Technical Content Optimization

Presented by: Michael King
Contact: mike@ipullrank.com
Twitter: @iPullRank
No one knows Google’s algorithms.
No one knows Google's algorithms.

Patents are not an indication that something is currently in use.
No one knows Google's algorithms.

Patents are not an indication that something is currently in use. I'm not aiming to offend anyone with this presentation.
Search Engine Expectations
How do Search Engines See Content?
My Hypothesis

Just my educated opinion...

Thinking in terms of hacking, I wonder if Google's algos have always been more elegant than we thought and we've just been using links as the equivalent of brute force password cracking.

There have probably always been better "exploits" we've just always overindexed on links.

11:21 AM - 31 Jul 2018

6 Retweets 41 Likes
Our Correlation Studies Showcase Our Bias, Not Google's

I believe that because we are biased by the impact of links, we focus heavily on link acquisition. As a result our correlation studies reflect that.
PageRank certainly changed the game. However, at its simplest, it is largely just about link volume.
Google Cares About Content Relevance and Parity

In addition to the volume of links, the source and target content relevance is examined.
Our Metrics Mostly Measure Volume

First, our link metrics are, at best, just approximations. They are not necessarily indicative of anything that Google is actually doing or measuring.

Second, our link metrics are mostly some measure of link volume.
Except for Majestic.
Topical PageRank

It is not realistic to expect that Google has been using pure PageRank. Rather, it makes way more sense to expect they have defaulted to some form of Topical PageRank for several years.

h/t @CyrusShepard
But we just want the link, no matter what?
How Content is Understood in Information Retrieval

Content is tokenized and then relationships between terms are determined statistically.
Yet The Question is Always How Long Should Content Be?
How Content is Understood by Google

Google has a more sophisticated and living system for doing the same thing.
But we just “make great content?”
Content Scoring and Storing

Search engines can score documents based on a variety of features. Google’s file system suggests that they have the capability to store multiple versions of any document over time.

Figure 1: A slice of an example table that stores Web pages. The row name is a reversed URL. The contents column family contains the page contents, and the anchor column family contains the text of any anchors that reference the page. CNN’s home page is referenced by both the Sports Illustrated and the MY-look home pages, so the row contains columns named anchor:cnni.com and anchor:my.look.ca. Each anchor cell has one version; the contents column has three versions, at timestamps $t_3$, $t_5$, and $t_6$. 
Google’s Scoring Functions

There’s more than one scoring function. Google scores content and links a variety of different ways and then chooses the best results.

There is not just one “algorithm.”
Document Scoring Simplified

Content Factor + Content Factor + Speed Factor + Link Factor + Link Factor = Document Score
Each Component to the Equation Has a Weight

\[ a + b + c + d + e = \text{Document Score} \]
The Weights May Look Like This

3 (Content Factor) + 6 (Content Factor) + 1 (Speed Factor) + 2 (Link Factor) + 2 (Link Factor) = Document Score
This is What SEOs Do

Content Factor + Content Factor + Speed Factor + Link Factor + Link Factor = 148
So, Then, Google Turns Down the Weights on Links

Content Factor 3 5 + Content Factor 6 2 + Speed Factor 1 4 + Link Factor 95 .25 + Link Factor .01 74 = 55.49
When Really, Content Optimization May Go Further

\[3 \times 100 + 6 \times 100 + 1 \times 4 + 0.25 \times 95 + 0.01 \times 74 = 928.49\]
Google Constantly Evaluates and Experiments

DS: Last month, we had the Quality Update. How is Google assessing the quality? How do clicks factor in?

GI: We use clicks in different ways. The main things that we use clicks for evaluation and experimentation. There are many, many people who are trying to induce noise in clicks. Rand Fishkin, for example is experimenting with clicks. Using clicks directly in ranking would be a mistake. In personalized results, if you search for apple, we would most likely serve you a disambiguation box. We have to figure out if you mean the company or the food. Then, we’d look at the click you made.
There is Evidence of CTR Expectations

Granted, the screenshot is from the Paid side, but this suggests that Google has at least considered an expected CTR. This is also consistent with the idea of evaluating a successful search based on click activity.

The hypothesis being that if a result does not consistently meet the expected CTR, it should be demoted.
Experiments.

