Commonsense Properties from Query Logs and Question Answering Forums

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, Gerhard Weikum

Goal

- Mine Commonsense Knowledge (CSK) about :
  - Object properties
  - Human behavior
  - General concepts
- Focus on salient properties
- Examples :
  - (bananas, are, edible)
  - (children, like, bananas)
- Applications : Chatbot, Question Answering, Visual content understanding, Search engine queries interpretation, ...

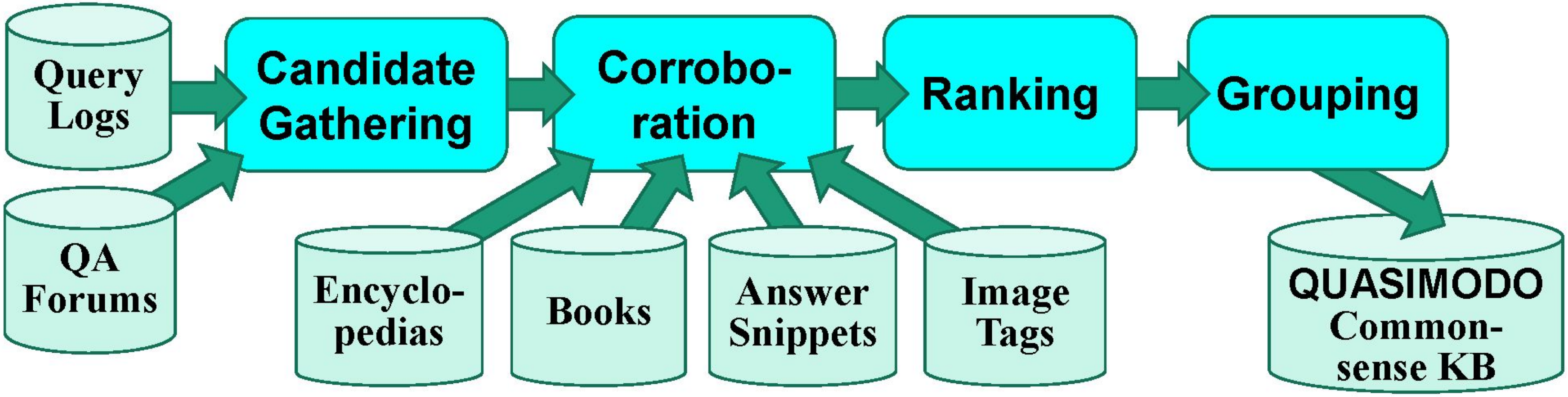
Challenges

- Sparseness and bias
- Rarely expressed
- Non-encyclopedic (no Wikipedia)
- Noise and high bias on online content

Previous Work

- Traditional Knowledge Bases
  - No commonsense
- ConceptNet
  - Manual, does not scale
- Webchild
  - Focus on possible properties, not salient ones
- TupleKB
  - Domain specific

General Pipeline



Candidate Gathering

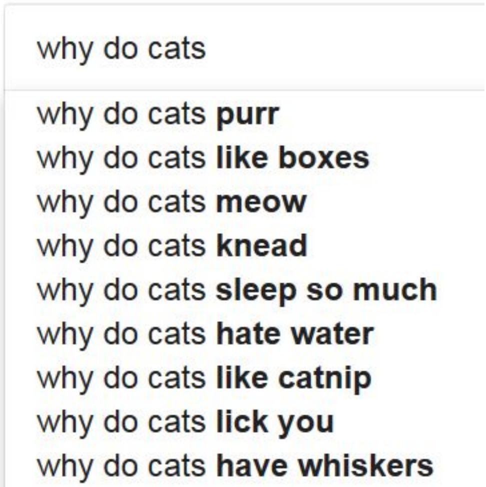
- Main idea : Extract facts from questions
  - When asking a question, make assumptions about the world

Why are bananas yellow? → Bananas are yellow!

