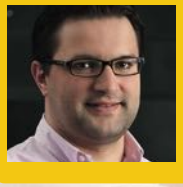
Tobias Aubele, Mario Fischer

Suchmaschinenoptimierung wird zweifelsfrei immer anspruchsvoller. Während es früher genügte, die Keywords gut verteilt in den Text zu streuen oder möglichst viele Links einzusammeln, selbst zu bauen oder einfach zu kaufen, erkennt Google immer mehr dieser Täuschungsversuche und ist bestrebt, sich insbesondere vom lästigen Joch der Linkmanipulationen zu befreien. Der jüngste Coup ist, eine eigene Wissensbasis aufzubauen, mit der u. a. die inhaltliche Vertrauenswürdigkeit von Websites und Webseiten besser bewertet werden kann. Das funktioniert offenbar recht gut – und diese Wissensbasis wird vom Umfang her in absehbarer Zeit wohl nahezu explodieren. Website Boosting erklärt, was auf ambitionierte Webmaster zukommt.

SEO 4.0

FAKTEN, FAKTEN, FAKTEN!

DER AUTOR



Tobias Aubele ist Professor für E-Commerce, insbesondere Conversion-Optimierung und Usability an der Hochschule Würzburg-Schweinfurt. Er promovierte im Bereich Konsumpsychologie an der University of Gloucestershire.

DER AUTOR

Mario Fischer ist Professor für E-Commerce an der Hochschule Würzburg-Schweinfurt und Herausgeber der Website Boosting. Er beschäftigt sich u. a. seit 1996 intensiv mit dem Thema Suchmaschinen.

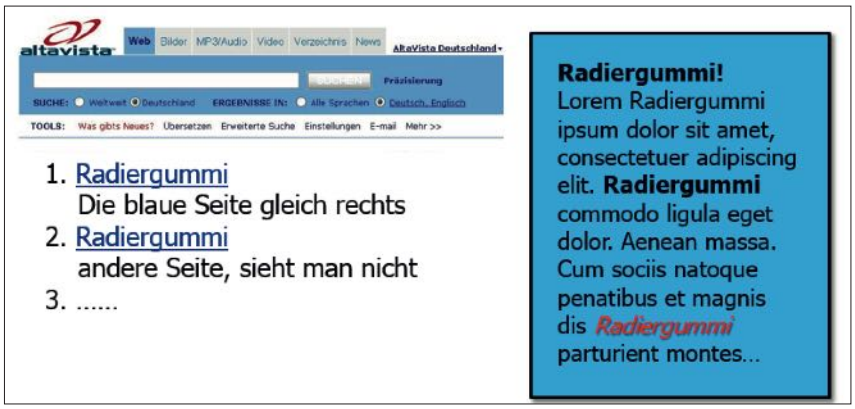


Abb. 1: SEO 1.0: Früher war alles einfacher – ein paar Keywords reichten

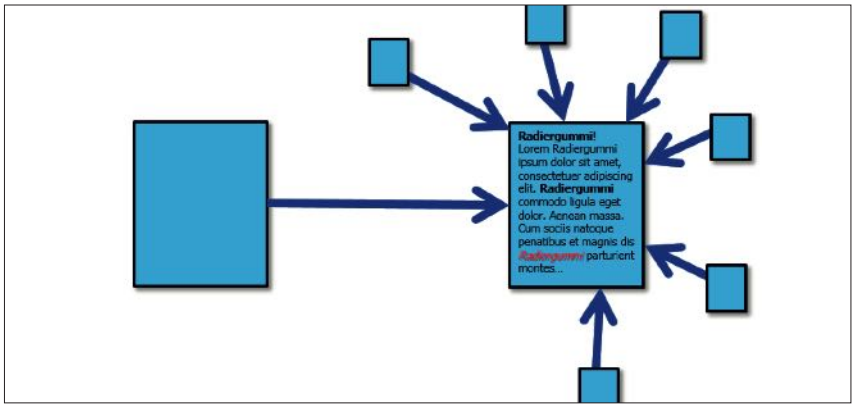


Abb. 2: SEO 2.0: Google betritt die Bühne und Backlinks werden wichtig

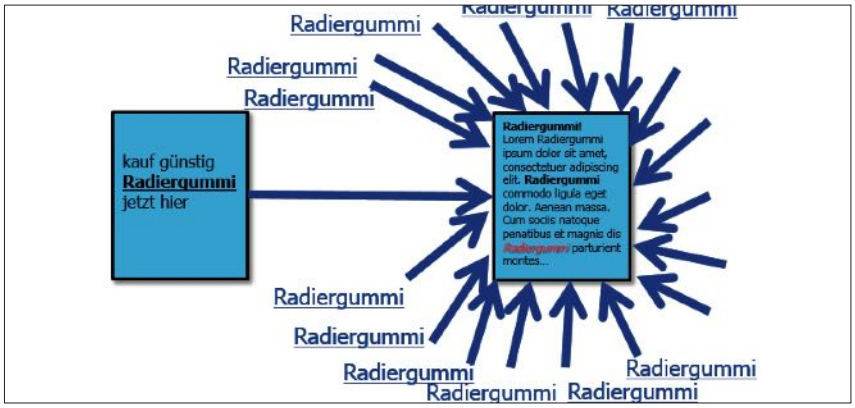


Abb. 3: SEO 3.0: Massiver Backlinkaufbau mit passenden Ankertexten

SEO 1.0

Wie einfach war SEO zu Zeiten, als es diese Bezeichnung bzw. Abkürzung noch gar nicht gab. Webseiten in Suchmaschinen wie Fireball, Excite oder AltaVista zu einem guten Ranking zu verhelfen, war damals, wie auf sitzende Enten zu schießen. In der Regel genügte es, das gewünschte Keyword nur häufig genug im Text zu nennen, und schon waren die erste Seite und oft sogar der erste Platz fest gebucht.

SEO 2.0

Am 04. September 1998 betrat dann Google die Bühne mit der Idee, Backlinks als externes Votum für eine Website als einen wesentlichen Rankingfaktor mit einfließen zu lassen. Der sog. PageRank verhinderte relativ zuverlässig, dass hochgejubelte Keyworddichten die begehrten Plätze stürmen konnten. Plötzlich wurde alles ein ganzes Stück anspruchsvoller und vor allem aufwendiger. Man muss sich auch vor Augen halten, dass es zu diesem Zeitpunkt weder eine SEO-Branche oder Community geschweige denn Publikationen oder Konferenzen zu diesem Thema gab. Was man wusste, hatte man ausprobiert und die Erkenntnisse in kleinen und engen Kreisen ausgetauscht.