But we don't A/B test our metadata?
Queries Become Entities

Google has explicitly indicated that they break a query into a series of entities first.
Attributes of Entities Inform Context

Google builds context around the query leveraging the information about the entity to inform what documents to consider.
Entities are Extracted From Content to Inform Rankings

When Google is processing content, they extract ad score entities.

Ranking entity realizations for information retrieval

Abstract

Methods, systems, and apparatus, including computer programs encoded on a computer storage medium, for identifying and ranking entities for reference as search results. In one aspect, a method includes receiving data identifying resources that are relevant to a query. The data for each resource can include a relevance score, a list of references to entity realizations included in the resource, and for each reference to an entity realization in the list, one or more resource reference scores. For each resource and for each reference to an entity realization in the resource, a partial score for the reference can be determined from the resource reference scores for the reference and the relevance score for the resource. For each reference to an entity realization, a reference score for the reference is determined from each of the partial scores for the reference. Search results can be ranked based on the reference scores.

Images (7)

Classifications

G06F17/3053
Entity Salience.

But our industry doesn’t develop entity strategies?
Intent.

Intent for a given keyword can change over time.
Intent for a given keyword can change over time.

Yet we set and forget our keyword research?
Technical Content Optimization is a Multi-dimensional Word Game, but...
..we don't know what we’re doing
Mind the Text Analysis Knowledge Gap
What is missing from the SEO skillset
We're Entering Different Fields

We are no longer talking about SEO. Rather, we are talking about natural language processing, computational linguistics, information retrieval and graph theory. All things that conceptually underpin the computer science behind how search engines model topics in content.
Natural Language Processing Use Cases
A “Corpus” is Built from Crawling

Googlebot crawls the web and builds a corpus of web documents. This is how the process begins.
Documents are Parsed

Documents are broken down into words and then tagged with parts of speech.
Topics are Modeled

A series of transformations and analyses are done on the content to understand what they are about.
Language is Understood

Once the content is understood, and features are extracted, it can then be scored and used for rankings. This is how search engines develop their expectation of what should be featured in content.
Google Has Something Called BERT

BERT is Google’s technique for pre-training NLP for classification, question answering and named entity recognition.


Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing

Friday, November 2, 2018

Posted by Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language

One of the biggest challenges in natural language processing (NLP) is the shortage of training data. Because NLP is a diversified field with many distinct tasks, most task-specific datasets contain only a few thousand or a few hundred thousand human-labeled training examples. However, modern deep learning-based NLP models see benefits from much larger amounts of data, improving when trained on millions, or billions, of annotated training examples. To help close this gap in data, researchers have developed a variety of techniques for training general purpose language representation models using the enormous amount of unannotated text on the web (known as pre-training). The pre-trained model can then be fine-tuned on small-data NLP tasks like question answering and sentiment analysis, resulting in substantial accuracy improvements compared to training on these datasets from scratch.

This week, we open sourced a new technique for NLP pre-training called Bidirectional Encoder Representations from Transformers, or BERT. With this release, anyone in the world can train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU. The release includes source code built on top of TensorFlow and a number of pre-trained language representation models. In our associated paper, we demonstrate state-of-the-art results on 11 NLP tasks, including the very competitive Stanford Question Answering Dataset (SQuAD v1.1).
Finally, a Machine That Can Finish Your Sentence

Completing someone else’s thought is not an easy trick for A.I. But new systems are starting to crack the code of natural language.
In other words, it's the thing that does this.
Learn More about BERT (and ELMo)

Jay Alammar has a great article with illustrations that walks you through BERT and some of the other pre-existing NLP mechanisms

http://jalammar.github.io/illustrated-bert/

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

The year 2018 has been an inflection point for machine learning models handling text (or more accurately, Natural Language Processing or NLP for short). Our conceptual understanding of how best to represent words and sentences in a way that best captures underlying meanings and relationships is rapidly evolving. Moreover, the NLP community has been putting forward incredibly powerful components that you can freely download and use in your own models and pipelines (It’s been referred to as NLP’s ImageNet moment, referencing how years ago similar developments accelerated the development of machine learning in Computer Vision tasks).
BERT Q&A Demo

Get a preview of what Google might extract from a page for a given question with this BERT demo.

https://www.pragnakalp.com/demos/BERT-NLP-QnA-Demo/
More than Keywords: 7 Concepts of Advanced On-Page SEO

By: Cyrus Shepard

October 21st, 2014

The author's views are entirely his or her own (excluding the unlikely event of hypnosis) and may not always reflect the views of Moz.

