Quality to Quantity to Quality on the Web

Andraž Tori, CTO at Zemanta
@andraz
Topics

- a bit about Zemanta
- how advanced “data tools” and spammers interact
We are all trying to organize the web
Making it right,
making it useful
and linked
Not so long time ago, in a city not far away...
some other people
are trying to do the opposite
trying to disorganize it,
make it confusing,
and to profit from that
using the tools we have built!
Their motives are not sinister (mostly)
it is about profit
Profit

- publish as much content as possible
- quality is not (that) important
- get traffic or high page ranking for certain terms
- sell clicks, links or whole “fully built” sites to the highest bidder
- users and search engines are necessary evil to be tricked as cheaply as possible
So, why do I care?
Job opening

You will get a spreadsheet with 180 blog url’s and logins. You will log into each blog and schedule 2 posts per week ...

You will spice up every post with images and/or related links within the content, using a Wordpress plugin called Zemanta

https://www.odesk.com/jobs/Wordpress-Blog-Poster_~~c8c04549b8e6b600
And why might you care?

- the organized information is great tool for those that try to disorganize it
- they are poisoning “our web”
- using resources & services you help provide
- and it is just matter of time when they start poisoning linked data too
What do we do at Zemanta
Zemanta is a “personal writing assistant” suggesting content while you write (your blog) analyzing your text connecting it with background knowledge, other stories on the web, images you choose what suggestions to include to make your writing more informative, vivid and useful
Harry Potter book appeared, in 1997, it was just a year before the universal search engine Google was launched. And so Hermione Granger, that charming grind, still goes to the Hogwarts library and spends hours and hours working her way through the stacks, finding out what a basilisk is or how to make a love potion. The idea that a wizard in training might have, instead, a magic pad where she could inscribe a name and in half a second have an avalanche of news stories, scholarly articles, books, and images (including images she
Path: p

IN-TEXT LINKS

Harry Potter Granger

Harry Potter movies

Harry Potter

Hogwarts

Quidditch

Google

universal search

Word count: 148
Harry Potter book appeared, in 1997, it was just a year before the universal search engine Google was launched. And so Hermione Granger, that charming grind, still goes to the Hogwarts library and spends hours and hours working her way through the stacks, finding out what a basilisk is or how to make a love potion. The idea that a wizard in training might have, instead, a magic pad where she could inscribe a name and in half a second have an avalanche of news stories, scholarly articles, books, and images (including images she
Harry Potter book appeared, in 1997, it was just a year before the universal search engine Google was launched. And so Hermione Granger, that charming grind...
Opening up the hood
the reality
How it works

Plain text (article)

Analysis

Semantic search

Content suggestions

Linked data

RSS feeds
Main design goals

- Input is meaningful chunk of text (not a keyword or a phrase)
- Input is English language
- Has to work across all domains in the open world
  - music, celebrities, finance, entertainment, politics…
  - even when talking about semantic web
Analysis pipeline

Named Entity Extraction

Known phrases extraction (aho-corasick)

Surface form features evaluation

Statistical comparison to background knowledge

Semantic coherence and hand-tuned heuristics

Disambiguated entities

Triple store

Amazon.com

Wikipedia

Freebase

CrunchBase

e etc.
Analysis pipeline

- **Named Entity Extraction**
  - Known phrases extraction (aho-corasick)
  - **Surface form features evaluation**
    - Statistical comparison to background knowledge
    - Semantic coherence and hand-tuned heuristics

**Categorization to Dmoz**

- **Categories**
- **Ambiguous named entities**
- **Disambiguated entities**

**Triple store**

- Amazon.com
- Wikipedia
- Freebase
- CrunchBase
- etc.
Background knowledge

- Data from Wikipedia, MusicBrainz, Freebase… and world wild web

- Includes linguistical and semantical properties
  + unstructured data

- Present in two forms:
  - in “original” custom built triple store on top of MySQL (150 GB)
  - processed into 7 GB optimized “memory mapped dump”
Background knowledge

- 7M mined and linked up entities and concepts
- 30M aliases
- Refreshed about once a month
  - want to make it real-time
- Input data quality is really important
After analysis

Text → SOLR articles → Related articles

Text → SOLR images → Images

Related articles

Images
Solr

- We adapted Solr for “query by document”
- 52% precision (at 10) on internal evaluations
  - plain Lucene MLT comes to 44%
  - difference is from “bag of terms” approach over “bag of words” (terms coming from analysis step)
- Our live index is 5M articles
- Solr is really not optimized to handle 50 terms in a single query
Lucene plain “More Like This”

- MLT (maxqt=5)
- MLT (maxqt=10)
- MLT (maxqt=15)
- MLT (maxqt=22)
- MLT (maxqt=30)
Every part of the system is being constantly evaluated

Precision/recall at 5 different points in the system

Mostly bi-weekly releases of new datasets and the engine
Overview

- We do pretty deep processing to deliver simple user experience of “personal authoring assistant”

- And everything is available over the web API
  - tagging
  - named entity recognition and disambiguation to Linked Open Data URIs
So mash-ups happen...
We are just one of the many people offering services based on large amounts of web data, each spending man-years trying to organize their data, trying to offer best possible service.
now back to the bad guys
Job opening

You will get a spreadsheet with 180 blog url’s and logins. You will log into each blog and schedule 2 posts per week ...

