Social Data Management
Applications of Social and Graph Data

Silviu Maniu[1]

January 7th, 2019

[1]Université Paris-Sud

Obtaining Data on the Web

Crawling: the operation of obtaining a "picture" of the pages on the Web.

Obtaining Data on the Web

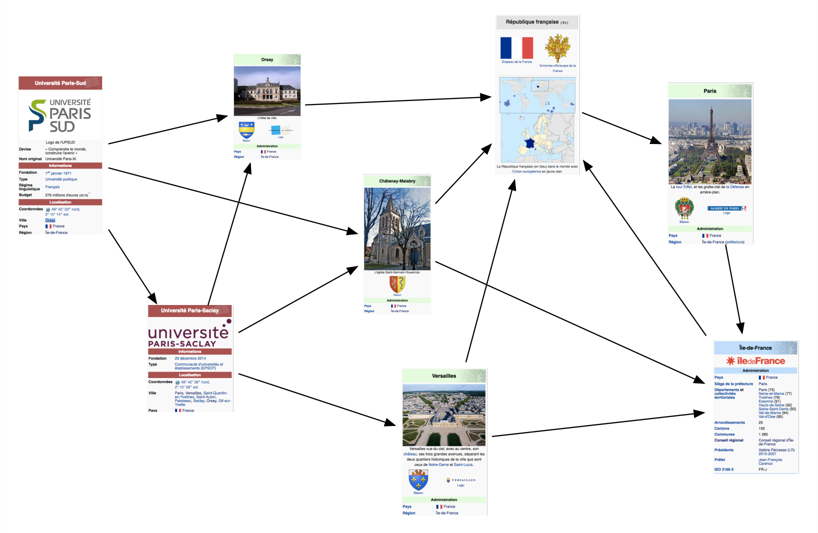
Crawling: the operation of obtaining a "picture" of the pages on the Web.

An iterative process:

1. get a set of pages on the Web called seeds, and process their outgoing links,
2. for each outgoing link, extract it from the Web and process its outgoing links,
3. repeat step 2 until no pages are left.

The set of pages to be processed is called the frontier.

Crawling: Illustration



Focused Crawling

When we have a budget and objective – focused crawling:

- budget – limited Web API calls (Twitter, Foursquare, Facebook), limited money
- objective – crawl only the news related to a subject, obtain the pages that are relevant to a query, etc.

Applications: Web crawling, deep Web mining, social network querying, peer-to-peer gossip.

As opposed to classical crawling (BFS is enough), there must be a way to estimate the worth of each node to be crawled.

Algorithms for Focused Crawling

As opposed to classical crawling (BFS is enough), there must be a way to estimate the worth of each node to be crawled.

Estimation algorithm amount to probabilistic processing: estimating the worth of each node (topic centered PageRank), or probabilistically choosing the best nodes (multi-armed bandits).

Table of contents

Some tasks cannot be performed effectively by computers (*Which?*)

Some tasks cannot be performed effectively by computers (*Which?*)

Crowdsourcing: asking the answers to data from Internet workers, and not from computers

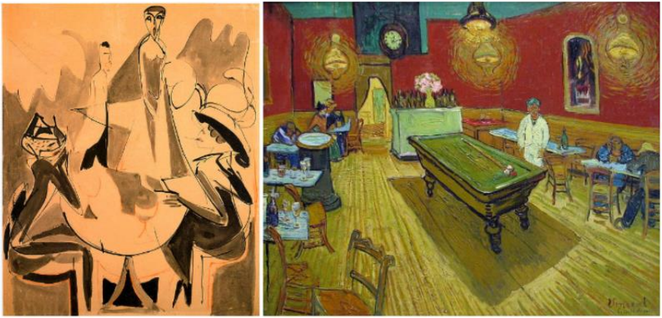
Some tasks cannot be performed effectively by computers (*Which?*)

Crowdsourcing: asking the answers to data from Internet workers, and not from computers

Applications:

- image recognition
- entity resolution
- data cleaning

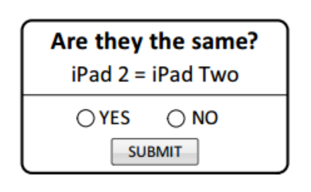
Image Recognition

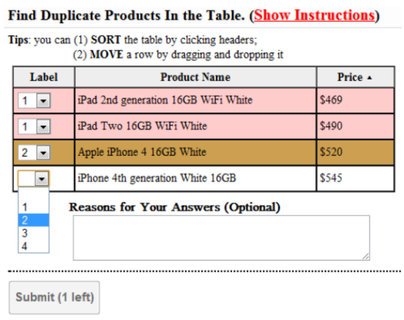


How similar is the artistic style in the paintings above?