"What is this page about?"

As marketers, helping search engines answer that basic question is one of our most important tasks. Search engines can't read pages like humans can, so we incorporate structure and clues as to what our content means. This helps provide the relevance element of search engine optimization that...
Keyword Usage

As you might imagine, leveraging Text Statistics to inform rankings indicates that you might want to use the keywords on the page.

This is where having the opportunity to rank begins.
Tokenization

Search Engines break paragraphs into sentences and sentences into “tokens” or individual words.

This better positions content for statistical analysis
N-grams

N-grams are terms or phrases of N-length.

<table>
<thead>
<tr>
<th>Full sentence</th>
<th>It does not, however, control whether an exaction is within Congress’s power to tax.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>“It”; “does”; “not.”; “however.”; “control”; “whether”; “an”; “exaction”; “is”; “within”; “Congress’s”; “power”; “to”; “tax.”</td>
</tr>
<tr>
<td>Bigrams</td>
<td>“It does”; “does not.”; “not, however.”; “however, control”; “control whether”; “whether an”; “an exaction”; “exaction is”; “is within”; “within Congress’s”; “Congress’s power”; “power to”; “to tax.”</td>
</tr>
<tr>
<td>Trigrams</td>
<td>“It does not”; “does not, however”; “not, however, control”; “however, control whether”; “control whether an”; “whether an exaction”; “an exaction is”; “exaction is within”; “is within Congress’s”; “within Congress’s power”; “Congress’s power to”; “power to tax.”</td>
</tr>
</tbody>
</table>
All Our N-gram are Belong to You

Thursday, August 3, 2006

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google’s datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there’s no data like more data, and scaled up the size of our data by one order of magnitude and then another, and then one more - resulting in a training corpus of one trillion words from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That’s why we decided to share this enormous dataset with everyone. We processed 1,024,906,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,911 unique words, after discarding words that appear less than 200 times.

Watch for an announcement at the Linguistics Data Consortium (LDC), who will be distributing it soon, and then order your set of 6 DVDs. And let us hear from you - we’re excited to hear what you will do with the data, and we’re always interested in feedback about this dataset, or other potential datasets that might be useful for the research community.

Update (22 Sept. 2006): The LDC now has the data available in their catalog. The counts are as follows:

<table>
<thead>
<tr>
<th>File sizes: approx. 24 GB compressed (gzip’ed) text files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tokens: 1,024,906,267,229</td>
</tr>
<tr>
<td>Number of sequences: 95,119,469,564</td>
</tr>
<tr>
<td>Number of unigrams: 13,588,911</td>
</tr>
</tbody>
</table>
Synonyms and Close Variants

Statistical relevance is computed using both synonyms and close variants of words.
Stemming

Search Engines break words into their stems to better determine relationships and relevance.

```r
text <- "love loving lovingly loved lover lovely love"
text_tokens(text, stemmer = "en") # english stemmer
```

```
[[1]]
[1] "love"  "love"  "love"  "love"  "lover"  "love"  "love"
```
Lemmatization

Similar to stemming, lemmatization is the grouping of inflected forms of the same word or idea.

For example, “better” and “good” have the same lemma.

Semantic Distance & Term Relationships

Search engines look for how physically close words are together to better determine their relationship in statistical models.
Latent Dirichlet Allocation

