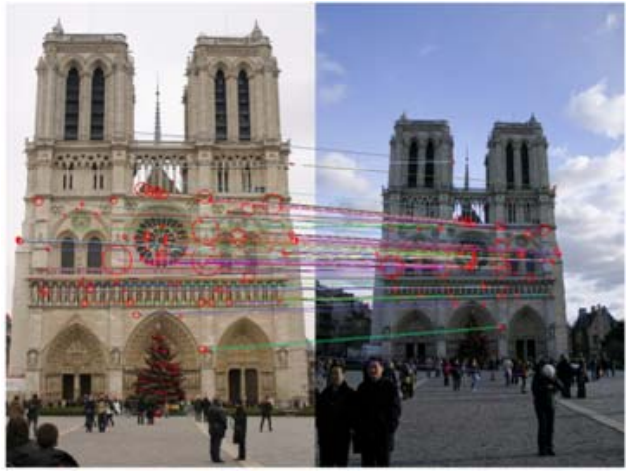
基于哈希的特征匹配

- 三维重建主要包括以下四步:



# Images	# Cores	Match Time	Reconstruction Time	Largest Component
150,000	496	13 Hours	8 Hours	2,106

基于哈希的特征匹配



特征匹配

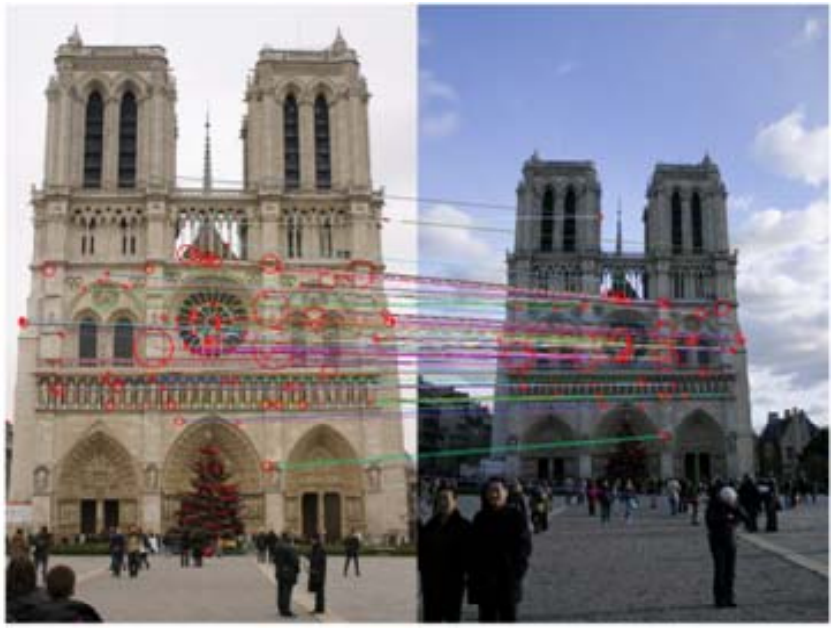
- 图像搜索
- 三维重建
- 图像分类
- 目标识别

典型应用

- **匹配复杂度太高**：$O(N*(N-1)*M^2)$
- **硬件加速成本高**：GPU、并行

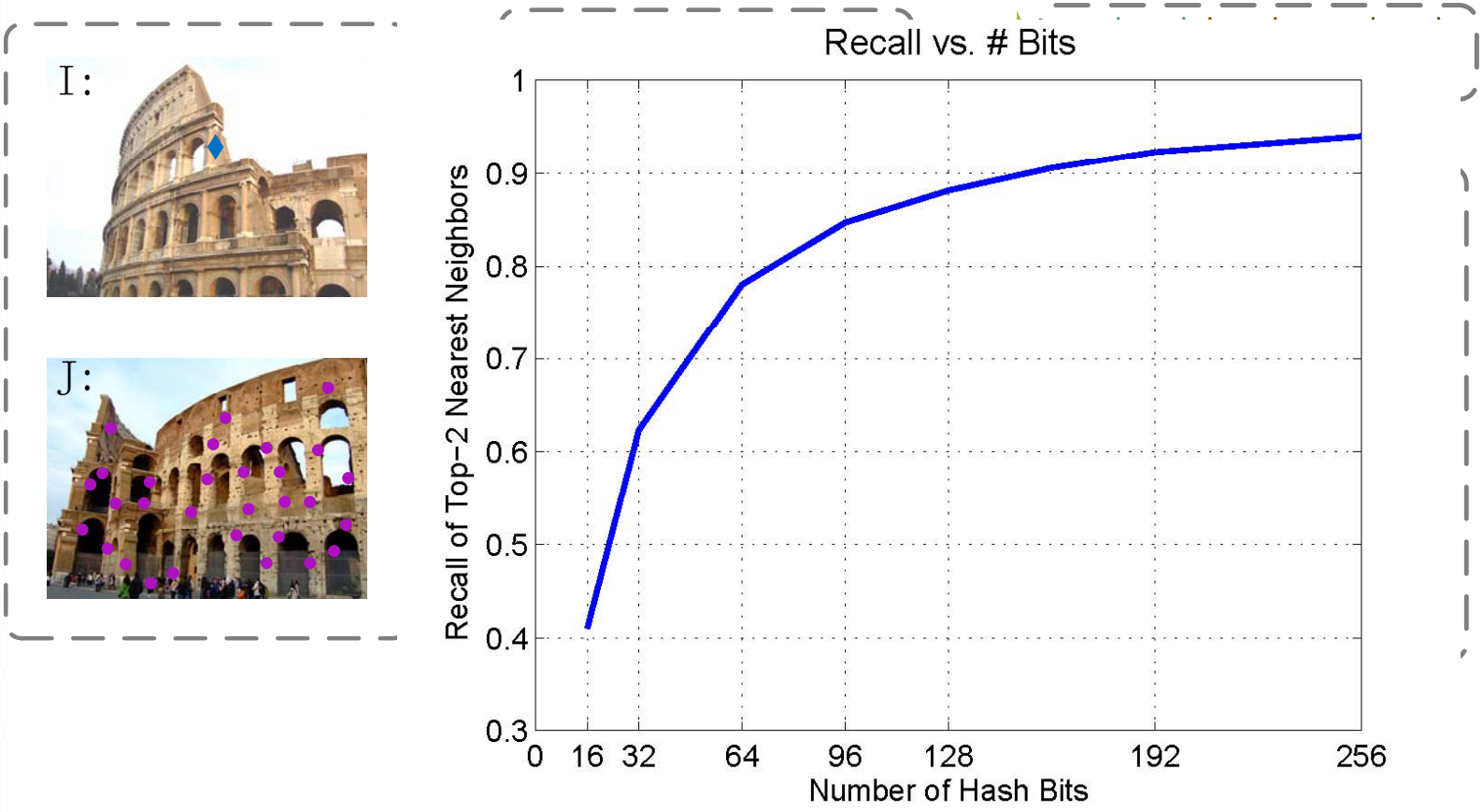
基于哈希的特征匹配

- 现有特征匹配方法可以分为三大类:
  - Point matching
  - Line matching
  - Region matching



Point matching is searching in essence!

基于哈希的特征匹配



Jian Cheng et al., “Fast and Accurate Image Matching with Cascade Hashing for 3D Reconstruction ”. CVPR 2014

基于哈希的特征匹配

(a) Taj.Mahal: 189 images.

(b) Statue.of.Liberty: 674 images.

(c) Notre.Dame: 715 images.

(d) Colosseum: 1357 images.

基于哈希的特征匹配

Method	Taj.Mahal			Statue.of.Liberty		
	T_{Match}	Speed-Up	Points	T_{Match}	Speed-Up	Points
Brute	52575s	1.00×	**124038**	30608s	1.00×	88001
KDTree	5136s	10.24×	119165	9765s	3.13×	98674
LDAHash	1774s	29.64×	120353	874s	35.02×	123274
CasHash-8Bit	301s	174.67×	118331	358s	85.50×	203587
CasHash-10Bit	**183s**	**287.30**×	116224	**231s**	**132.50**×	**246206**
Method	Notre.Dame			Colosseum		
	T_{Match}	Speed-Up	Points	T_{Match}	Speed-Up	Points
Brute	396729s	1.00×	358121	12307s	1.00×	**540308**
KDTree	60663s	6.54×	347056	2430s	5.06×	445774
LDAHash	13136s	30.20×	413348	851s	14.46×	492040
CasHash-8Bit	2266s	175.08×	**484960**	222s	55.44×	393408
CasHash-10Bit	**1354s**	**293.01**×	400673	**196s**	**62.79**×	512508