Nach und nach wurde das „Geheimwissen" dann breiter geteilt und SEO betrat als Begriff die Bühne des Online-Marketings. Mehr und mehr setzte sich bei einzelnen Unternehmen die Erkenntnis durch, dass man mit einer optimierten Website oder einem optimierten Online-Shop über Google mehr Kunden gewinnen und damit mehr Geld verdienen konnte. Die ersten ernst zu nehmenden Agenturen schossen aus dem Boden und die SEO-Welle kam ins Rollen. Das war dann auch gleichzeitig der Auslöser für eine nächste Stufe, denn viel zu viele clevere Menschen nutzten die Erkenntnisse, um durch immer mehr Manipulationen immer mehr Spam in den Index von Google zu drücken. Einige wenige starke Links mit den passenden Ankertexten reichten aus, um eine Seite selbst für Keywords mit vielen Treffern auf die ersten Plätze der Ergebnisliste zu katapultieren – damals noch die puren, „ten blue links", mit Ausnahme einiger AdWords-Anzeigen auf der rechten Seite. Dort blieb sie dann so um die vier Wochen festgenagelt, denn vor dem sog. Caffeine-Hardware-Update bei Google gab es noch den regelmäßigen „Google Dance", mit dem die in zeitlich längeren Abständen neu gerechneten Ergebnisse auf alle Datencenter ausgerollt wurden. Bei Google hatte man mittlerweile erkannt, dass man schärfer gegen Manipulationen vorgehen musste, häufigere Updates in den Algorithmen brauchte und auch ein schlagkräftiges Spam-Abwehrteam.

SEO 3.0

In den Jahren um 2005 herum beginnend bis heute wurde es dann zunehmend schwieriger, Google etwas über Manipulationen unterzuschieben. Dies bedeutet natürlich nicht, dass es aktuell nicht mehr möglich ist. Der Aufwand ist aber ungleich höher und die Preise für den Kauf wirklich guter Backlinks sind mittlerweile in der Regel durchaus bereits vierstellig, weil nicht selten bereits mehrere Vermarkter in der Kette die Hand aufhalten und die Linkplätze jeweils mit saftigem Aufschlag weiterverkaufen.

Google machte den Manipulatoren mit zahlreichen Updates das Leben schwerer. Rund 400 Updates spielt Google pro Jahr in etwa ein und nur die größeren darunter erhalten Namen wie Allegra, Florida, Burbon, Austin, Venice, Vince, Pirate, Panda oder Pinguin und erlangen entsprechende Aufmerksamkeit.

Panda – gegen miesen Content

Die wohl bekanntesten – und diejenigen mit den größten Auswirkungen – sind das Panda- und das Pinguin-Update. Panda war und ist nach Meinung vieler Experten eher userverursacht und zielt auf sog. „Thin Content", also „dünne" bzw. inhaltsschwache Seiten ab. Da die inhaltliche Qualität bisher maschinell nur schwer wirklich verlässlich zu messen war, zog man offenbar die Bewegungen der Surfer mit zur Beurteilung heran. Wenn zu viele Besucher nach einem Klick auf ein Suchergebnis innerhalb kurzer Zeit zu den Suchergebnissen zurückkehren und auf andere Ergebnisse klicken, ist es sehr wahrscheinlich, dass auf der entsprechenden Seite wohl nichts wirklich Vernünftiges gefunden wurde. Google nennt solche schnellen Bounces „Short Click". „Long Clicks" sind dann im Gegenzug solche, bei denen der Suchende länger auf der Ergebnisseite verweilt oder sogar dort bleibt. Da

WAS IST EIN BROWSER-FINGERPRINT?

Wenn ein Browser eine Webseite von einem Server abfragt, gibt er im Gegenzug umfangreiche Informationen an diesen. Weitere Informationen können durch zusätzliche aktive Abfragemethoden ermittelt werden. Beim Datenaustausch mit einem Webserver wird nicht nur die IP-Adresse übermittelt, die sich bei dynamischer Vergabe, z. B. bei privaten DSL-Anschlüssen, ja ständig ändert, sondern auch Typ und Version des Browsers, welche Plug-ins installiert sind und welche Versionen diese haben. Auch die installierten Schriftarten (die z. T. deutlich variieren), das Betriebssystem und dessen Version sowie die darstellbaren Systemfarben können zur Identifikation eines Computers herangezogen werden. Je mehr Merkmale ermittelt werden, desto eindeutiger wird dieser „Fingerabdruck".

mittlerweile fast jeder zweite Webnutzer den hauseigenen Chrome-Browser nutzt, steht Google eine extrem verlässliche Datenquelle zur Verfügung. Durch die (fast) 50 %-Abdeckung aller Bewegungsdaten muss man gar nicht mehr von einer hochzurechnenden Stichprobe sprechen. Marktforscher würden für die Möglichkeit, die Hälfte aller potenziellen Kunden „befragen" zu können, wahrscheinlich töten. Der Fehler liegt hier praktisch bei null.

Kruden Ideen, wie sie immer wieder mal in Foren diskutiert werden, man könne seinen Mitbewerbern schaden, indem man ständig für diese über Suchergebnisse Short Clicks produziere, muss man allerdings eine Absage erteilen. Webnutzern lässt sich über sog. Browser-Fingerprints auch ohne die IP-Adresse relativ genau eine Nutzer-ID zuordnen und somit wird erkennbar, dass solche Klicks nur von einer oder wenigen Personen und nicht von einer breiten Masse kommen.

Pinguin – gegen miese Methoden

Über das Uservotum via Short Click kann man also als minderwertig empfundene Seiten erkennen, natürlich immer durch andere Signale abgesichert. Noch besser ist es natürlich, wenn solche Seiten erst gar nicht in den Suchergebnissen auftauchen. Um diesem Ziel näherzukommen, folgte dann das Pinguin-Update, mit dem man relativ umfangreich gegen übertriebene und manipulative Suchmaschinenoptimierung vorging, insbesondere gegen manipulierte und gekaufte Backlinks.

Topicerkennung

Wer die bisherigen Berichte über die sog. WDF/IDF-Formel in der Website Boosting verfolgt hat, weiß, dass Suchmaschinen auch in der Lage sind, das Hauptthema einer Webseite zumindest grob zu identifizieren und über Kookurrenzen anderer Wörter zu prüfen, wie treffsicher diese Einordnung ist. Das Prinzip dahinter ist einfach zu erklären, was nicht darüber hinwegtäuschen soll, dass es sehr schwer in Algorithmen zu gießen ist. Ein Beispiel: Auf einer Webseite tritt der Begriff „Läufer" häufiger auf. Handelt es sich hier nun um eine Seite, auf der es um Jogging geht? Bei der Analyse weiterer Worte tauchen allerdings Begriffe wie Spielzug, Bauer und Rochade auf und somit geht es wohl um das Spiel Schach bzw. eine spezielle Figur daraus, nämlich den Läufer. Kämen Wörter wie z. B. persisch, Muster, Orient und rutschfest vor, handelt es sich mit hoher Wahrscheinlichkeit nicht um das Spiel, sondern um einen länglichen Teppich für einen Flur. Suchmaschinen können also prüfen, ob ein Dokument mit den verwendeten Wörtern aus sich selbst heraus beweisen kann, dass es tatsächlich im Kern um ein bestimmtes Thema geht.