You will spice up every post with images and/or related links within the content, using a Wordpress plugin called Zemanta

https://www.odesk.com/jobs/Wordpress-Blog-Poster_~~c8c04549b8e6b600
There's more than meets the eye
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Use Zemanta or OpenCalais to add tags, images, links

Pull additional content from Freebase

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Pull additional content from Freebase

Cover your tracks

Use Zemanta or OpenCalais to add tags, images, links

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?
Finding their keywords, niches

- Domain expertise
- Users like to install extensions and say “yes”
- You observe referrers on sites you control
- You buy the data on the black market
Firefox add-on with 7m downloads can invade privacy
Ant Video secretly tracks every website visited

By Dan Goodin • Get more from this author

Posted in ID, 20th May 2011 04:00 GMT

A high-rated Firefox extension with more than 7 million downloads secretly collects data about every website the open-source browser visits and combines it with uniquely traceable information tied to the user, an independent security researcher said.

The undisclosed behavior of the Ant Video Downloader and Player add-on takes place even when the Firefox private browsing mode is turned on or when users are availing themselves of anonymity services such as Tor. The add-on carries a rating of four out of five possible stars and gets an average of almost 7,000 downloads per day, according to official Mozilla statistics.

The revelations raise new questions about the safety of extensions offered on Mozilla's website. A spokeswoman for the open-source developer said the media player, like all public
The sophisticated part of the market

“Demand Media relies on a proprietary algorithm to help editors best determine what subjects their writers should tackle.”

Factors:

- Keyword competition
- Revenue
- Driving traffic to/from existing content

http://emediavitals.com/article/16/demand-media-s-content-assembly-line
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Use Zemanta or OpenCalais to add tags, images, links

Pull additional content from Freebase

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?

Publish
Find / create content

- Steal
- Take from “open article directories”
- Have your own “content assembly line” like Demand Media
Submit your **high-quality, original** articles for more **exposure, credibility and traffic** back to your website.

**Article Categories**

- Arts and Entertainment
- Business
- Computers and Technology
- Gaming
- Home Based Business
- Internet and Businesses Online
- Legal
- Real Estate
- Relationships
- Travel and Leisure
- Automotive
- Cancer
- Finance
- Health and Fitness
- Home Improvement
- Investing
- News and Society
- Recreation and Sports
- Self Improvement
- Womens Interests
- Book Reviews
- Communications
- Food and Drink
- Home and Family
- Insurance
- Kids and Teens
- Pets
- Reference and Education
- Shopping and Product Reviews
- Writing and Speaking

**Recent Articles**

**Research And Evaluation Studies That Work: Why Evaluate?**

Evaluations are essential elements in a company's or organization's planning and development cycles. This article...
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Pull additional content from Freebase

Use Zemanta or OpenCalais to add tags, images, links

Amazon Mechanical Turk to post comments and links back to your site

Profit?

Use Zemanta to find similar blogs

Publish
This is not the text you are looking for.
This is not the text you are looking for.
Translate it to random language and back to English

Übersetzen sie zufällig Sprache und wieder auf Englisch

Language and translate it happen again in English

Μεταφράστε αυτό σε δειγματοληπτικούς γλώσσα και πίσω στην αγγλική γλώσσα

Translate this random language back to English

Traduisez au langage aléatoire et revenir à l'anglais

Translate to random language to English and back

它翻译成随机的语言和回英文

Translate it back into the English language and random
Covering their tracks

- Trying to fool search engines or people?
- Search engines are catching up
- Google Translate API is being closed due to “abuse”?
- The trend is “rewriting” by human editors, procured on the global market
Blog Articles

I need 25 blog articles. Topic would be about radiant floor heating. I will be sending you the articles to rewrite, must pass copyscape or you will not be paid. It would just be for blog posts so the article should just be 300 to 350 words. The work will be FREE OF ERRORS, NO GRAMMAR OR PUNCTUATION MISTAKES. The work should be in perfect English. No spun articles. Any work that is poorly done, has erro ...

Looking for long Term Article writers (Start Immediately)

I have 7000 products that I need the 100 words descriptions to be written for. Most of the products belong to Electronics & Furniture category. Initially, around 20 articles/descriptions will be required on a daily basis. As we move on, the numbers may go up to 50 articles per day. You must be able to write perfectly in UK/Indian English and pay attention to grammar, spelling, Typo and punctuation. You should have decent re ...