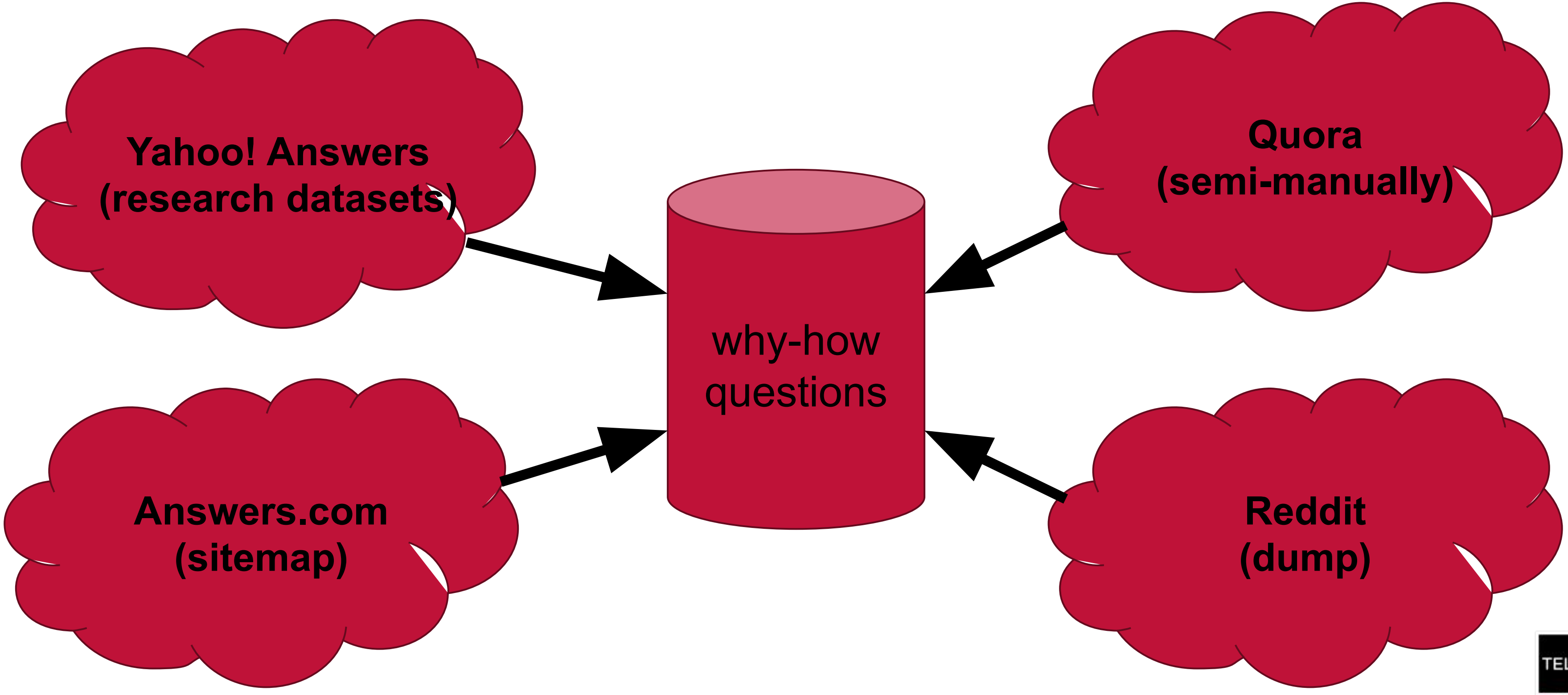
  - Harvest human curiosity, « wisdom of the crowds »

Candidate Gathering – Query Logs

- **Indirect access to the query logs through autocompletion**



Candidate Gathering – QA Forums



Candidate Gathering – Statistics

Pattern	In Query Logs	In QA Forums
how does	19.4%	7.5%
why is	15.8%	10.4%
how do	14.9%	38.07%
why do	10.6%	9.21%
how is	10.1 %	4.31%
why does	8.97%	5.46%
why are	8.68%	5.12%
how are	5.51%	1.8%
how can	3.53%	10.95%
why can’t	1.77%	1.40%
why can	0.81%	0.36%

Candidate Gathering – Results

- Questions transformed to statements then to triples using OpenIE techniques

Why do lions often hunt zebras? → **Q2S** → Lions often hunt zebras

OpenIE → (lions, often eat, zebras)

Modality → (lions, eat, zebras, often)

Positivity → (lions, eat, zebras, often, positive)