Manuelle Gütecontainer und manuelle Aktionen

Natürlich lässt sich nicht alles über Algorithmen realisieren. Zumindest jetzt noch nicht. Daher beschäftigt Google ein Heer sog. Qualitätstester, die im Nebenjob auf 450-Euro-Basis vorgegebene Webseiten absurfen und

„Google's investment in artificial intelligence does more than just create better search results; it allows Google's engineers to make constant algorithm changes right under our noses" **– Nate Dame**

ein menschliches Votum zu diesen Seiten abgeben. Dieses Votum erlaubt es Google seit vielen Jahren, Webseiten und Websites in verschiedene Kategorien wie z. B. cool, uncool, nützlich, spammy oder andere einzuteilen. Mithilfe dieser Cluster können dann neu entwickelte Algorithmen auf ihre Wirkungen hin live getestet werden. Ein weltweit agierendes Team an Spamfightern komplettiert, wohin die Algorithmen heute noch nicht hinreichen oder wo sie noch zu unscharf Spam von Nicht-Spam trennen. Die Spamfighter recherchieren, analysieren und prüfen manuelle und automatisch eingehende Meldungen und vergeben in begründeten Fällen unterschiedlich wirkende Strafen.

Diese und einige andere Maßnahmen machen die Suchmaschinenoptimierung deutlich anspruchsvoller, weil der „schnelle" Weg immer weniger funktioniert – und schon gar nicht auf mittlere bis längere Sicht hin. Und nun setzt Google wohl in absehbarer Zeit nochmals eins drauf! Die inhaltliche Erkennung von Webcontent schreitet mit großen Schritten voran und die Wirkungen könnten durchaus ein Hochzählen der Version auf 4.0 rechtfertigen.

SEO 4.0

Semantik

Bereits kurz nach der Existenz des Webs wurde von Vordenkern darüber diskutiert, wie man Inhalte so darstellen könnte, dass mehr Informationen über diese Inhalte und ihre Art vorliegen. Die Idee nannte man „Semantic Web". HTML ist ja eine Auszeichnungssprache. Mit ihr legt man fest, wie und wo eine Zahl wie „13,50" dargestellt wird. Wir Menschen erkennen aus dem Kontext die semantische Bedeutung, so z. B., dass es sich um eine Preisinformation handelt, wenn sie an den gewohnten Stellen auf einer Produktseite eines Online-Shops angezeigt wird. Eine Suchmaschine kann das nicht. Für sie ist es eine Zahl wie jede andere. Daher ist sie für eine weitergehende Interpretation auf zusätzliche Informationen angewiesen. Dies kann man mit zusätzlichen Tags erreichen, wie sie Frameworks wie Microdata oder RDF (Rich Data Format) zur Verfügung stellen. Die Zahl wird im Quelltext also angereichert um die Information, dass es sich um einen Preis handelt, dieser in Euro angegeben wurde und dass die Verkaufssteuer (MwSt.) schon enthalten ist, es sich also um eine Bruttopreisangabe handelt. Werden jetzt der Produktbezeichnung und anderen beschreibenden Eigenschaften noch solche semantischen Tags mitgegeben, kann eine Maschine den wesentlichen Inhalt extrahieren und vor allem strukturiert im Index ablegen. Im strengen Sinne „versteht" sie damit natürlich immer noch nicht, worum es hier genau geht – aber sie ist in der Lage, über Querverbindungen Dinge besser zu vergleichen und am Ende auch auffindbar zu machen.

Die Vision der Forscher von einem semantischen Web konnte leider bis heute nicht so richtig verwirklicht werden. Im Wesentlichen liegt das wohl daran, dass die Menschen die Möglichkeiten solcher Zusatzkennzeichnungen in der breiten Masse noch immer igno-

DAS PROBLEM MIT DER SPRACHE:

Semantik sprachübergreifend zu erfassen bzw. zu interpretieren, ist gar nicht so leicht, wie man gemeinhin glauben möchte. Es genügt in der Regel nicht, einfach nur Begriffe zu übersetzen, weil sich die die mentalen Modelle oftmals unterscheiden. „I go by car or by foot" als Beispiel zeigt, dass die englische Sprache zum Teil für aktive Bewegungen oft schlicht ein „go" verwendet. Im Deutschen würde man aber niemals „ich gehe mit dem Auto" sagen. Während man in Italien streng übersetzt die Zähne „wäscht", „putzt" man sie bei uns. Zähne zu waschen ruft bei uns ein komisch seifiges Gefühl im Mund hervor, ein Italiener hätte wohl eher das Bild einer Putzfrau im Kopf, wenn er unseren deutschen Satz übersetzen würde. Im Russischen fügt man einem Verb oft noch die Information hinzu, ob man etwas nur einmal oder häufiger macht. Wir spielen ein Instrument, Tennis oder „Verstecken" – was dem Italiener wiederum komisch vorkommt. Er hätte im Kopf, dass wir mit einer Geige im Sandkasten spielen. Der Franzose spielt „von" einem Instrument und „an" einem Sport. Es genügt also nicht, nur Bedeutungen von Wörtern zu übersetzen, sondern es müssen auch die mit Wörtern oft unterschiedlich verknüpften mentalen Modelle im Kopf der Sprechenden mit berücksichtig werden. Dies mag auch der Grund sein, warum sich beobachten lässt, dass sprachbasierte Filter bei Suchmaschinen in anderen Sprachen oft zeitverzögert eingesetzt werden. Sätze wie „Ihre roten Haare lockten sich, aber nicht mich" sind für Menschen mit Deutsch als Muttersprache problemlos zu verstehen – für Maschinen stellen sie eine extrem hohe Herausforderung dar!

„Semantisches Taggen durch die breite Masse scheint eine Illusion zu sein."

rieren. In Zeiten, in denen das Metatag „Keywords" immer noch fleißig und mit viel Aufwand befüllt wird, um sich einen vermeintlichen Vorteil beim Ranking in Suchmaschinen zu verschaffen, darf man wohl nicht auf ein breites Verständnis solcher semantischer Tags hoffen. Und selbst wenn: Würden Webmaster überall auf der Welt Elemente auf Webseiten wirklich einheitlich genug vertaggen? Wahrscheinlich nicht. Darauf wollte und konnte man bei Google wohl nicht warten und überlegte sich schon vor vielen Jahren, wie man Inhalte denn auch ohne solche Tags besser verstehen könnte.

Bei der Inhaltserkennung ist man aber nicht nur auf Text angewiesen. Auch Bilder oder Videos, die ja auch nichts anderes sind als eine schnelle Folge von Bildern, können dazu beitragen. Im Lauf der Zeit konnte man beobachten, wie die Bildersuche bei Google immer besser wurde. Nur ist hier nicht unmittelbar klar, aufgrund welcher Informationen Google eine blaue Rose identifiziert und ins Suchergebnis einreiht. Die Bilder befinden sich ja auf Webseiten und sind dort in der Regel in einen textuellen Bezug eingebunden. Mittlerweile können Inhalte auf Bildern aber auch ohne diese Informationen recht zuverlässig erkannt werden.

Das kann man selbst recht einfach testen. Wer z. B. seine Smartphone-Bilder mit seinem G+-Account synchronisieren lässt oder auch eigene Fotos auf Google Drive ablegt, kann in G+ eine eigene Bildersuche verwenden. Diese Bilder kommen nicht von Webseiten und sind, sofern man sie selbst gemacht hat, noch nicht mit erklärendem Text versehen. Abbildung 4 und Abbildung 5 zeigen so eine Suche. Die dort eingegebenen Suchbegriffe sind selbst gemacht und noch nicht im Web aufgetaucht. Es existieren keine Exif- oder ähnliche Daten, aus welchen eine Maschine Rückschlüsse auf die Inhalte ziehen könnte. Man kann hier zweifelsfrei sehen, dass Google die Inhalte „erkennt" und sie passenden Suchworten zuordnen kann. Die Trefferqualität ist mittlerweile erstaunlich gut.