LDA is a recursive algorithm for mapping keywords to topics.
<table>
<thead>
<tr>
<th>Row ID</th>
<th>Document</th>
<th>T Term</th>
<th>D topic_0</th>
<th>D topic_1</th>
<th>D topic_2</th>
<th>D topic_3</th>
<th>D topic_4</th>
<th>D topic_5</th>
<th>D topic_6</th>
<th>D topic_7</th>
<th>D topic_8</th>
<th>D topic_9</th>
<th>S Assign...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row0</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>SEO</td>
<td>0.231</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.546</td>
<td>0.222</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row1</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Tools</td>
<td>0.259</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.491</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row2</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>The</td>
<td>0.056</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.722</td>
<td>0.222</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row3</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Complete</td>
<td>0.074</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td>0.185</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row4</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>List</td>
<td>0.139</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.694</td>
<td>0.167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row5</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>(2019...</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.582</td>
<td>0.159</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row6</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Update</td>
<td>0.157</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.666</td>
<td>0.176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row7</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>SECodepat</td>
<td>0.287</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.472</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row8</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>site</td>
<td>0.204</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.639</td>
<td>0.157</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row9</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>probably</td>
<td>0.213</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.676</td>
<td>0.111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row10</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>most</td>
<td>0.111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row11</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>advanced</td>
<td>0.194</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.518</td>
<td>0.278</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row12</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Complete</td>
<td>0.167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.676</td>
<td>0.157</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row13</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>research</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.629</td>
<td>0.278</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row14</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>tool</td>
<td>0.222</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.713</td>
<td>0.065</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row15</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>on</td>
<td>0.083</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.685</td>
<td>0.231</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row16</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>the</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.602</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row17</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>market</td>
<td>0.342</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.565</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row18</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Yes</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.583</td>
<td>0.176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row19</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>Close</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.777</td>
<td>0.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row20</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>enter</td>
<td>0.102</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.62</td>
<td>0.278</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row21</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>all</td>
<td>0.102</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.648</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row22</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>seed</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.611</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row23</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>and</td>
<td>0.194</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.657</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row24</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>get</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.537</td>
<td>0.222</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row25</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>suggest</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.685</td>
<td>0.167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row26</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>like</td>
<td>0.157</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.731</td>
<td>0.102</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row27</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>any</td>
<td>0.046</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.611</td>
<td>0.343</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row28</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>other</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.509</td>
<td>0.352</td>
<td>0.046</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row29</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>tool</td>
<td>0.194</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.602</td>
<td>0.204</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row30</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>But</td>
<td>0.213</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.62</td>
<td>0.12</td>
<td>0.046</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Row31</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>can</td>
<td>0.241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.648</td>
<td>0.111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row32</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>do</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.676</td>
<td>0.176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row33</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>so</td>
<td>0.213</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.657</td>
<td>0.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row34</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>much</td>
<td>0.333</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.555</td>
<td>0.111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row35</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>more</td>
<td>0.148</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.759</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row36</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>than</td>
<td>0.213</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.611</td>
<td>0.176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
<tr>
<td>Row37</td>
<td>SEO Tools: The Complete List (2019 Update...)</td>
<td>list</td>
<td>0.065</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.759</td>
<td>0.176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>topic_6</td>
</tr>
</tbody>
</table>
Zipf’s Law

Zipf’s law is a theory that words will be similarly distributed across documents in the same corpus (or document set).
Zipf’s Law Applied

When run on an actual dataset, Zipf’s law tends to hold up in high rankings.
Term Frequency Inverse Document Frequency

Term frequency, inverse document frequency (TF-IDF) identifies the importance of keywords with respect to other keywords in documents in a corpus.

**TF-IDF**

Term frequency–inverse document frequency (TF-IDF) measures the importance of a keyword phrase by comparing it to the frequency of the term in a large set of documents. Many advanced textual analysis techniques use a version of TF-IDF as a base.
Algorithmical Inspirations for Better Writing

or why Text Optimization with TF-IDF is the key to success.

Search Engine Optimization has been around for many years. So far, however, our focus has been primarily on rankings and backlinks. We have monitored and exchanged links, bought links and built links by ourselves. But for years now, the search engines have been trying to convince us of a slightly different approach: "You need good and relevant content that is especially valuable for your users". And so we started to write content.
Phrase-based Indexing & Co-occurrence

Google specifically looks for keywords that co-occur with a target keyword in a document set.

Usage of those keywords is an indication of relevance for subsequent documents.

Phrase-Based Indexing and Co-occurrence

Using the concept of co-occurrence, search engines know that certain phrases tend to predict other phrases. Presence of these co-occurring phrases can strengthen topic focus. Links from pages with co-occurring phrases can also help.
But You Heard TF-IDF is Old And You Shouldn’t Focus On it...