Source → (lions, eat, zebras, often, positive, Google, 0.4)

Corroboration

- Reduce noise thanks to additional signals from :
  - Wikipedia and Simple Wikipedia
  - Answer snippets from search engines
  - Google Books
  - Image Tags from OpenImages and Flickr
  - Google's Conceptual Captions dataset
- Train Naive Bayes from all signals from 700 manually annotated triples (TuplesKB requires 70.000)
  - Precision of 61%

Ranking + TODO Example

- From Corroboration, get plausibility score π
- Define a probability from it:

$$P[s, p, o] = \frac{\pi(spo)}{\Sigma_{x \in KB}\, \pi(x)}$$

- Derive a typicality τ and a saliency σ:

$$\tau(\mathrm{s, p, o}) = P[p, o \mid s] = \frac{P[s, p, o]}{P[s]}$$

$$\sigma(\mathrm{s, p, o}) = P[s \mid p, o] = \frac{P[s, p, o]}{P[p, o]}$$

Grouping

- **Reduce redundancy**
- **Clustering method based on tri-factorization**
- **Groups of (Subject, Object) and Predicate**

P clusters	SO clusters
make noise at, be loud at, make noises at, croak in, croak at, quack at	fox-night, frog-night, rat-night, mouse-night, swan-night, goose-night, chicken-night, sheep-night, donkey-night, duck-night, crow-night
misbehave in, talk in, sleep in, be bored in, act out in, be prepared for, be quiet in, skip, speak in	student-class, student-classes, student-lectures
diagnose, check for	doctor-leukemia, doctor-reflexes, doctor-asthma, doctor-diabetes, doctor-pain, doctor-adhd