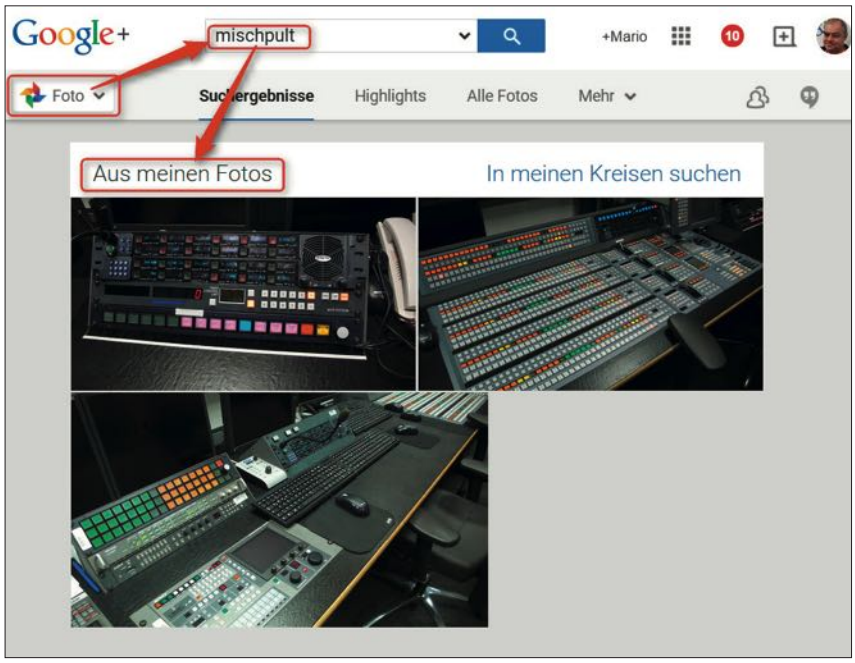


Abb. 4: Google kann auch eigene Fotos im eigenen G+-Account ohne jede begleitende Bildinformation gut zuordnen und erkennt zweifelsfrei den Inhalt

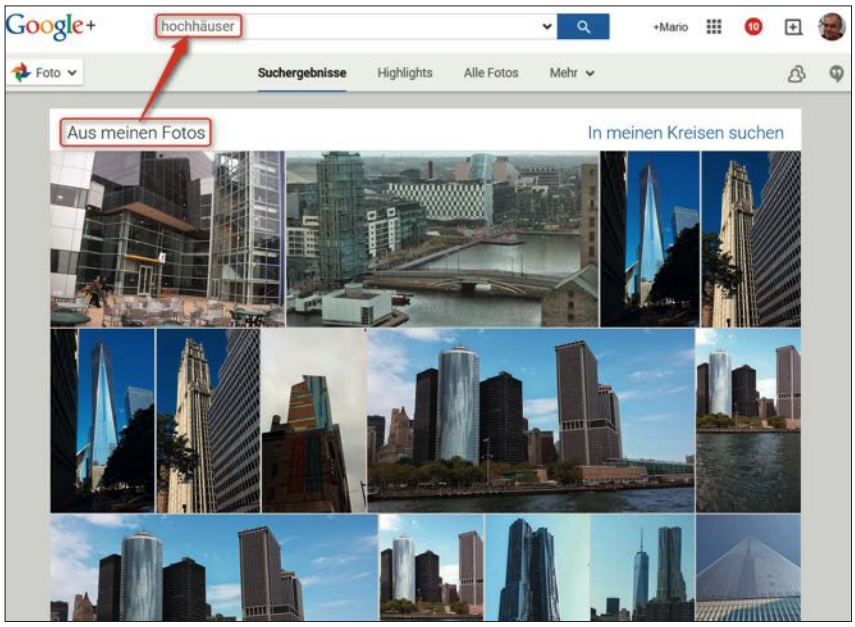


Abb. 5: Was ist auf einem selbst gemachten Foto zu sehen? Die Inhaltserkennung bei G+ funktioniert mit noch kleinen gelegentlichen Ausrutschern bestens!

Auch das Ziehen von Bildern von der eigenen Festplatte auf den Suchschlitz in der Bildersuche bei Google bringt qualitativ gute Ergebnisse. Zum Teil tauchen sogar beschreibende Worte als Vermutung auf, wie Abbildung 6 zeigt.

„Object detection, classification and labeling"

Im Research Blog von Google (*http://einfach.st/grbp* und *http://einfach.st/grbp2*) kann jeder nachlesen, wie Inhaltserkennung, Klassifizierung und das „Labeling" von Bildern funktionieren und wie weit sie fortgeschritten sind. Erkenn- und separierbare Objekte (auch Teile davon) werden mittlerweile relativ problemlos erkannt, klassifi-

Abb. 6: Vermutung für dieses Bild? Katze im Bett!

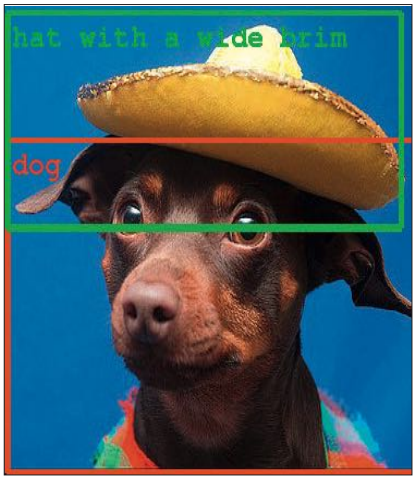


Abb. 7: Objekterkennung und Klassifizierung *(Quelle: Google, http://einfach.st/grbp2)*

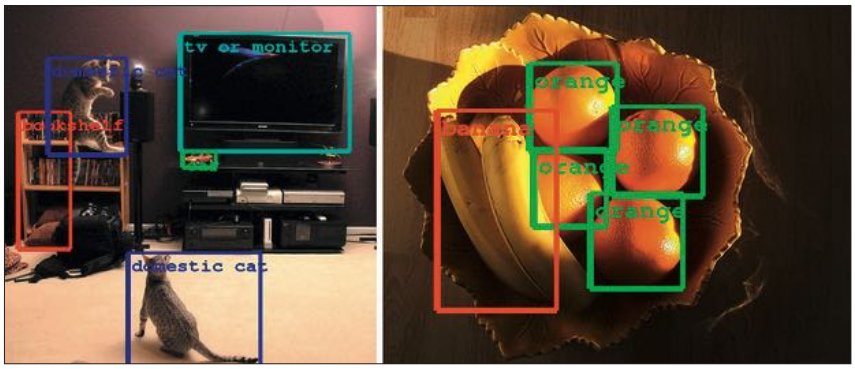


Abb. 8: Auch komplexere Bilder lassen sich nach Objekten zerlegen *(Quelle: Google Reseach Blog)*

