Mining Subjective Properties on the Web

Immanuel Trummer, Alon Halevy, Hongrae Lee, Sunita Sarawagi, Rahul Gupta
immanuel.trummer@epfl.ch, {halevy, hrlee, sarawagi, grahul}@google.com

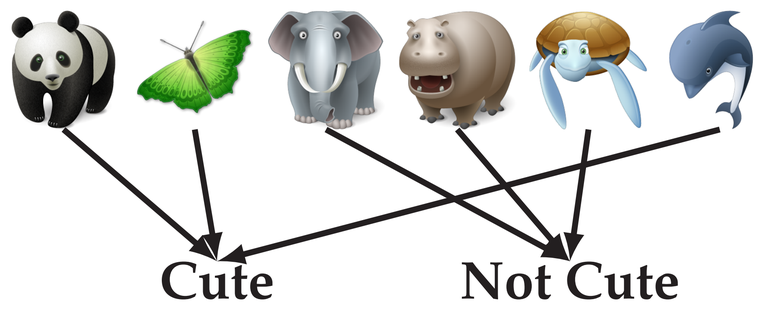
Context

Google receives queries for CUTE ANIMALS and BIG CITIES. To answer such queries from structured knowledge bases, we must know which subjective properties (e.g., CUTE or BIG) are commonly associated with which entities (e.g., specific animals or cities).

Goal

Given an entity and a subjective property, find out whether the majority of Web users associate the property with the entity.

Example Problem: Which animals are cute?



Subjective vs. Objective

Objective Properties are non-controversial. There exists a ground truth (e.g., whether a city is AMERICAN) that we can extract from a few Web sites.

Subjective Properties are controversial. There is no ground truth (e.g., whether a city is BIG) but often a dominant opinion. We can find the dominant opinion by extracting opinions from many Web sites and weighting conflicting statements against each other.

Challenges

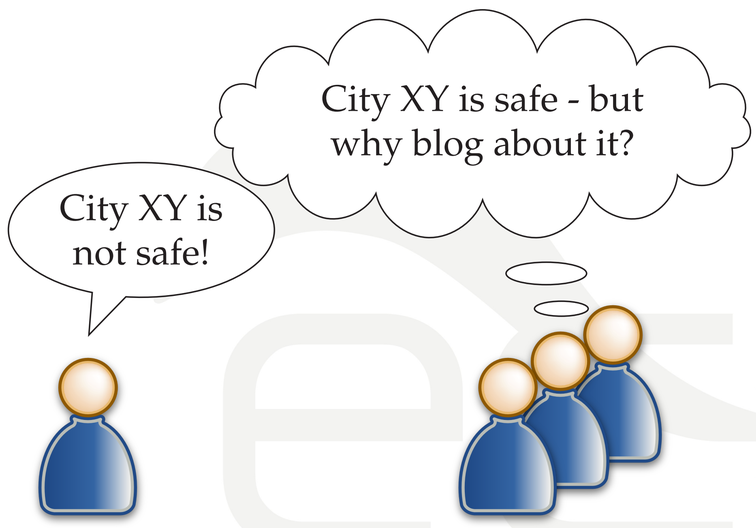
The following approach might seem natural:

Naive Majority Vote Approach. We count all statements claiming that a property applies to an entity and all statements claiming the opposite. We assume that the majority vote reflects the dominant opinion.

It works poorly due to two effects:

Sparseness. Many knowledge base entities are never mentioned on the Web. This can provide useful information: an unmentioned city is probably not big. But such correlations are entity type and property specific.