I actually agree with this, you should not focus on TF-IDF in isolation.

I suspect that Google is using a series of text analysis functions to calculate relevance, but the age of the technique does not make it any less valuable.
Heap’s Law

Heaps law indicates that vocabulary within a corpus grows at a predictable rate.
**Entity Salience**

Search engines leverage the features of the entity to understand the context in which the entity is referenced. Implicit features of the entity may be used to understand the document.
Hidden Markov Models

Hidden Markov Models allow search engines to extract implicit entities.
Entities are Also Used in Verification

Google double checks what you’re talking about by using entities.

H/t @TomAnthonySEO

Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang
Wilko Horn, Camillo Lugaresi, Shaohua Sun, Wei Zhang
Google Inc.
{lunadong.gabrilovich@kpmurphy.vandang.wilko.camillo@sunsh.weizhibl@gmail.com}

ABSTRACT

The quality of web sources has been traditionally evaluated using exogenous signals such as the hyperlink structure of the graph. We propose a new approach that relies on endogenous signals, namely, the correctness of factual information provided by the source. A source that has few false facts is considered to be trustworthy.

The facts are automatically extracted from each source by information extraction methods commonly used to construct knowledge bases. We propose a way to distinguish errors made in the extraction process from factual errors in the web source per se, by using joint inference in a novel multi-layer probabilistic model.

We call the trustworthiness score we computed Knowledge-Based Trust (KBT). On synthetic data, we show that our method can reliably compute the true trustworthiness levels of the sources. We then apply it to a database of 2.8B facts extracted from the web, and thereby estimate the trustworthiness of 1/3M webpages. Manual evaluation of a subset of the results confirms the effectiveness of the method.

1. INTRODUCTION

"Learning to trust is one of life’s most difficult tasks."
- Isaac Watts

Quality assessment for web sources is of tremendous importance in web search. It has been traditionally evaluated using exogenous signals such as hyperlinks and browsing history. However, such signals mostly capture how popular a webpage is. For example, the gossip websites listed in [16] mostly have high PageRank scores [4], but would not generally be considered reliable. Conversely, some less popular websites nevertheless have very accurate information.

In this paper, we address the fundamental question of estimating how trustworthy a given web source is. Informally, we define the trustworthiness or accuracy of a web source as the probability that it contains the correct value for a fact (such as Barack Obama’s nationality), assuming that it mentions any value for that fact. (Thus we do not penalize sources that have few facts, so long as they are correct.)

We propose using Knowledge-Based Trust (KBT) to estimate source trustworthiness as follows. We extract a plurality of facts from many pages using information extraction techniques. We then jointly estimate the correctness of these facts and the accuracy of the sources using inference in a probabilistic model. Inference is an iterative process, since we believe a source is accurate if its facts are correct, and we believe the facts are correct if they are extracted from an accurate source. We leverage the redundancy of information on the web to break the symmetry. Furthermore, we show how to initialize our estimate of the accuracy of sources based on authoritative information, in order to ensure that this iterative process converges to a good solution.

The fact extraction process we use is based on the Knowledge Vault (KV) project [10]. KV uses 16 different information extraction systems to extract (subject, predicate, object) knowledge triples from webpages. An example of such a triple is (Barack Obama, nationality, USA). A subject represents a real-world entity, identified by an ID such as nuis in Freebase [2]; a predicate is pre-defined in Freebase, describing a particular attribute of an entity; an object can be an entity, a string, a numerical value, or a date.

The facts extracted by automatic methods such as KV may be wrong. One method for estimating if they are correct or not was described in [11]. However, this earlier work did not distinguish between factual errors on the page and errors made by the extraction system. As shown in [11], extraction errors are far more prevalent than source errors. Ignoring this distinction can cause us to incorrectly distrust a website.

Another problem with the approach used in [11] is that it estimates the reliability of each webpage independently. This can cause problems when data are sparse. For example, for more than one billion webpages, KV is only able to extract a single triple (other extraction systems have similar limitations). This makes it difficult to reliably estimate the trustworthiness of each source.
Page Segmentation

Modern Search engines determine both prominence of content and review the visual hierarchy to determine how valuable the content is to the page.
Search engines are recursive.
We need to understand Google's statistical expectation of content and add it to our content's requirements.
What Should We Be Doing?

