Data Collection

——Data Acquisition——

"Data Mining: Methods and Applications"

1

Contents

- Background
- Data Acquisition
  - Data Discovery
  - Data Augmentation
  - Data Generation
- Data Labeling
- Improvement of Existing Data and Models
- How to Decide which Data Collection Techniques to Use When
- Interesting Future Research Challenge

2

Data Acquisition

- **The goal of data acquisition is to find datasets that can be used to mine data.**
- **Three approaches in the literature: data discovery, data augmentation, and data generation.**
  - **Data discovery** is necessary when one wants to share or search for new datasets and has become important as more datasets are available on the Web and corporate data lakes.
  - **Data augmentation** complements data discovery where existing datasets are enhanced by adding more external data.
  - **Data generation** can be used when there is no available external dataset, but it is possible to generate crowdsourced or synthetic datasets instead.
- **Classification of Data Acquisition**

3

Data Acquisition

A classification of data acquisition techniques. Some of the techniques can be used together. For example, data can be generated while augmenting existing data.

Task	Approach	Techniques
Data discovery	Sharing	Collaborative Analysis [9]–[11] Web [12]–[17] Collaborative and Web [18]
	Searching	Data Lake [19], [19]–[23] Web [24]–[34]
Data augmentation		Deriving Latent Semantics [35]–[37] Entity Augmentation [30], [31] Data Integration [38]–[44]
Data generation	Crowdsourcing	Gathering [45]–[54] Processing [49], [50], [55], [56]
	Synthetic Data	Generative Adversarial Networks [57]–[62] Policies [63], [64] Image [65]–[71] Text [72]–[74]

4

Data Acquisition——Data Discovery

- **Two Steps of Data Discovery:**
  - **Data sharing**: the generated data must be indexed and published for sharing. (*Post-hoc Method*)
  - **Data searching**: search the datasets for the data mining tasks. (how to scale the searching and how to tell whether a dataset is suitable for a given machine learning task)

5

Data Acquisition——Data Discovery

- **Data Sharing:** The data sharing systems are platforms for sharing datasets.
  - *Collaborative Analysis* : In an environment where data scientists are collaboratively analyzing different versions of datasets, **DataHub** can be used to host, share, combine, and analyze them.
    - 2 system components: dataset version control system inspired by Git (a version control system for code) and a hosted platform on top of it, which provides data search, data cleaning, data integration, and data visualization.
  - *Web* : A different approach of sharing datasets is to publish them on the Web. **Google Fusion Tables** is a cloud based service for data management and integration.
  - *Collaborative and Web* : **Kaggle** is a merging of collaborative and Web-based systems

6

Data Acquisition——Data Discovery

- **Data Searching:** We want to explore systems that are mainly designed for searching datasets.
  - ***Data Lake***: Providing a way to search datasets and analyze them has significant business value because the teams or individuals do not have to make redundant efforts to re-generate the datasets for their data mining tasks.
    - IBM estimates that **70%** of the time spent on analytic projects is concerned with discovering, cleaning, and integrating datasets that are scattered among many business applications. Thus, IBM takes the stance of creating, filling, maintaining, and governing the data lake where these processes are collectively called *data wrangling* (数据整理).

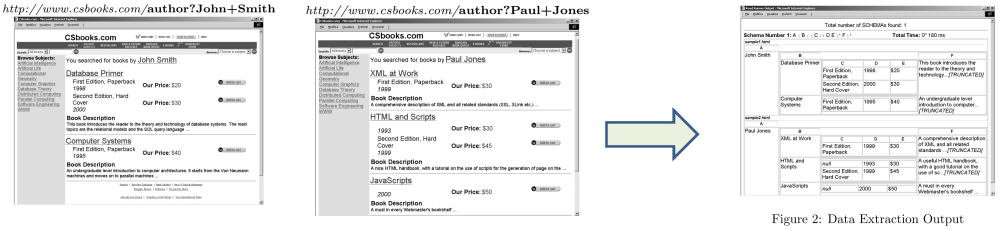
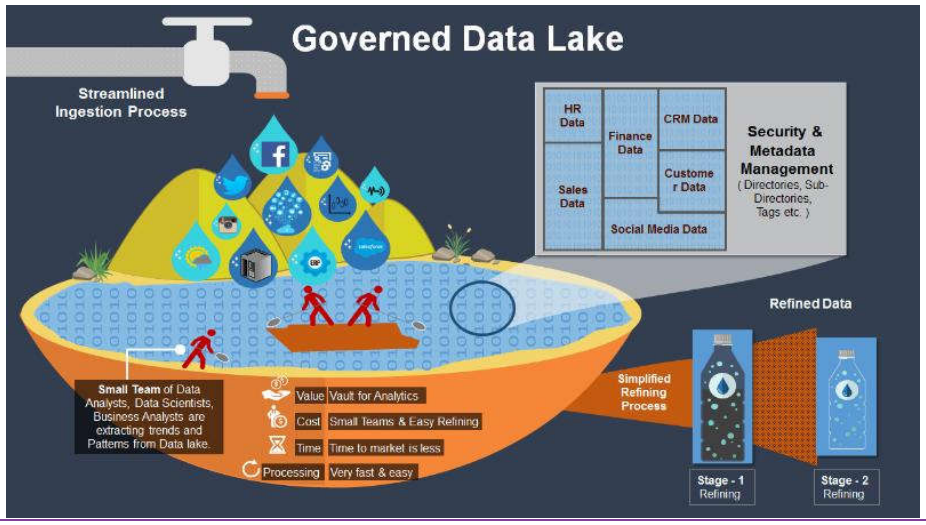