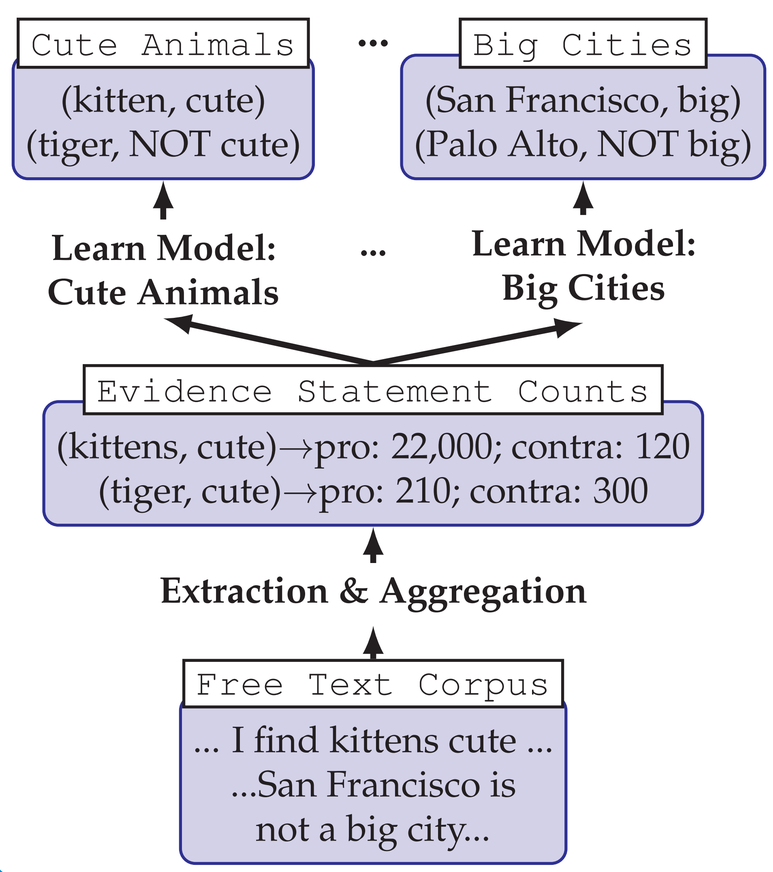
Skew. Users who think for instance that a specific city is not safe might be more likely to blog their opinion than users who perceive the same city as safe. Hence users with one specific opinion might be over-represented on the Web.



The Surveyor System

We built the SURVEYOR system which extracts statements about entities and properties from the Web to associate entities to properties. SURVEYOR parsed an entire Web snapshot resulting in 4 billion entity-property associations.

We count statements that associate a property to an entity and compare to the number of statements doing the opposite. We learn a statistical model of how users generate content on the Web that is specific to an entity type and a property. We interpret statement counts for specific entities and properties based on that model.

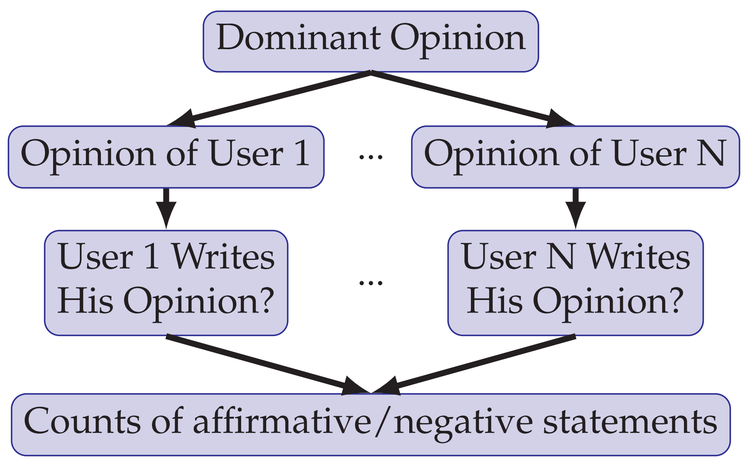


Extraction

We transform free text sentences into a dependency tree representation and identify mentions of knowledge base entities. We detect dependency tree patterns that indicate a statement claiming that an entity does or does not have a specific property.

User Model

Our goal is to infer the dominant opinion from the number of affirmative and negative statements counted on the Web. We model the relationship between statement counts and dominant opinion as Bayesian network:

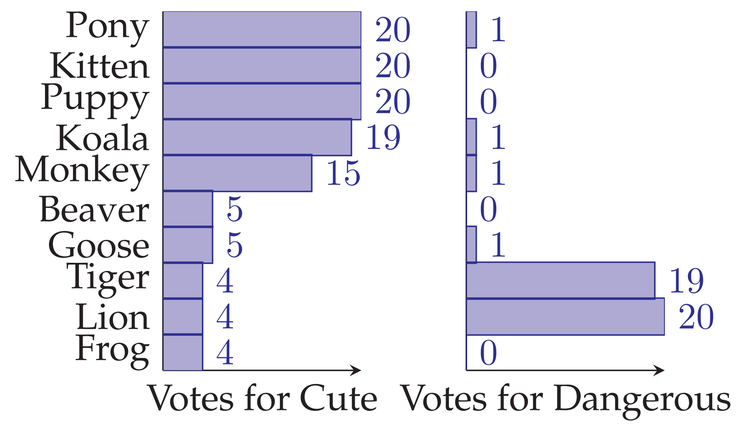


Our model has parameters that describe how user opinion depends on the dominant opinion and how the probability that a user expresses himself on the Web depends on his opinion. We learn the best parameter values for each type-property combination by an unsupervised expectation-maximization approach. Thereby we overcome sparseness and skew.