- Very similar
- Somewhat similar
- Neither similar nor dissimilar
- Somewhat dissimilar
- Very dissimilar

Entity Resolution



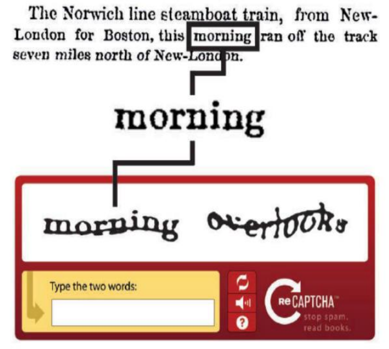


CAPTCHA

CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart

ReCAPTCHA



Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum. ReCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321: 1465-1468, 2008

Crowdsourcing on the Internet

Crowdsourcing Terms

Workers: users, bloggers, Merchanical Turk workers

Requesters: persons who need their data cleaned or need new knowledge

Crowdsourcing Terms

Workers: users, bloggers, Merchanical Turk workers

Requesters: persons who need their data cleaned or need new knowledge

Tasks – also known as HITs (human interface tasks): questions, comments, Wikipedia edits,

Crowdsourcing Terms

Workers: users, bloggers, Merchanical Turk workers

Requesters: persons who need their data cleaned or need new knowledge

Tasks – also known as HITs (human interface tasks): questions, comments, Wikipedia edits,

Incentives: usually money, but can be reputation, recognition in the community

Tasks

Types of tasks:

- binary questions: is Paris the capital of France?
- open questions: what is the address of Télécom?
- comparisons: which image is "better"

Data Issues in Crowdsourcing

Answers from crowds are unreliable, due to the workers' answers

Why?

Data Issues in Crowdsourcing

Answers from crowds are unreliable, due to the workers' answers
Why?

- the workers' answers have to be biased by their reliability (*how to measure?*)
- the data has to be stored and processed in databases (*what kinds of databases?*)

Data Issues in Crowdsourcing

Answers from crowds are unreliable, due to the workers' answers
Why?

- the workers' answers have to be biased by their reliability (*how to measure?*)
- **the data has to be stored and processed in databases** (*what kinds of databases?*)

Qurk

For tasks on Amazon Mechanical Turk, they can be expressed as an workflow:

- SQL queries on the data existing in the database
- UDFs (User Defined Functions) on missing data

Qurk

Qurk

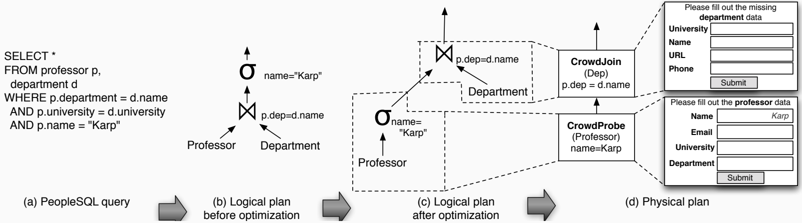
Users give different and conflicting answers – *how can we solve this?*

Qurk

Users give different and conflicting answers – *how can we solve this?*

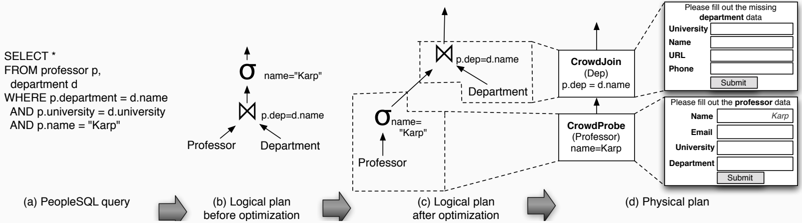
- Qurk uses resolution rules, such as majority voting