基于哈希的特征匹配

通过测试确信它在不同数据集上都非常有效

OpenMVG (open Multiple View Geometry)

I'm actually testing it to ensure it works well on different datasets and that I have a code that works as good as your original version.

https://github.com/openMVG/openMVG/issues/194

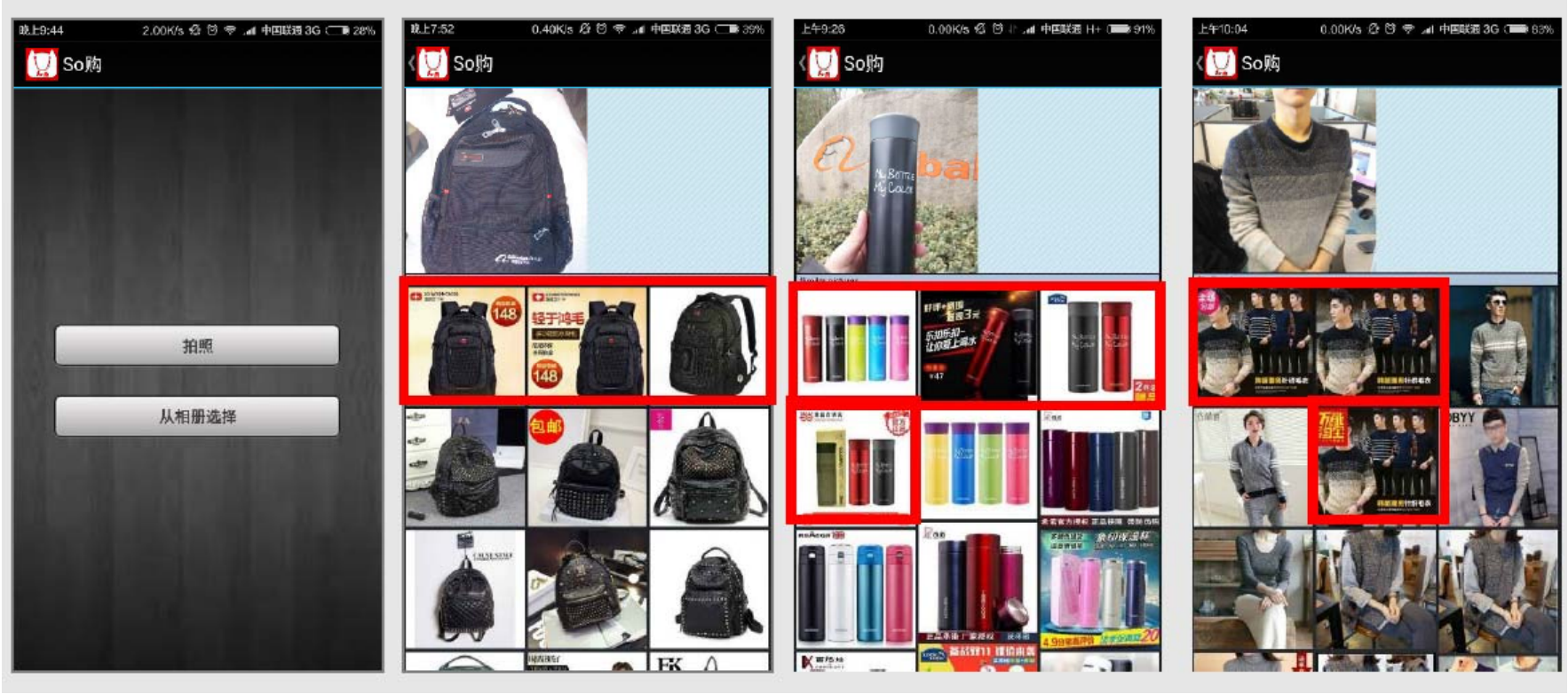
Theia Vision Library

这是非常有用的见解和足够简单的想法

it's a really useful insight and is a simple enough idea.

https://github.com/kip622/Theia

手机搜索



- 数据库：约300万幅，含衣服、鞋、箱包、配饰、瓶饮、美妆、零食、家具8类
- 硬件配置：NVDIA TITAN Black 6GB / INTEL Core i7 3.5GHz / 16GB 内存
- 检索时间：100ms