Let's Get Actionable
Step 1 - Research & Create:
Create content that fulfills expectations
We need to create or optimize content that balances user needs and search engine expectations. This requires researching the query space based on what is currently ranking in Google’s corpus.

Step 2 - Build Relevant Links:
Build Links with Content Parity
Building links needs to be a more detailed process. You need to only queue up link prospects that are topically relevant. These are the links that will propel you further than links from random unrelated content.

Step 3 - Test & Optimize:
Measure performance and experiment with metadata
You need to continually review the performance of your content against that of the average performance in your space. If you are performing below average, this an opportunity to perform experiments on your metadata.
We need to do more of Column B

Our industry talks a lot about the concepts in the column B, but in practice we do not have enough SEO tools to facilitate the latter.
It Starts with Keyword Research

This cannot be the extent of your keyword research in 2019.
Let's Take it back to 2013

At minimum you should be pulling, keywords, search volume, ranking difficulty, co-occurring keywords, target personas, need states, entities.

Check out my Persona-driven Keyword research deck for how to pull most of this.

https://www.slideshare.net/ipullrank/persona-driven-keyword-research/
The ultimate goal is to determine a series of co-occurring terms and entities to incorporate into your content for each target keyword.
Meet Knime Analytics

https://www.knime.com/knime-software/knime-analytics-platform
Knime Allows for Drag & Drop Text Analysis

More than just text analysis, Knime has a series of features for.
TextProcessing Extension

https://www.knime.com/knime-text-processing
Palladian HTTP Extension

https://www.knime.com/community/palladian
Download a “Corpus” for a Keyword

Ideally, we’d review all of the results for a given keyword to achieve parity with Google, but in this case we’ll download the top 100 results for “seo tools” from SEMRush.
Here’s a Starter Workflow

This workflow allows you to get started with LDA and computes Term Frequency.
1. Put your domain into SEMRush.
2. See what you don't rank well for in organic research.
### 3. Pick a keyword.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Pos</th>
<th>Search Volume</th>
<th>KD</th>
<th>CPC (USD)</th>
<th>URL</th>
<th>Traffic %</th>
<th>Costs %</th>
<th>CTR</th>
<th>Results</th>
<th>Trend</th>
<th>SERP</th>
<th>Last Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEO consultant</td>
<td>6</td>
<td>4,800</td>
<td>62.40</td>
<td>19.02</td>
<td>moz.com/community/recommended</td>
<td>0.05</td>
<td>0.18</td>
<td>0.39</td>
<td>81,300,000</td>
<td></td>
<td></td>
<td>21 hr ago</td>
</tr>
<tr>
<td>SEO agency</td>
<td>3</td>
<td>4,800</td>
<td>69.41</td>
<td>25.90</td>
<td>moz.com/community/recommended</td>
<td>0.10</td>
<td>0.44</td>
<td>0.67</td>
<td>129,600,000</td>
<td></td>
<td></td>
<td>22 hr ago</td>
</tr>
<tr>
<td>SEO optimization</td>
<td>2</td>
<td>3,600</td>
<td>75.90</td>
<td>14.44</td>
<td>moz.com/seo-primer-guide-to-seo</td>
<td>0.11</td>
<td>0.29</td>
<td>0.52</td>
<td>87,900,000</td>
<td></td>
<td></td>
<td>19 hr ago</td>
</tr>
<tr>
<td>SEO expert</td>
<td>19</td>
<td>3,000</td>
<td>55.07</td>
<td>20.33</td>
<td>moz.com/seo-expert-guide</td>
<td>0.02</td>
<td>0.09</td>
<td>0.35</td>
<td>57,300,000</td>
<td></td>
<td></td>
<td>21 hr ago</td>
</tr>
<tr>
<td>Local SEO</td>
<td>5</td>
<td>3,000</td>
<td>65.30</td>
<td>12.26</td>
<td>moz.com/local-seo</td>
<td>0.42</td>
<td>0.90</td>
<td>0.41</td>
<td>126,900,000</td>
<td></td>
<td></td>
<td>19 hr ago</td>
</tr>
<tr>
<td>SEO website</td>
<td>13</td>
<td>2,000</td>
<td>75.50</td>
<td>10.96</td>
<td>moz.com/learning-actors</td>
<td>&lt; 0.01</td>
<td>0.01</td>
<td>0.51</td>
<td>381,000,000</td>
<td></td>
<td></td>
<td>13 hr ago</td>
</tr>
<tr>
<td>SEO robots</td>
<td>12</td>
<td>2,000</td>
<td>75.50</td>
<td>10.96</td>
<td>moz.com/seo-primer-guide-to-seo</td>
<td>&lt; 0.01</td>
<td>0.01</td>
<td>0.51</td>
<td>381,000,000</td>
<td></td>
<td></td>
<td>13 hr ago</td>
</tr>
<tr>
<td>SEO website</td>
<td>14</td>
<td>2,000</td>
<td>75.50</td>
<td>10.96</td>
<td>moz.com/learning-actors</td>
<td>&lt; 0.01</td>
<td>0.01</td>
<td>0.51</td>
<td>381,000,000</td>
<td></td>
<td></td>
<td>13 hr ago</td>
</tr>
</tbody>
</table>