CrowdDB



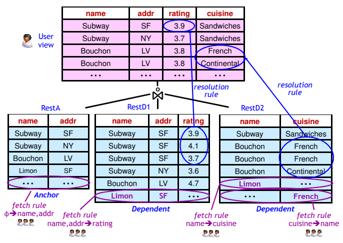
SELECT *
FROM professor p,
department d
WHERE p.department = d.name
AND p.university = d.university
AND p.name = "Karp"
σ name="Karp"
⋈ p.dep=d.name
Professor
Department
⋈ p.dep=d.name
Department
σ name= "Karp"
Professor
CrowdJoin
(Dep)
p.dep = d.name
CrowdProbe
(Professor)
name=Karp
Please fill out the missing department data
University
Name
URL
Phone
Submit
Please fill out the professor data
Name
Karp
Email
University
Department
Submit
(a) PeopleSQL query
(b) Logical plan before optimization
(c) Logical plan after optimization
(d) Physical plan

CrowdDB



- same principle as Qurk, but allows for the generation of new tuples

Deco



- separation between crowd and user views
- defines fetch and resolution rules
- fetch: how data is obtained from the crowd
- resolution: how data is aggregated

Data Issues in Crowdsourcing

Answers from crowds are unreliable, due to the workers' answers

Why?

Data Issues in Crowdsourcing

Answers from crowds are unreliable, due to the workers' answers
Why?

- **the workers' answers have to be biased by their reliability** (*how to measure?*)
- the data has to be stored and processed in databases (*what kinds of databases?*)

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

What is the capital of France?

worker	**answer**
Anne	Paris
Richard	Lyon
Jean	Lyon
Pauline	Paris
Benoit	Paris

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

What is the capital of France?

worker	**answer**
Anne	Paris
Richard	Lyon
Jean	Lyon
Pauline	Paris
Benoit	Paris

Aggregation rules: majority vote, average, . . .

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

What is the capital of France?

worker	answer
Anne	Paris
Richard	Lyon
Jean	Lyon
Pauline	Paris
Benoit	Lyon

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

What is the capital of France?

worker	**answer**
Anne	Paris
Richard	Lyon
Jean	Lyon
Pauline	Paris
Benoit	Lyon

In some cases aggregation rules can fail

Simple Aggregation Rules

Resolution Rules: aggregating the answer from the crowd

What is the capital of France?

worker	answer
Anne	Paris
Richard	Lyon
Jean	Lyon
Pauline	Paris
Benoit	Lyon

Assume that Anne and Pauline give correct answers in 90% of the cases, and Richard, Jean and Benoit only in 50% of the cases – what is the correct answer?

Worker Accuracy

Let us assume labelling questions, where each worker needs to give an answer with only one true value

A simple model: a worker w_i has accuracy π_i – a probability of π_i to give the correct answer and a probability of $1 - \pi_i$ to give the incorrect one

Worker Accuracy

Let us assume labelling questions, where each worker needs to give an answer with only one true value

A simple model: a worker w_i has accuracy π_i – a probability of π_i to give the correct answer and a probability of $1 - \pi_i$ to give the incorrect one

How to get the worker accuracies?:

- estimate their accuracy on a set of control questions
- sometimes, possible to do it without any ground truth input

Example of Crowdsourced Worker Accuracy

worker	Italy	France	U.K.	Spain
Anne	Rome	Paris	London	Madrid
Jean	Milan	Paris	London	Madrid
Pauline	Milan	Lyon	Manchester	Barcelona

Example of Crowdsourced Worker Accuracy

worker	Italy	France	U.K.	Spain
Anne	Rome	Paris	London	Madrid
Jean	Milan	Paris	London	Madrid
Pauline	Milan	Lyon	Manchester	Barcelona

What is the correct answer? – **truth discovery**

Truth Discovery in Crowdsourcing

Assume a set of k facts in $\{0, 1\}$, a set of n workers w_i

Every worker answer for every fact:

$$\boldsymbol{a} = \{a_{11}, \cdots, a_{1n}, \cdots, a_{kn}\}$$

Each worker has an accuracy π_i which is the probability that they answer 1 correctly

We want to derive the labels/answers, $\boldsymbol{l}$

Maximum Likelihood

A standard approach to optimize probabilities – computing the likelihood given the answers:

Maximum Likelihood

A standard approach to optimize probabilities – computing the likelihood given the answers:

$$\mathcal{L}(\boldsymbol{\pi}, \phi \mid \boldsymbol{a}) = \prod_{i}^{n} \prod_{j}^{w} \phi_i^{l_i} (1 - \phi_i)^{1 - l_i} \pi_j^{y_{ij}} (1 - \pi_j)^{1 - y_{ij}}$$

where

$$y_{ij} = a_{ij} l_i + (1 - a_{ij})(1 - l_i)$$

Maximum Likelihood

A standard approach to optimize probabilities – computing the likelihood given the answers:

$$\mathcal{L}(\boldsymbol{\pi}, \phi \mid \boldsymbol{a}) = \prod_{i}^{n} \prod_{j}^{w} \phi_i^{l_i} (1 - \phi_i)^{1 - l_i} \pi_j^{y_{ij}} (1 - \pi_j)^{1 - y_{ij}}$$

where

$$y_{ij} = a_{ij} l_i + (1 - a_{ij})(1 - l_i)$$

We want to estimate $\boldsymbol{\pi}$ and ϕ by maximizing the likelihood

Maximum Likelihood

Maximizing it gives us the following estimates

$$\hat{\phi}_i = \frac{\sum_j^n a_{ij}\pi_j + \sum_j^n (1 - a_{ij})(1 - \pi_j)}{n}$$

$$\hat{\pi}_i = \frac{\sum_i^k a_{ij}\phi_i + \sum_i^k (1 - a_{ij})(1 - \phi_j)}{k}$$

Maximum Likelihood Estimation (MLE)

The estimations are recursively defined – to maximize it, we can use the EM algorithm:

1. initialize the facts and the worker accuracies (assume workers are 100% accurate)
2. estimation (E-step) estimate the labels l_i based on the probabilties $\hat{\phi}_i$
3. maximization (M-step) compute the worker and fact probabilities based on the labels
4. iterate 2 and 3 until convergence

Example of Crowdsourced Worker Accuracy

worker	**Italy**	**France**	**U.K.**	**Spain**
Anne	Rome	Paris	London	Madrid
Jean	Milan	Paris	London	Madrid
Pauline	Milan	Lyon	Manchester	Barcelona

Exercise: What is the correct answer?

Using BID Databases

country	capital	answers
France	Paris	7
France	Lyon	3
Italy	Rome	5

0.7

country	capital
France	Paris
Italy	Rome

0.3

country	capital
France	Lyon
Italy	Rome

Using BID Databases

country	capital	prob
France	Paris	0.7
France	Lyon	0.3
Italy	Rome	1

0.7

country	capital
France	Paris
Italy	Rome

0.3

country	capital
France	Lyon
Italy	Rome

Add a `REPAIR-KEY` construct to SQL to transform raw answers to probabilistic databases

Using BID Databases

To answer queries like *What is the correct capital of country X?* we can add a `WHILE` operator / fixpoint operator

Using BID Databases

To answer queries like *What is the correct capital of country X?* we can add a `WHILE` operator / fixpoint operator

- this will result in a Markov chain of instances, for which we need to compute the stationary distribution for a class of queries
- this is a known #P-hard problem

Using BID Databases

To answer queries like *What is the correct capital of country X?* we can add a `WHILE` operator / fixpoint operator

- this will result in a Markov chain of instances, for which we need to compute the stationary distribution for a class of queries
- this is a known #P-hard problem

Approximation:

- additive approximation is PTIME
- multiplicative approximation is NP-hard

Acknowledgments

Figures in the crowdsourcing section are taken from the following references.

References i

Dawid, A. P. and Skene, A. M. (1979).
Maximum likelihood estimation of observer error-rates using the em algorithm.
Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1).

Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011).
Crowdsourcing systems on the world-wide web.
Commun. ACM, 54(4).

Li, G., Wang, J., Zheng, Y., and Franklin, M. J. (2016).
Crowdsourced data management: A survey.
IEEE Transactions on Knowledge and Data Engineering, 28(9).

Liu, Q., Steyvers, M., and Ihler, A. (2013).
Scoring workers in crowdsourcing: How many control questions are enough?
In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13.