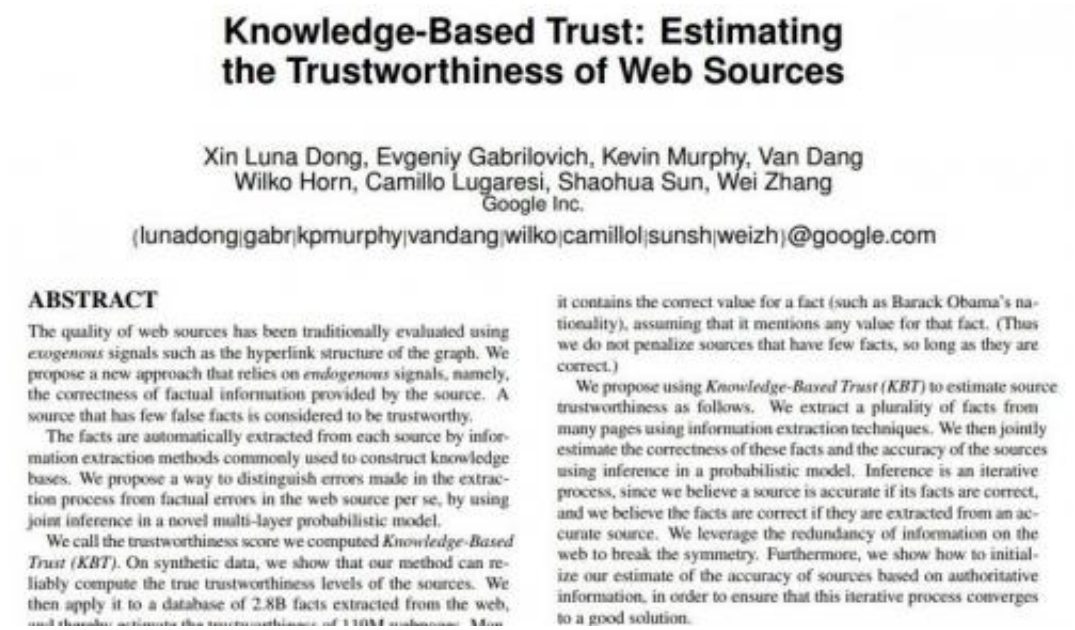
Google looking at ways to rate websites based more on trustworthiness

March 2 2015, by Bob Yirka



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|gabr|kpmurphy|vandang|wilko|camillol|sunsh|weizh)@google.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,

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.

A team of researchers at Google has been looking into ways to change the way links are retrieved by its famous search engine—instead of ranking them based on popularity, the researchers are looking into ways of ranking based on the trustworthiness of the site, which would be based on information the web agrees is factual. In their paper they have uploaded to the *arXiv* preprint sever, the team describes their ideas and what they have found thus far.

Information returned by Google's search engine has become more and more important over the past several years—where once it was considered entertaining or merely useful, now it is big business. Companies put a lot of money into making sure links for their products rank high on the list, which in turn means that the actual real-world value of the information Google returns has increased dramatically. Also at stake is real-world money attached to page views, aka "clicks"—higher rankings on Google generally translate to more people clicking on a link, which means more money for the owner of that link. Recognizing the value their search data represents, Google is responding by looking into ways to provide more value to people that use their search engine. Instead of simply ranking by a site based on how many other sites link to it, Google wants to factor in whether in information on sites it lists is actually truthful.

We all know that there is a plethora of links on the web that take us to places we do not trust, or want be at in the first place—we get taken in by come-on's or by headings that promise one thing and deliver something different, or find ourselves visiting a site that it is very obviously bogus and feeling foolish for it. Because of its stature in the search community, Google wants to change this. Their idea is to count the number of purported facts on a given web site and then compare those against a knowledge-based trusted source—returning a number (they call it a Knowledge-Based Trust index) that represents the trustworthiness of the site. Those with a higher trustworthiness number would appear before those with less trustworthiness in Google searches. The team notes that Google already has a "Knowledge Vault" that can be used as the trusted source. They report that their research thus far has revealed that their method can "reliably compute the true trustworthiness levels of the sources."

At this point, it is not clear if Google actually intends to implement such a change—if so, it could mean a round of misery for web site owners

who post bogus material for the express purpose of reaping cash rewards.

More information: Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources, arXiv:1502.03519 [cs.DB] arxiv.org/abs/1502.03519v1

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 119M webpages. Manual evaluation of a subset of the results confirms the effectiveness of the method.

via Newscientist