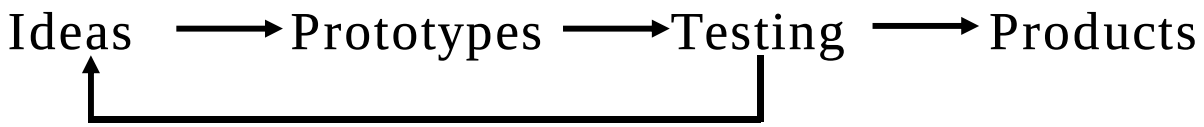
Demo

大 纲

- 背景介绍
- 索引与排序
- 近似近邻搜索
- **总结与展望**

总结与展望

- 网络技术的发展模式 – 呈螺旋式上升

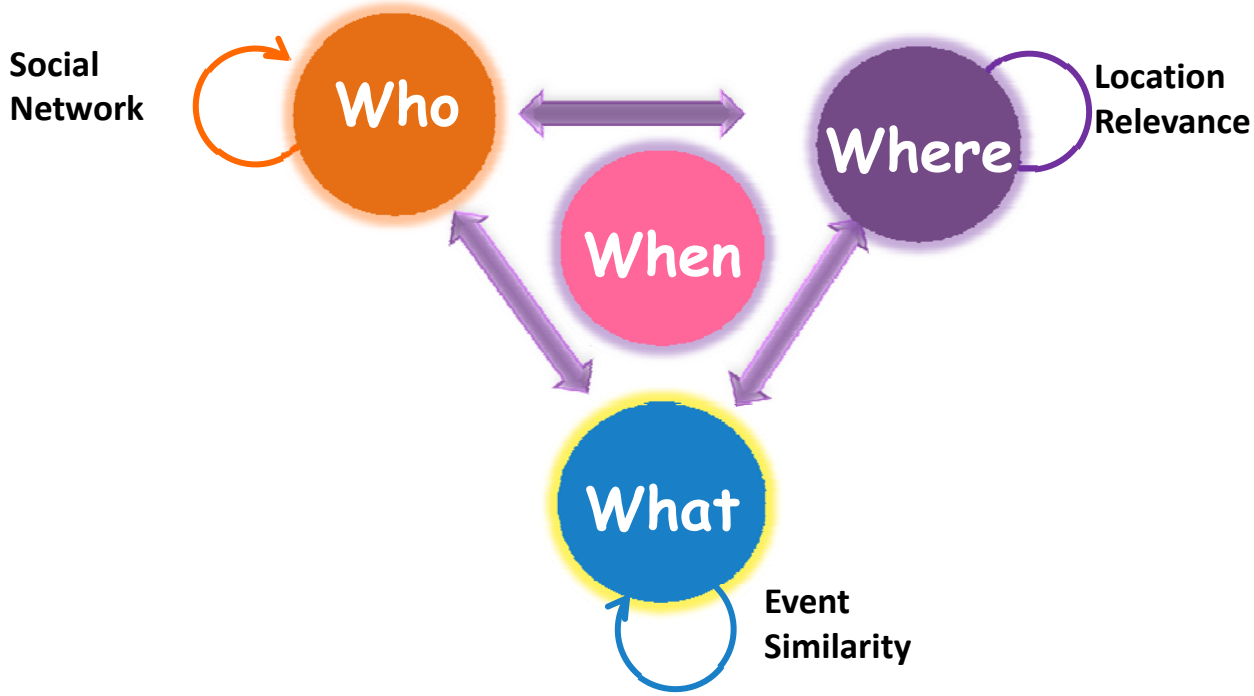


- 面向Billion量级数据的高效实验平台
  - 算法的自动化与可扩展性 /存储空间/内存空间/计算效率
- 机器学习与数据挖掘技术
  - Learning to ……
    - Learning to crawling
    - Learning to extracting
    - Learning to ranking
    - ……

发展趋势 - 协同化

多关系协同的搜索

4W – who, where, what, and when



发展趋势 – 社交化


facebook
Home
Videos
Channels
Community
Search
to President Bush
About This Video
From: TheFactor26
Related Videos
Promoted Videos
Rate:
Views: 900,048
& Responses
Post a video response
Post a text comment
twitter
follow me
开心网
Linked in
新浪微博1周年
t.sina.com.cn
校内网
xiaonei.com

发展趋势 – 泛在化

编辑

分享

家庭录像

发布

分享

检索

网络视频

数字媒体自动标注和个性化定制

录制

定制

摘要

电视机顶盒

浏览

定制

传播

移动通信设备

Thanks!

Q&A