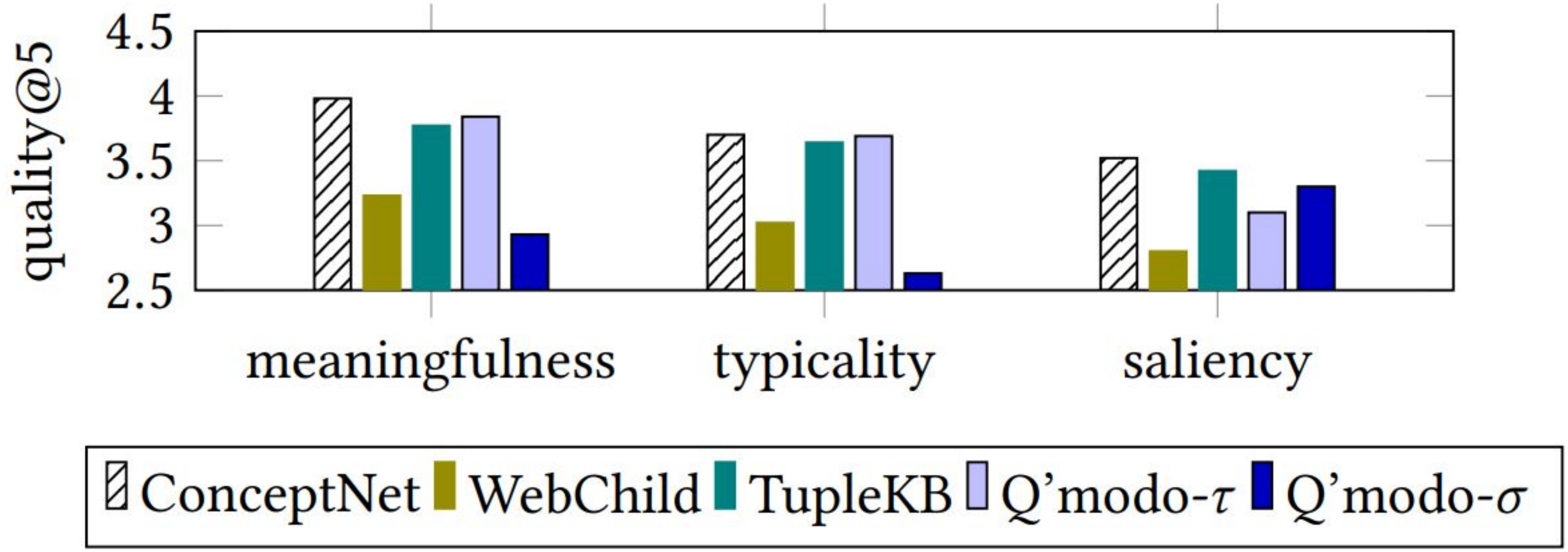
Statistics

	Full KB					animals		occupations	
	#S	#P	#P≥10	#SPO	#SPO/S	#S	#SPO	#S	#SPO
ConceptNet-full@en	842,532	39	39	1,334,425	1.6	50	2,678	50	1,906
ConceptNet-CSK@en	41,331	19	19	214,606	5.2	50	1,841	50	1,495
TupleKB	28,078	1,605	1,009	282,594	10.1	49	16,052	38	5,321
WebChild	55,036	20	20	13,323,132	242.1	50	27,223	50	26,257
Quasimodo	80,145	78,636	6084	2,262,109	28.2	50	39,710	50	18,212

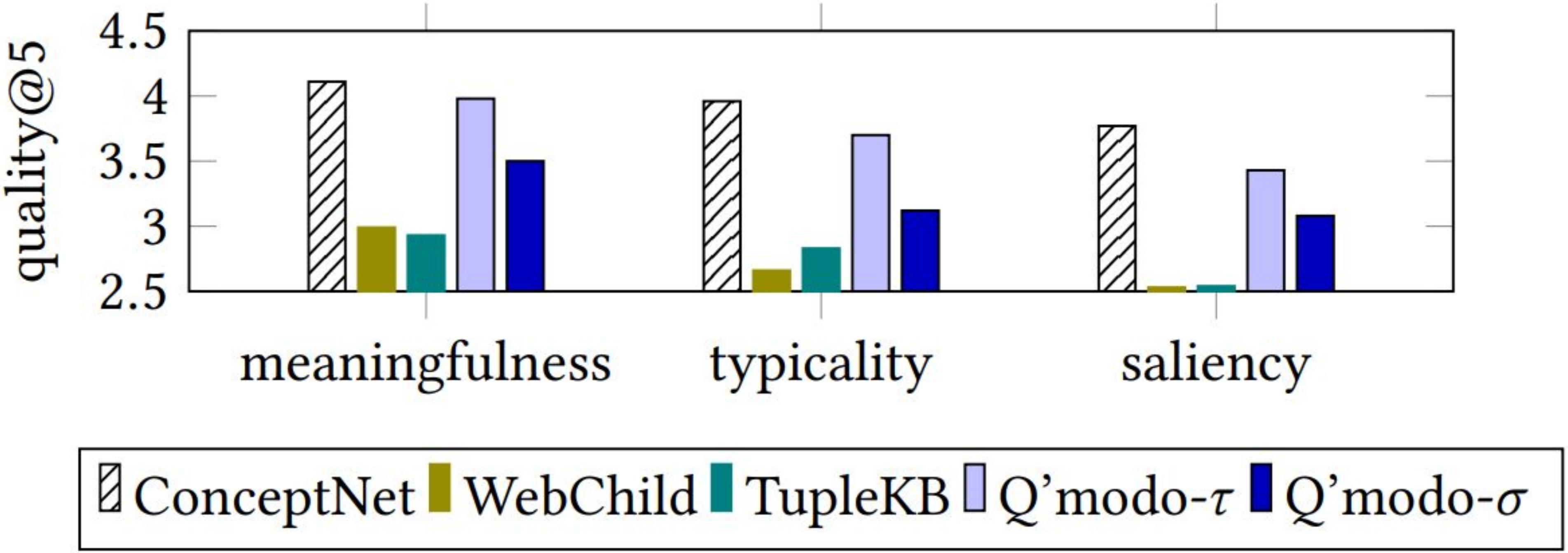
Examples of facts

- Practical knowledge from human, e.g. : **(car, slip on, ice)**
- Problems linked to a subject, e.g.: **(pen, can, leak)**
- Emotions linked to events. e.g.: **(divorce, can, hurt)**
- Human behaviors. e.g.: **(ghost, scare, people)**
- Negative knowledge, e.g.: **Not (elephant, can, jump),**
- Salient modalities, e.g.: **Always (doctor, have, unreadable handwriting)**
- Trivial facts, e.g.: **(road, has_color, black)**
- Newest facts. e.g.: **(trump, build, wall)**
- Cultural knowledge (here U.S.) e.g.: **Always (school, have, locker)**
- Comparative knowledge, e.g.: **(light, faster than, sound)**