Figure 2: Data Extraction Output

Data Acquisition

Data Lake



Data Acquisition——Data Discovery

- Recently, scalability has become a pressing issue for handling data lakes that consists of most datasets in a large company. **Google Data Search** (*GOODS*) is a system that catalogues the metadata of tens of billions of datasets from various storage systems within Google.
- Expressive queries are also important for searching a data lake. While GOODS scales, one downside is that it only supports simple keyword queries: The DATA CIVILIZER system, DATARAMAN (*discovery queries*) and AURUM.

9

Data Acquisition——Data Discovery

- **Data Searching**
  - *Web* : automatically extract the useful ones. **WebTables** extracts all Wikipedia infoboxes. A service called *Google Dataset Search* was launched for searching repositories of datasets on Web. The motivation is that there are thousands of data repositories on the Web that contain millions of datasets that are not easy to search.
  - In comparison to *GOODS*, Dataset Search targets the Web instead of a *data lake*.

Figure 1 - The process to make an image search.

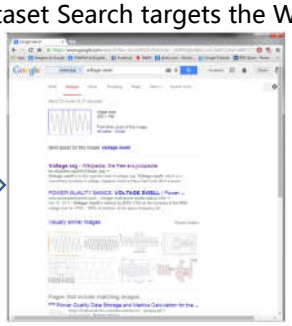


Figure 2 - The search result.

Google Image Search

10

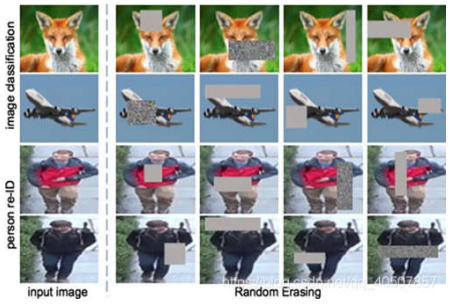
Data Acquisition——Data Augmentation

- **Data Augmentation**: Another approach to acquiring data is to augment existing datasets with external data.
  - ***Adding pre-trained embeddings*** is a common way to increase the features to train on.
  - ***Entity augmentation techniques*** have been proposed to further enrich existing entity information.
  - ***Data integration*** is a broad topic and can be considered as data augmentation if we are extending existing datasets with newly-acquired ones.
- *Deriving Latent Semantics*
  - General Language Model*: Word2Vec* (CBOW , Skip-gram) *, GloVe, Doc2Vec.*
  - Latent Topic Modeling: Latent Dirichlet Allocation (LDA)
- *Entity Augmentation*
  - *Octopus and InfoGather:*

11

Data Acquisition——Data Augmentation

- *Data Integration:* especially if we are extending existing data sets with other acquired ones.
  - *Hamlet System, Hamlet++ systems*