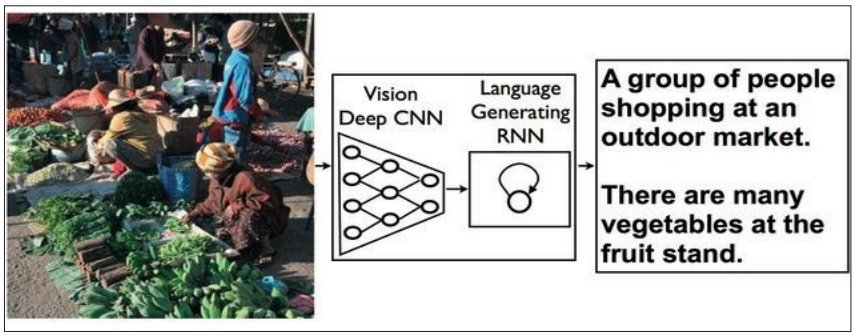


Abb. 9: Forschungsergebnisse zeigen, wie präzise Maschinen heute bereits Bilder beschreiben können *(Quelle: Google, http://einfach.st/grbp)*

ziert und dann mit natürlicher Sprache beschrieben. Abbildung 7 und Abbildung 8 zeigen dies exemplarisch. Das verwendete neuronale Netzwerk (CNN – Convolutional Neural Network) wird dabei ständig weiter trainiert und lernt durch die enorme Rechenpower und das praktisch unendliche Bildmaterial im Google-Speicher überproportional schnell dazu. Nimmt man die Testmöglichkeiten noch mit dazu, die ja prinzipiell über die Google-Suche jederzeit als A/B-Test mit menschlichen Suchenden machbar sind bzw. wären, offenbart sich das wahre Potenzial – und warum das wohl kein anderes Unternehmen derartig schnell und präzise verbessern oder gar nachbauen kann. Die Masse der Daten und die Rechenpower stehen in diesem Umfang niemandem sonst auch nur ansatzweise zur Verfügung.

Natürlich ist die verbale Beschreibung von Bildern durch Algorithmen noch nicht perfekt und es gibt immer wieder vereinzelt auch Fehlinterpretationen. Deren Minimierung ist aber nur eine Frage der Zeit, ebenso wie die Verwendung von Bildern auf Webseiten zur Signalerkennung für das Ranking. Wenn es in einem Beitrag um Klavier- oder Saxofonspielen geht bzw. die textlichen Signale (OnPage) dies vermuten lassen, aber die verwendeten Bilder erkennbar andere Themen haben (Abbildung 10), wäre ein Punktabzug für das Ranking durchaus denkbar. Deswegen muss eine Seite nicht gleich als Spamversuch per Flag gebrandmarkt werden oder komplett abstürzen. Aber wenn es noch weitere ähnlich gute Beiträge zum gleichen Topic gibt, bei denen die begleitenden Bilder besser passen, dann ist wohl für jeden vernünftigen Menschen einsich-

Ein maschineller Albtraumsatz:
„Ihre roten Haare lockten sich – aber nicht mich."

tig, dass diese im Ranking weiter oben stehen sollten. Ob Google mittlerweile solche Techniken schon einsetzt, um ein harmonisches Verhältnis von Text und Visualisierung zu messen und ggf. mit zu bewerten, ist nicht bekannt. Technisch wäre es durchaus machbar.

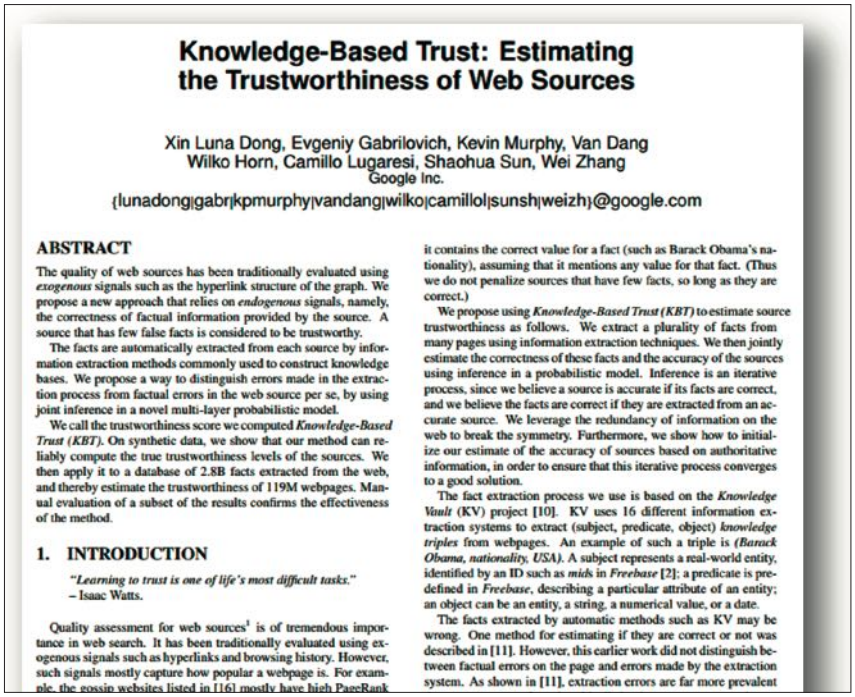
Aber die Algorithmen zielen nicht nur auf die Erkennung von Bildinhalten ab, sondern auch darauf, ob die Aussagen eines Textes vertrauenswürdig sind bzw. der allgemeinen Auffassung entsprechen oder nicht. Fälschlicherweise wurde mitunter auch in renommierten Zeitschriften wie der FAZ oder der Welt Google unterstellt, man beanspruche dort für sich, die „Wahrheit" zu kennen bzw. danach zu entscheiden. Hier wurde u. a. falsch verstanden, dass es um Wahrscheinlichkeiten geht und wie gesagt um den „Common Sense". Absolute Wahrheiten kennt ja auch der Mensch nicht, da sich Wissen und auch die Missinterpretation von Wissen ständig fortentwickeln. War man kurz vor den Gebrüdern Wright noch der Meinung, Maschinen, die schwerer als Luft sind, könnten niemals fliegen (Common Sense damals), weiß man auch dies heute besser. Bei Google erkannte man wie erwähnt schon vor Längerem, dass man mit dem Auswerten von Backlinks als gewichtigem Rankingfaktor wohl auf Dauer nicht weiterkommt. Und genau deswegen ist man auf der Suche nach alternativen, vertrauenswürdigen Signalen.

Abb. 10: Die Überschrift sagt Klavier, das Bild eindeutig Gitarre – solche Unstimmigkeiten können Maschinen mittlerweile erkennen

Abb. 11: Ergebnis einer „refinded search" durch einen Nutzerdialog *(Screenshot Google)*

Keywords: „not provided" – und das ist auch gut so?