Experimental Setup

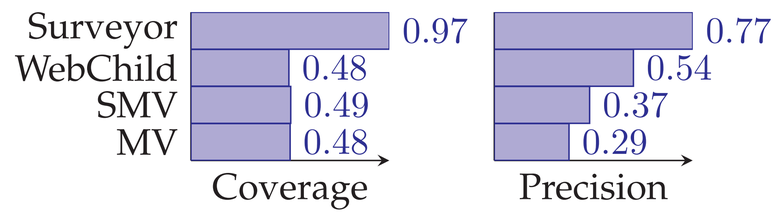
Coverage is the percentage of entities for which SURVEYOR infers a dominant opinion with high confidence. **Precision** is the percentage of entities for which SURVEYOR's predictions match the majority opinions of Amazon Mechanical Turk (AMT) workers.

Example. We asked 20 AMT workers: Is the animal cute? Is it dangerous? SURVEYOR predicts the majority opinions.

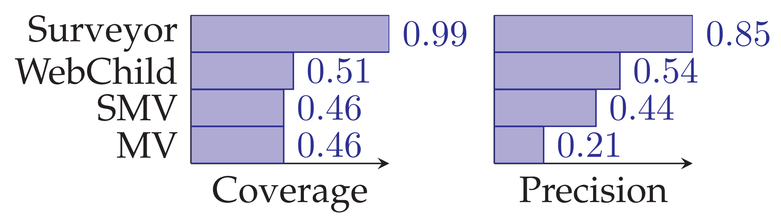


Experimental Results

We compared SURVEYOR against two naive baselines and the WEBCHILD system:



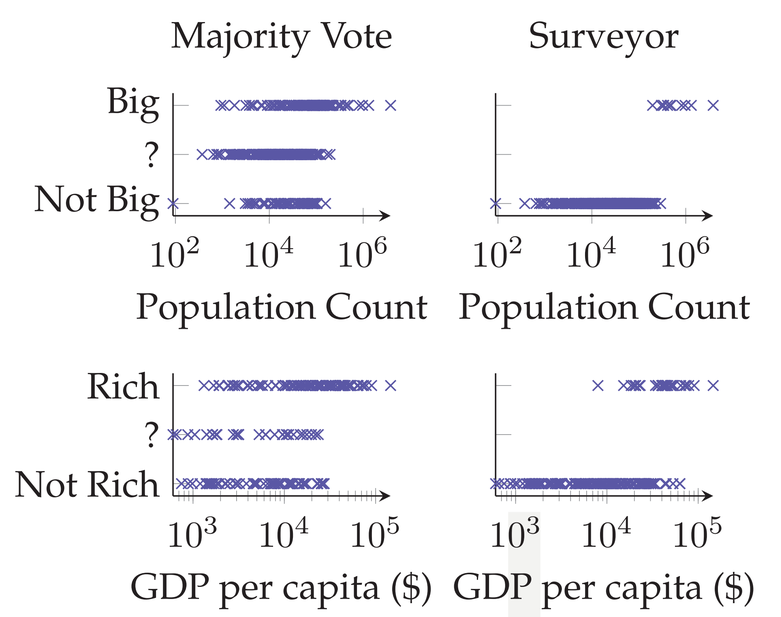
The performance of SURVEYOR improves for test cases with high AMT worker agreement:



SUVEYOR outperforms WEBCHILD since only SURVEYOR is specialized to subjective properties. The two baselines do not use entity type and property specific user models and hence suffer from poor precision and recall.

We study how well the output of SURVEYOR correlates with objective properties. We compare against a simple majority vote approach.

Unlike majority vote, SURVEYOR categorizes cities with many inhabitants as big and countries with high GDP per capita as rich:



Conclusion

Mining subjective properties from the Web is possible but requires specialized systems.