ARTICLE RE-WRITER

I am looking for someone to simply RE-WRITE existing articles. Hello (POTENTIAL EMPLOYEE), Thanks for the Job inquire. What I am seeking is for someone to RE-WRITE articles for me based on a theme I provide. Currently, I would like the articles seen here: http://EzineArticles.com/?expert=Fazila_Adam Re-Written. There are hundreds of Articles, but I ask that you start at the TOP of the list and work your way down. I would ...

Skills: Article Writing, *Being able to write in English a Must*
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Use Zemanta, OpenCalais to add tags, images, links

Cover your tracks

Pull additional content from Freebase

Amazon Mechanical Turk to post comments and links back to your site

Profit?

Use Zemanta to find similar blogs

Publish
Spammers say darndest things

“Thanks guys, I scrape content most of the time and then I build (sites) using WordPress. tagpig is the perfect plugin for this kind of work.”

Fernan
“TagPiG is a serious time saver.”

Joe B

What Do You Get?

Automated Content Analysis
Leverage 5 powerful content analysis engines - Yahoo, Tagthe.net, Open Calais, Alchemy and Zemanta to extract the key topics and phrases from your posts.
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Pull additional content from Freebase

Use Zemanta or OpenCalais to add tags, images, links

Publish

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?
As traditional media companies are trying to grasp the power of online communities and user-generated content, Demand Media TM remains a step ahead as it works at lightning speed to redefine this space. Demand Media's executive team has a vision for the future of new media and a proven track record for delivering on its vision. At the core of the company's unique approach is the combination of domain services, brandable domains and niche content web sites. Founded in May of 2006, Demand Media is privately held and is based in Santa Monica, CA, with offices in San Francisco and Seattle, WA.

Source: Demand Media on Freebase, licensed under CC-BY
Remixing linked data and spam

- Currently mostly the good guys are using Linked Data
- However, it's just too tempting to be left alone
- Automatic publishing of full articles using factual information from linked data?
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Pull additional content from Freebase

Use Zemanta or OpenCalais to add tags, images, links

Cover your tracks

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?

Publish
Publish

- On hosted third party platforms
  - eating their resources
- Platforms have hard time killing spammers
- Smaller ones don't necessarily have the incentive
- If they remove spammer too fast, it is easier for spammer to probe the limits
  - “kill with delay”
- Spam detection is resource intensive
Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Use Zemanta or OpenCalais to add tags, images, links

Pull additional content from Freebase

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?
Valuable comments

As I write this post, Zemanta is showing me 5 different articles that are related to my post. I could visit each one of these sites and reach out to the owner to see if they would be interested in linking to my post, or I could leave a valuable comment on the page and include a link back to my post.

Gather search terms (extensions, logs, guess)

Analyze → what people search for?

Find / create such content

Cover your tracks

Pull additional content from Freebase

Use Zemanta or OpenCalais to add tags, images, links

Use Zemanta to find similar blogs

Amazon Mechanical Turk to post comments and links back to your site

Profit?
Profit?

- sell ads
- sell links
- sell “fully developed site”
- to the highest bidder
Search engines to the rescue?

- Mahalo cut 10% of the staff the day after Google announced ranking changes
- Demand Media's stock isn't doing that well anymore
- However this is a never-ending story
Ecosystem

- Very sophisticated, large players
  - moving to more high quality content, video?
- Small time operations
  - using more and more sophisticated tools available on the market cheaply (modern asymmetric warfare?)
- Others specifically building tools to poison the web and sell them to small time operators
Food for thought
Can we make spammers (and others) work for us, making linked data better?

(think reCAPTCHA)
Could article directories be fruitfully used?
eZineArticles.com, GoArticles.com, etc...
Find rewritten articles and use them as parallel corpuses?
Could we use global workforce market more efficiently to get more linked data?
Thesis, antithesis, synthesis?

Spammers are breaking traditional CAPTCHAs with AI, so I've built a new system. It asks users to rate a slate of comments as "constructive" or "not constructive."

Then it has them reply with comments of their own, which are later rated by other users.

But what will you do when spammers train their bots to make automated constructive and helpful comments?


http://xkcd.com/810/
Thank you!

Questions?
Image sources

- http://www.flickr.com/photos/dzingeek/4587871752/
- http://www.flickr.com/photos/25101572@N02/4393474025/
- http://www.flickr.com/photos/jurvetson/542500748
- http://www.flickr.com/photos/legofenris/4288913574
- http://www.flickr.com/photos/ekilby/3733627940
- http://www.flickr.com/photos/ekilby/3732799269/
- http://www.flickr.com/photos/cipherswarm/38354452
- http://xkcd.com/810/