**What does SEO stand for?**

- SEO Analysis
- Best SEO Company
- Best SEO Company
- WordPress SEO
- WordPress SEO
- Squarespace SEO
- SEO Specialist
- SEO Ranking

**SEOs**

- Moz.com/free-seo-tools
- Moz.com/buying-guides
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/option
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
- Moz.com/community/recommended
4. Look at the SERP.
What does SEO stand for?

What does SEO mean, and what do I REALLY need to know about it?

Really Google?
IS THIS YOUR #1 RESULT?!
CHAPTER THREE

WHY SEARCH ENGINE MARKETING IS NECESSARY

An important aspect of SEO is making your website easy for both users and search engine robots to understand. Although search engines have become increasingly sophisticated, they still can’t see and understand a web page the same way a human can. SEO helps the engines figure out what each page is about, and how it may be useful for users.

A Common Argument Against SEO

We frequently hear statements like this:

"No smart engineer would ever build a search engine that requires websites to follow certain rules or principles in order to be ranked or indexed. Anyone with half a brain would want a system that can crawl through any architecture, parse any amount of complex or imperfect code, and still find a way to return the most relevant results, not the ones that have been developed by unscrupulous search engine marketers."
5. See what else could rank for our site.
Content Success

Plan and write better content

Forget endless keyword research and hours of writing drafts upon drafts. Content Success cuts that work in half by assisting you at each stage of the content creation process. It monitors your content's performance, present new topic ideas and helps you write new content for your target audience. Guarantee content your users will love with Content Success.

Enter Ryte's Content Success.
6. Put in the keyword and compare TF-IDF to ranking pages.
You can see the corpus they pulled for review.
7. Put the copy in. See what important keywords are missing and make edits.
Content Optimization implemented

Results with no Link Building (Competitive Medicare Terms Min MSV 8100)
Content Optimization implemented

Results with No Link Building
(Competitive Medicare Terms Min MSV 8100)
Agile Content Development

How do you make great revenue-driving online content?

Iterate, iterate, iterate. With Searchmetrics, agile content development backs the creative process for the first time with deep learning and the latest in data science.

Create lively content that is also optimized to meet user intent and stand atop the search page rankings. Drive predictable revenue gains using the new future of web content development.
The topic explorer helps you identify semantically clustered keyword opps.
You can generate a new page around the core topics.
It abstracts a lot of the topic modeling and gives you detailed direction as you write.
TECHNICAL CONTENT OPTIMIZATION

OVERVIEW

Check your texts for SEO friendliness on-the-go

SEO Writing Assistant is an extension that provides instant recommendations for content optimization based on the best-performing articles in Google top 10. Content professionals, bloggers, and marketing experts can quickly assess if their texts or the texts of external writers meet various content-quality requirements in just a few seconds.
**A/B Test Metadata**

A/B test your page titles and meta descriptions by creating different buckets of pages, making changes, and reviewing their CTR performance.

