Google

The Data Divide

Luke Segars - Google
11/7/2012 @ CS10

Statement
Google's advantage is not in writing drastically better software; it's in having more data.

Question
Can any problem be solved by computers if enough data is available?

One major change that's come about from the digital revolution is the fact that *MUCH more data is available* for consumption now than ever before.

Internet
~4.5M URLs / month

Twitter
~5.5B tweets / month

Blogs
lots / month

Google Maps
>= 5M miles of road

Project Gutenberg
40,000 free books

And that's just the visible stuff...

iClicker Question

Approximately how many web pages did Google have in its search index earlier this year?

a. 200 million
b. 1 billion
c. 12 billion
d. 45 billion
e. 100 billion

That's a lot to search through, but it's also a lot to learn from.

All of the text, images, video and other media generated on the Internet aren't entirely independent, and *finding regular occurrences generally implies a correlation between concepts*.

Let's consider an example:

*Say you have **10,000** news articles from diverse sources about Hurricane Sandy.*
***7,000** of them contain the phrase "New York."*
***3** of them contain the phrase "Arizona."*

Assuming that your news sources are actually telling a story, it is reasonable to assume that Hurricane Sandy is more closely related to New York than Arizona.

The core idea is based in statistics:

1. Many people, places, things, and ideas are somehow related to each other.
2. Ideas that are more closely related to each other are more likely to co-occur.
3. Co-occurring once means nothing, but co-occurring millions of times suggests that the two ideas are related.

Let's look at three places where Google uses this principle to make great things.

PageRank
How do we objectively measure a site's reputation?

Spell Checking
How can we build a system that automatically learns new words in any language?

Image Composition
How do we know what the subject of a picture is?

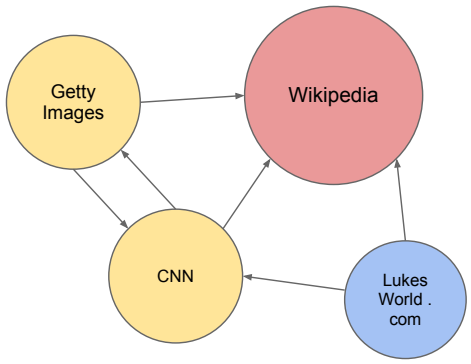
Example #1: Pagerank

The Internet is full of data sources, some more reputable and dependable than others.

Article on CNN > my blog post

Pagerank is an algorithm that estimates the reputation of a website by looking at the reputation of websites that provide links to it.

Example #1: Pagerank



Example #2: Spell checker

Splel checking is a fairly well-understood problem if you've got a reasonably static vocabulary. The Internet does not have a static vocabulary; new phrases are emerging all the time in all different languages.

How can Google keep its spell checker up to date in every language without a ton of work?

Example #2: Spell checker

Statistics!

Key observations:

1. A particular misspelling will likely be uncommon across the web, especially on reputable sites.
2. The context that a misspelling occurs in will be similar to the context of a similar but more common spelling.

Example #3: Image composition

Identifying what's going on in a picture is really hard to do algorithmically.

Google has a number of techniques for helping with this, one of which is using context clues that occur around the image in a document.

Captions, anchor text, titles, nearby paragraphs.

Example #3: Image composition

Biden casts vote, says 'it's always a kick'

By MATTHEW DALY, Associated Press – 23 minutes ago

GREENVILLE, Del. (AP) — Vice President Joe Biden has cast his vote in the 2012 election, saying "it's always a kick."

President Barack Obama's running mate says it was the eighth time he's run for election statewide in Delaware. Asked if he thought it was the last time he'd vote for himself, he told reporters: "No, I don't think so."

Biden arrived with his wife, Jill, at Alexis I. DuPont High School in Greenville, Del., shortly after the polls opened at 7 a.m. Tuesday and waited about 13 minutes.

Several voters offered to let Biden move ahead of them, but he said no. "I've never butted in line in Delaware. The idea of doing it on Election Day — whoa!" Biden said.

Biden shook hands with other voters, hugged several and posed for pictures as he waited. He called voting a great honor and urged Americans to "stand in line as long as you have to" in order to vote.

Also accompanying the vice president were his son, Delaware Attorney General Beau Biden; Beau Biden's wife, Hallie; and their 8-year-old daughter, Natalie.

AP ASSOCIATED PRESS

Photo 1 of 2

Vice President Joe Biden accompanied by his wife Jill Biden and son Beau Biden waves to members of the media after casting his ballot at Alexis I. duPont High School, Tuesday, Nov. 6, 2012, in Greenville, Del. (AP Photo/Matt Rourke)

There are many other significant applications of this principle, even without looking outside of Google.

1. Identifying synonyms and acronyms
2. Text translation
3. Speech recognition
4. Major event detection
5. Email spam detection
6. Reading speed limit signs in autonomous vehicles

Why do we have so many unsolved problems if all it takes is lots of data?

...either because you can't actually solve all problems with lots of data or because we don't have the right data.

Summary: I don't know if I buy that *all* problems can be solved with data, but our recent progress suggests that we should never say such things are impossible.

Considering the vastness of the Internet and the billions of people who still aren't connected, we've still got a lot of learning left to do.