12

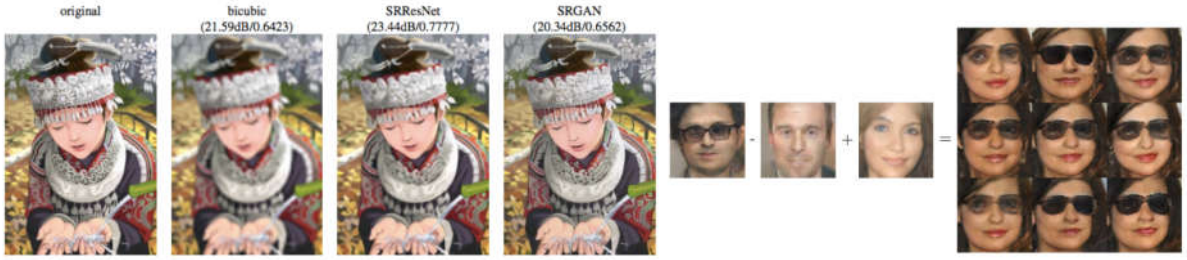
Data Acquisition——Data Generation

- **Data Generation:** If there are no existing datasets that can be used for training, then another option is to generate the datasets either manually or automatically.
  - Manual Construction : crowdsourcing (众包)
  - Automatic Techniques
- *Crowdsourcing:* Amazon Mechanical Turk (HITs) and **ImageNet** Project
  - Data generation using crowdsourcing can be divided into two steps: *gathering* data and *preprocessing* data.
    - *Data gathering :* is not limited to collecting entire records of a table.
    - *Preprocessing data :* Once the data is gathered, one may want to preprocess the data to make it suitable for machine learning purposes.
    - For both gathering and preprocessing data, ***quality control*** is an important challenge as well.

13

Data Acquisition——Data Generation

- *Synthetic Data Generation* : low cost and flexibility
  - Generative Adversarial Networks(GANs)
  - Application-specific generation techniques
- ***GANs*** The key approach of a GAN is to train two contesting neural networks: a generative network and a discriminative network. The generative network learns to map from a latent space to a data distribution, and the discriminative network discriminates examples from the true distribution from the candidates produced by the generative network.



14

Data Acquisition——Data Generation

- ***Policies*** Another recent approach is to use human-defined policies to apply transformations to the images as long as they remain realistic.
- ***Data-specific*** .We now introduce data-specific techniques for generation. Synthetic image generation is a heavily-studied topic in the computer vision.

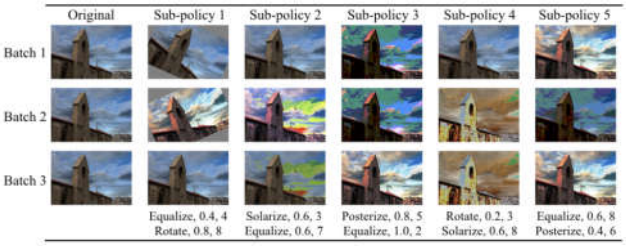
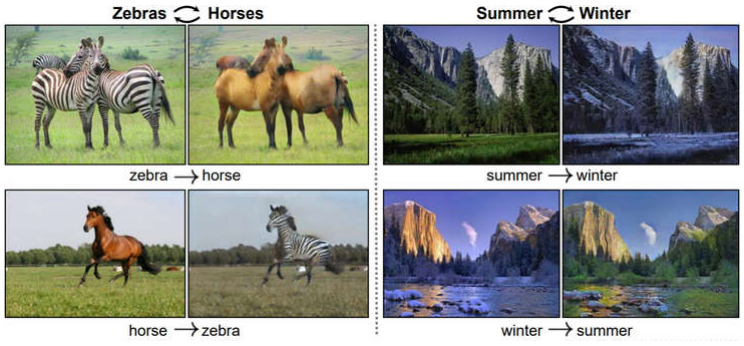


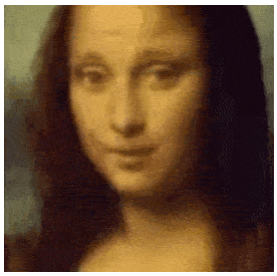
Figure 3. One of the successful policies on ImageNet. As described in the text, most of the policies found on ImageNet used color-based transformations.

Figure 6. Detections on the Amazon domain in Office, showing examples where our synthetic model (second row, green bounding box) improves localization compared to the model trained on real Webcam images (first row, red bounding box).

15

Data Acquisition——Data Generation





16

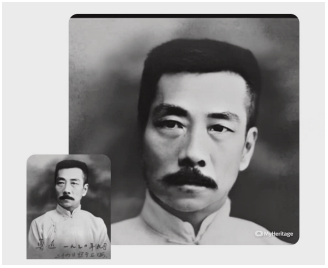
Data Acquisition——Data Generation

- 抖音“吗咿呀嘿”特效图

17

Data Acquisition——Data Generation

- MyHeritage 网站 深度怀旧项目



18

Data Acquisition——Data Generation

- MyHeritage 网站 深度怀旧项目

19

Contents

- Background
- Data Acquisition
- Data Labeling
- Improvement of Existing Data and Models
- How to Decide which Data Collection Techniques to Use When
- Interesting Future Research Challenge

20

Thanks !