„Lügen haben kurze Beine", das wissen die Kinder bereits in der Grundschule, und genau auf deren Sprachniveau befindet sich der Google-Algorithmus. Eine Suchmaschine, deren Rankingfaktoren vermeintlich auf Text und Backlinks basieren, wie passt dies in den Zusammenhang mit dem Sprachniveau? Relativ einfach, denn mittels maschinellen Lernens verstehen Suchmaschinen den Inhalt einer Nachricht. Es müssen somit nicht exakt die Keywords im Dokument enthalten sein, damit eine Webseite das beste Ergebnis zur Befriedigung der Suchintention sein kann. Dies ist aus Nutzersicht ein enormer Sprung in Richtung des Zieles „zeigt dem Nutzer das beste Ergebnis zu seiner Suche" und eine weitere Hürde für Suchmaschinenoptimierer. Es geht nicht mehr darum, das Keyword möglichst genau zu treffen, sondern die Suchintention inhaltlich zu verstehen und entsprechende Antworten aus einer holistischen Betrachtung des Themas zu liefern. Der Besuch der Webseite mit der Suchphrase „wie alt ist er" kann schlussendlich das Ergebnis einer Konversation mit Google sein, welches mit „O. k., Google, wie groß ist Dirk Nowitzki" begann (siehe Abbildung 11). Im „mobilen Zeitalter" wird diese Form des Suchens (Sprechen anstelle von Tippen) weiter zunehmen. Die Information „not provided" in den



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, and thereby estimate the trustworthiness of 119M 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[1] 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

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 *m/d* in *Freebase* [2]; a predicate is predefined 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

Abb. 12: Wissensbasiertes Vertrauen für Webseiten – ein Forschungspaper von Google, das es in sich hat! *(http://einfach.st/arxiv1)*

Abb. 13: Triple über Person *(Screenshot Google)*

Keywordberichten von Webcontrollingsystemen wie Google Analytics dürfte damit niemanden mehr zur Weißglut bringen, da auf „wie alt ist er" zukünftig nicht optimiert werden würde. Darüber hinaus würde die Analyse dieser Keywords ohne den vollständigen Zusammenhang keinen Mehrwert mehr bieten – die Customer Journey beginnt mit einem Dialog zwischen Google und dem Suchenden und endet mit einer kontextbehafteten Suchphrase auf dem Content der Webseitenbetreiber.

Die spannende Frage ist: „Woher weiß Google, dass Dirk aktuell 36 Jahre alt ist?" Halbwahrheiten bzw. Falschinformationen sollten aussortiert werden. Obwohl viele Ratgeberseiten bzw. Seiten über Prominente – von Google als „gossip"(engl. für Tratsch oder Geschwätz) bezeichnet – meist sehr viele Backlinks besitzen, kann das Ergebnis durchaus falsch sein und sollte damit nicht in den Toppositionen ranken. In einem wissenschaftlichen Bericht (Knowledge-Based Trust; siehe das PDF unter *http://einfach.st/arxiv1*) wurde von acht Google-Mitarbeitern ein Verfahren vorgestellt, welches eine Seite losgelöst von Links bewertet. Es geht darum, ein Ranking durch die Korrektheit von Informationen aufzubauen und damit unabhängig von externen, ggf. manipulativen Links zu sein. Damit ein Wert für den sog. Knowledge-Based Trust (KBT) auf Seiten- bzw. Domainebene ermittelt werden kann, bedarf es zweier Faktoren: erstens einer Vielzahl von Fakten, die aus den Webseiten extrahiert werden, und zweitens einer Bewertung dieser strukturierten Informationen, um zum einen Extraktionsfehler zu minimieren und zum anderen die Wahrscheinlichkeit eines wahren, d. h. richtigen Wertes der Information zu bestimmen.

Wie lernen Maschinen? Mit Entitäten und deren Beziehungen untereinander!

Eine Suchmaschine speichert Informationen in Entitäten, d. h. in eindeutigen Objekten mit ihren spezifischen Merkmalen. Dabei ist es unerheblich, ob das Objekt abstrakter, materieller oder immaterieller Art ist (Person, Gegenstand, Zustand etc.). Keywords beschreiben diese Entität und definieren deren Ausprägung. Dirk Nowitzki, als Prominenter eine facettenreiche Entität, ist zum Beispiel „Würzburger Basketballspieler", „NBA Profi", „reicher Sportler", „Teampartner von Monta Ellis", „Sportler des Jahres" sowie ein „großer Mensch". Die Auflistung zeigt, dass ein Keyword allein die Person Nowitzki sehr unzureichend beschreibt und deshalb mehrere unterschiedliche Keywords letztendlich die gleiche Person a) beschreiben und b) jeweils aus einem speziellen Blickpunkt betrachten. Ein weiterer negativer Aspekt von Keywords ist ihre Doppeldeutigkeit. Für Suchmaschinen sind sie für sich allein sehr unpräzise, da sie in unterschiedlichen Zusammenhängen benutzt

werden können. Mit dem Keyword „Jaguar“ kann eine Raubkatze oder ein Auto klassifiziert werden. Damit bedarf es der eindeutigen Zuordnung zum Themengebiet bzw. der Entität weiterer Bezug nehmender Wörter im unmittelbaren Zusammenhang („Savanne“ bzw. „Hubraum“). Dies ist der Hintergrund, weshalb bei der WDF*IDF-Analyse die beweisführenden Terme von großem Nutzen sind.

Tripel – die Mutter aller Logik

Der Lernprozess fußt auf der Bereitstellung (Extraktion) sogenannter Tripel. Ein Tripel besteht aus der Folge Subjekt – Prädikat – Objekt. Das Tripel liefert die Information: Wer oder was (Subjekt) hat welche Eigenschaft (Prädikat) in welcher Ausprägung (Objekt)? Die Tripel beschreiben jegliche Art von Entitäten: Ein Computer hat einen Arbeitsspeicher von 5 GB, ein Nike-Free-Turnschuh hat einen Preis von 100 €, eine Person kann hinsichtlich ihrer charakteristischen Merkmale beschrieben werden etc. (siehe Abbildung 13 und auch den englischen Blogbeitrag unter *http://einfach.st/seoskeptic*).

Das Ziel einer Suchmaschine ist, möglichst viele Tripel zu einer Entität zu besitzen. Die entsprechenden Identifikatoren, bspw. eine Domain mit der entsprechenden Deep-URL, liefern den exakten Ort dieser Wissensgenerierung. Die Maschine kann demnach die gesammelten Informationen diverser Quellen vereinen und verifizieren, da zu jedem Tripel ein eindeutiger Identifikator zur Verfügung steht. Sollten 100 Tripels von 100 Websites die Information „Dirk Nowitzki hat Größe 2,13 m“ liefern, könnte die Information als valide angesehen werden. Sofern eine Website die Information „2,10 m“ liefert, könnte diese Quelle, dieser Identifikator, bei einer Häufung fehlerhafter Informationen als nicht vertrauenswürdig eingestuft werden.