**SEO A/B TESTING IS PAGE ORIENTED**

- Pages 1, 2, 3: Control Layout
- All users see the same as each other
- Pages 4, 5, 6: Test Layout
Analysis: Compare your CTR

AWR publishes CTR stats by vertical [https://advancedwebranking.com/ctrstudy](https://advancedwebranking.com/ctrstudy)
Hypothesis: Identify Specifically Where You Underperform

Develop your hypotheses based keywords that underperform based on the expected CTR.
Use Wordnet to Scale Keyword Variant Discovery

https://wordnet.princeton.edu/
Automate Inclusion of Emerging Keywords from GSC in Rankings

Monitor your queries on a page-level week over week to identify new emerging keyword opportunities and automatically add them to your rank tracking.
11 New APIs for GrepWords Customers!

We have spent the last 2 weeks now rolling out a feature every single day. If you need a quick recap, we rolled out... Local Intent Keywords Commercial Intent Keywords Video Intent Keywords Suggestions for Amazon Suggestions for Ebay...
Review Competitors’ Usage of Entities

When Google is processing content, they extract and score entities. Compare your page’s entity salience and ordering against your competitors to identify what you need to feature.

Do the same as you do your keyword research.
Use Entities for Intelligent Internal Link Building

You can use the same data on your own site to determine how to automate internal link building.
Cloud Natural Language

Derive insights from unstructured text using Google machine learning.

GO TO CONSOLE

View documentation for this product.

Powerful text analysis

Google Cloud Natural Language reveals the structure and meaning of text both through powerful pretrained machine learning models in an easy to use REST API and through custom models that are easy to build with AutoML Natural Language BETA. Learn more about Cloud AutoML.

You can use Cloud Natural Language to extract information about people, places, events, and much more mentioned in text documents, news articles, or blog posts. You can use it to understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app. You can analyze text uploaded in your request or integrate with your document storage on Google Cloud.
This on-site investigation focused on the performance of the Certified Advanced 208-Compliant air bag system in a 2005 Ford Escape 4x4 sport utility vehicle. This two-vehicle crash occurred in July 2014 at 3:59 p.m. in the state of Colorado. The crash occurred on a curved portion of a three-lane interstate roadway. The Ford Escape lost control on an interstate highway and struck a concrete barrier on the right side of the roadway. The impact resulted in sufficient longitudinal deceleration of the Escape to command the deployment of the frontal air bag system and activation of the driver’s seat belt pretensioner. The vehicle rotated out from the initial wall impact and was subsequently struck by a 2013 BYD Qin BYD Qin pulling a single trailer. The restrained 48-year-old 48-year-old male driver of the Escape appears to have sustained a minor facial injury.
Knowledge Graph
- Query the web for rich connected entities
  
  Learn More

AI: X
- Extract structured data from any URL with AI Extractors
  
  Learn More

Crawlbot
- Scale up your extraction to 10,000s of domains with Crawlbot
  
  Learn More
Build High Accuracy Text Analysis

From topic classification and sentiment analysis to entity extraction. Train and integrate custom machine learning models in a matter of days, not months.

Customized text analysis

Obtain high accuracy results by training your custom...
Uncomfortable with APIs?

If APIs sound daunting to you, here's a Knime Analytics workflow that will make it easier for you to pull data from an API and format it into a table.

https://www.knime.com/blog/a-restful-way-to-find-and-retrieve-data
Natural Language Generation is the Next Big Thing

I suspect that in the near future, a lot of this will be available to automate.

There are already solutions that are more than just sophisticated content spinners.
Unleashing the Power of Words
Phrasetech Provides an Enterprise Solution for Automated Text Creation, Control and Optimization

Schedule a Demo  View Examples
Wrapping Up

Bringing it Home!
How to do effective Technical Content Optimization

A. Determine Search Engine Expectations

B. Optimize your Content

C. Test All the Things

D. Win the SERPs
I'm Mike King.

(@IPULLRANK)
We are iPullRank.
(@IPULLRANKAGENCY)
This is Our CEO.
(And, I'm #ZORASDAD)
At iPullRank, we're incorporating Natural Language Processing into our workflows to better understand the expectations of Google.
Thank You | Q&A

Mike King
Managing Director

Twitter: @iPullRank
Email: mike@ipullrank.com