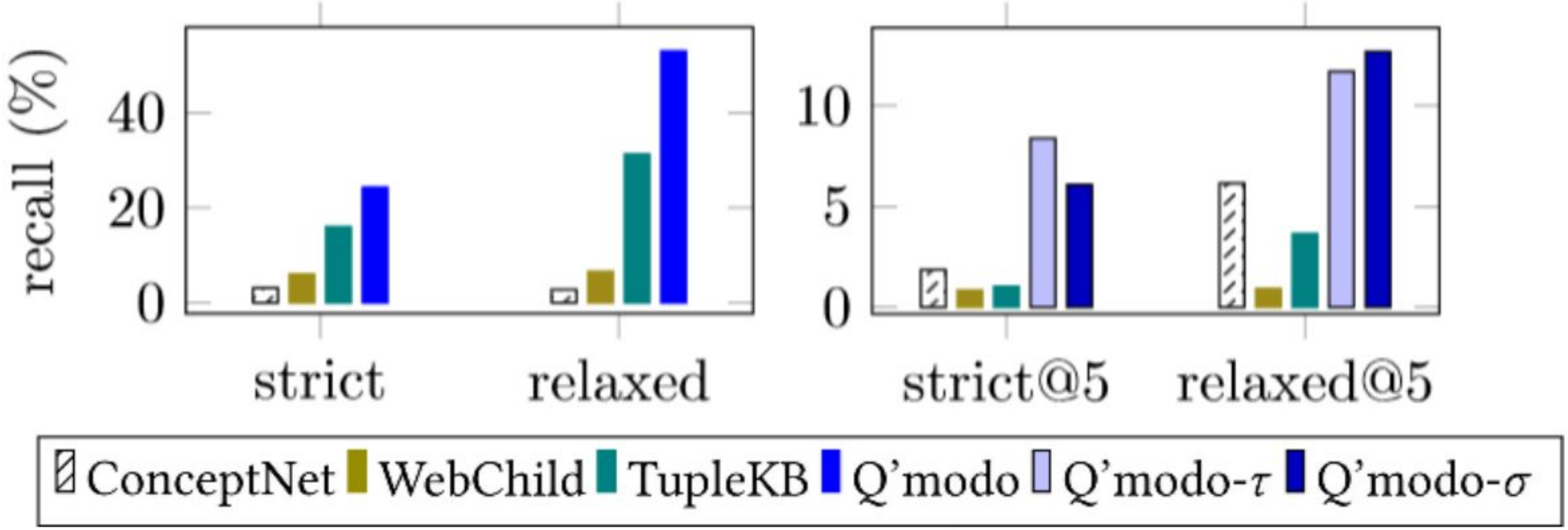
Precision – Entire CSKs



Precision – Same Subjects



Recall


recall (%)
40
20
0
strict
relaxed
10
5
0
strict@5
relaxed@5
ConceptNet
WebChild
TupleKB
Q'modo
Q'modo-τ
Q'modo-σ

Question Answering

KB	All
#Questions (Train/Test)	10974/3659
Random	22.0
word2vec	27.2
Quasimodo	**31.3**
ConceptNet	27.5
TupleKB	27.5
WebChild	24.1

TELECOM Paris IP PARIS

Conclusion

- We introduced a new methodology for acquiring CSK from non-standard sources
- Improve state of the art with better coverage of typical and salient properties, determined by Mturks
- Extrinsic evaluations illustrate advantages
- Data and code available: github.com/Aunsiels/CSK

Additional slides

Future Work

- Cultural knowledge
- Study of stereotypes
- Temporal evolution of the knowledge base
- Improve ranking methods
- Scale to the entire web

Litterature

- **Data: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/commonsense/quasimodo/**
- **Code: https://github.com/Aunsiels/CSK**
- **http://conceptnet.io/**
- **http://data.allenai.org/tuple-kb/**
- **https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/commonsense/webchild/**