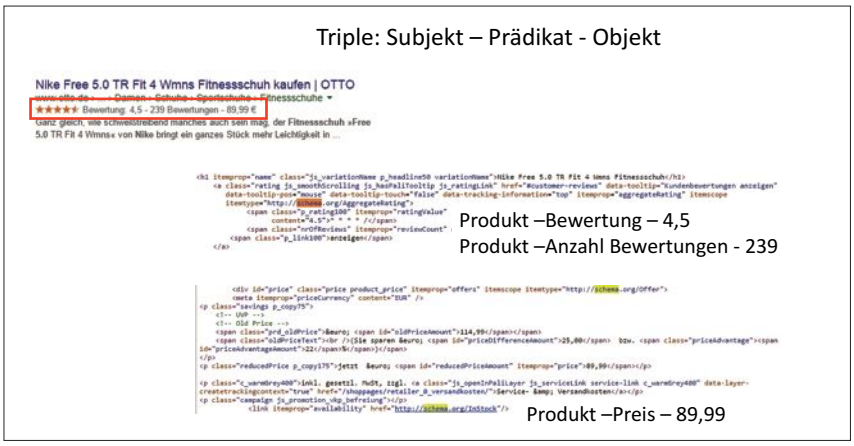


Abb. 14: Schema.org liefert Tripel

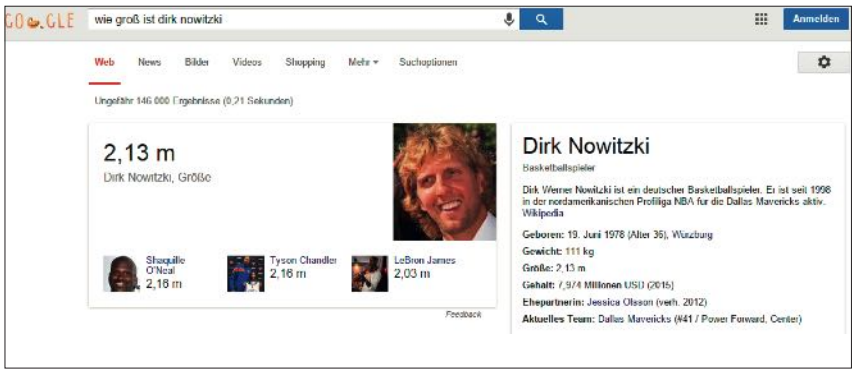


Abb. 15: Tripel als Faktenlieferant für verbundene Fragen *(Screenshot Google)*

Rich Snippets – Extrakt aus Semantik

Der Einsatz von Schema.org bzw. dem Data-Highlighter in den Google-Webmaster-Tools führt im Idealfall zur Ausspielung von Rich Snippets (siehe Abbildung 14). Durch die Taxonomie, die stringenten Vorgaben hinsichtlich des Formates, lernt die Suchmaschine. Sie versteht, dass dieses Produkt einen Preis von 89,99 €, 4,5 von 5 Sternen, 239 Bewertungen, einen Namen etc. hat.

Tipp: Durch die Implementierung dieser Mikroformate in die eigene Website interpretiert die Suchmaschine den Inhalt besser im Vergleich zu Crawling mit nachgelagerten Extraktoren. Dies ist neben der CTR-freundlichen Rich-Snippets-Darstellung ein unterschätzter Faktor. Der Einsatz des Open-Graph-Protokolls ist ebenfalls zu prüfen, damit in sozialen Netzwerken geteilte Informationen wunschgemäß dargestellt und maschinell interpretierbar werden.

SEO – Optimierung/Verifizierung der Beziehung

Ein weiterer Vorteil der Extraktion von Tripeln ist die damit verbundene Relation zur Entität. Das heißt, die Informationen der Tripel betrachten die Entität aus verschiedenen Blickwinkeln und ermöglichen demnach, eine Antwort auf eine bis dato noch nicht gestellte Frage zu liefern. Die Frage nach der Größe von Nowitzki bietet gleichzeitig Gewicht, Gehalt, Ehepartnerin, Team, Trikotnummer etc. Darüber hinaus werden auch die Körpergrößen weiterer Spieler angezeigt, von denen die Suchmaschine ähnliche Tripel besitzt sowie weiß (u. a. aus Suchketten, Browserverlauf), dass nach deren Körpergröße im Zusammenhang mit Nowitzki ebenfalls gesucht wird (Abbildung 15).

Wer hört Helene Fischer?

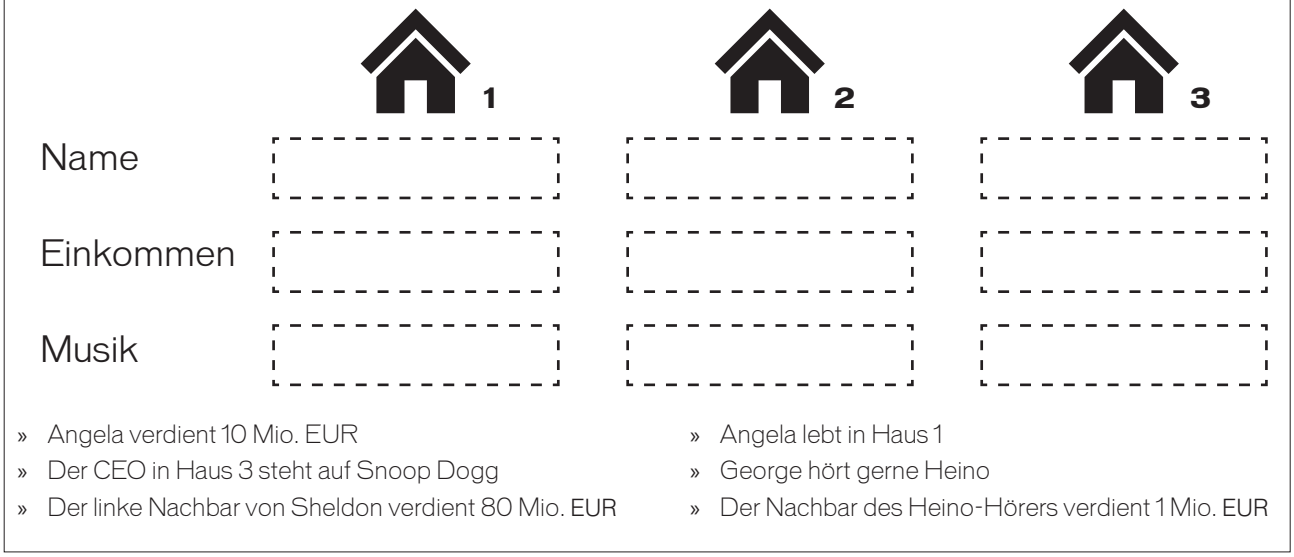


Abb. 16: Wie kann man von Fakten ausgehend fehlende Informationen ermitteln? Probieren Sie es einfach einmal aus! (Auflösung S. 98)

Abb. 17: Auszug aus der Datenbank dppedia.org über die Entität „Berlin“

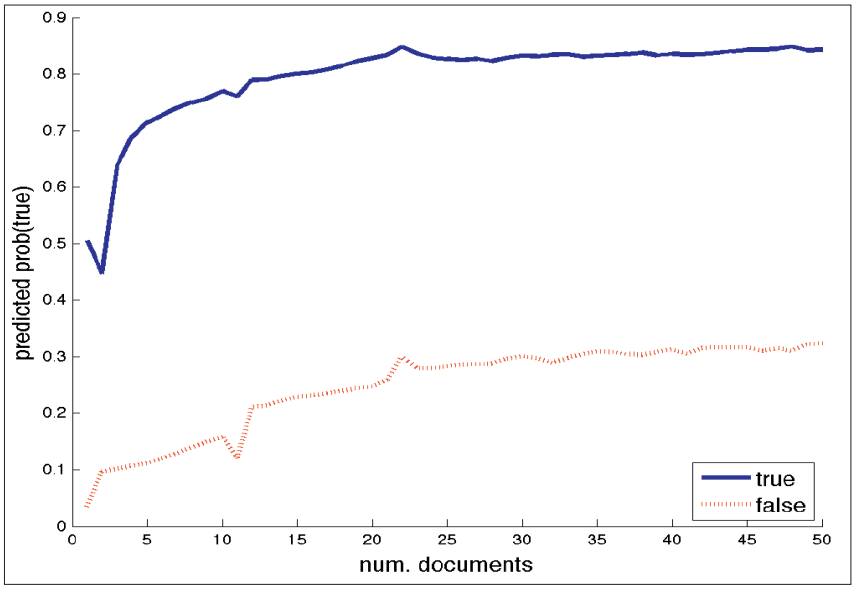


Abb. 18: Prognosegenauigkeit eines Tripels in Abhängigkeit von der Anzahl Dokumente, welche das Tripel enthalten (Quelle: Dong, Xin Luna et al.; Knowledge Vault; *http://einfach.st/kvault*)

Tripel – Basis für Verbindung und Logik

Websites können durch Entitäten bzw. Tripel in Verbindung zueinander gebracht werden. Die Vielzahl an Attributen, Eigenschaften und Beziehungen, welche mit einer Marke verknüpft werden, führt neben dem hohen Suchvolumen nach dem Markennamen dazu, dass ein bevorzugtes Ranking von Marken stattfindet, und nicht dadurch, dass es sich um einen Markennamen per se handelt. Durch Tripel werden die Objekte immer umfassender beschrieben, bestehendes Wissen/Informationen können durch weitere Quellen verifiziert werden. Die Maschine erhält dadurch ein faktenbasiertes Bild eines Objektes. Vorhandene Informationen können mittels Logik kombiniert werden, wodurch u. a. auch fehlende Informationen erschlossen werden. Welche Person in welchem Haus hört wohl Helene Fischer (Abbildung 16)? Der Mensch kann eine solche Aufgabe binnen Minuten lösen, ein Computer binnen Mikrosekunden.

Semantische Datenbanken – strukturiertes Wissen

Suchmaschinen können auf diverse semantisch strukturierte Datenbanken zugreifen und liefern ihrerseits die

gewonnenen und validierten Ergebnisse u. a. im Knowledge Graph aus. DBpedia, ein Gemeinschaftsprojekt von Universitäten und Instituten, enthält mehr als drei Milliarden Tripels, welche aus Wikipedia-Dokumenten in diversen Sprachen extrahiert wurden (siehe *http://blog.dbpedia.org*). Mittels der Abfragesprache SPARQL können gezielt Informationen aus Wikipedia ausgelesen, verarbeitet und in Relation zueinander gebracht werden. Weitere Datenbanken, welche ebenfalls sehr umfassende Themenbeschreibungen und Fakten liefern, sind beispielsweise *freebase.com* sowie YAGO vom Max-Planck-Institut für Informatik.

Tipp: Als Grundlage für die Texterstellung ist diese Datenbank durch die Betrachtung verknüpfter, inhaltlich verwandter Themenbereiche von großem Nutzen. Abbildung 17 zeigt beispielhaft einen Ausschnitt aus dbpedia über 5.000 aufbereitete (strukturierte!) Informationen über die Entität Berlin. In Kombination mit W-Fragen (wer, was, wie, warum, wo, wieso, ...) kann beim Texten eine holistische Sicht des zu beschreibenden Objektes (bspw. Festival in Berlin) eingenommen werden. Der Texter und später der Leser werden dadurch geleitet und umfassend informiert. Dabei sollte man natürlich immer auch im Kopf behalten, dass bei großen Textmengen ein Scannen im Sinne eines schnellen Überfliegens des Textes durch den Leser ermöglich wird (siehe Niesen „how users read the web"; *http://einfach.st/howread*).

Knowledge Vault – das Schatzkästchen von Google

Die bestehenden Möglichkeiten der textuellen Extraktion, insbesondere aus Wikipedia bzw. YAGO, wurden in den letzten Jahren stark erweitert. Vor allem Universitäten befassen sich mit dem Bereich Textmining/Natural Language Processing (NLP) und stellen umfassende Publikationen bzw. Programme

	Value	E_1	E_2	E_3	E_4	E_5
W_1	USA	USA	USA	USA	USA	*Kenya*
W_2	USA	USA	USA	USA	*N.Amer.*	
W_3	USA	USA		USA	*N. Amer.*	
W_4	USA	USA		USA	*Kenya*	
W_5	Kenya	Kenya	Kenya	Kenya	Kenya	Kenya
W_6	Kenya	Kenya		Kenya	*USA*	
W_7	-			*Kenya*		*Kenya*
W_8	-					*Kenya*

Abb. 19: Extraktion der Nationalität von Barack Obama von acht Webseiten *(Quelle: Dong, Xin Luna et al.; Knowledge Vault; a. a. O.)*

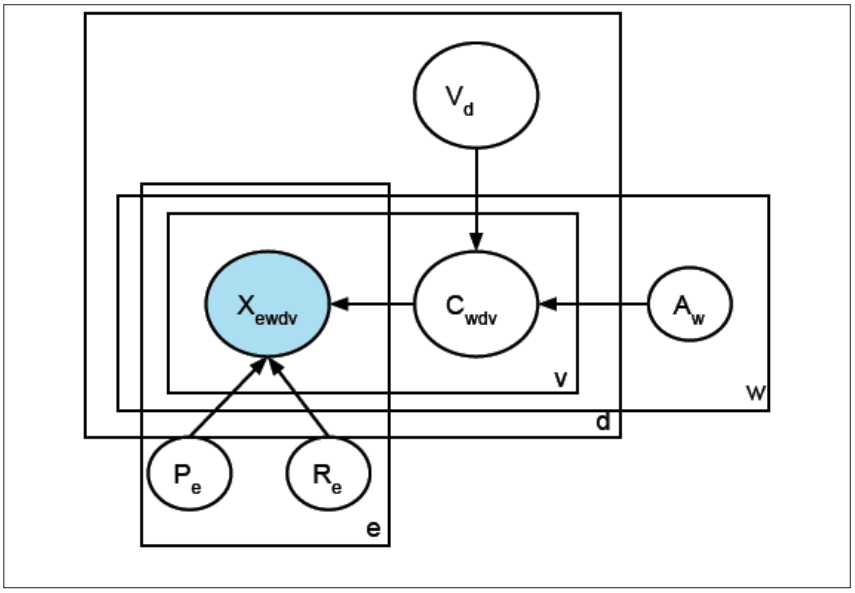


Abb. 20: Mehrschichtmodell zur Validierung des Tripels (Quelle: Dong, Xin Luna et al.; a. a. O.)

zur Verfügung (Stanford: (*http://nlp.stanford.edu/*); Sheffield: Open-Source-Software zum Textmining GATE (*https://gate.ac.uk/*)). Google-Mitarbeiter erläutern in einer Publikation (Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion – download unter *http://einfach.st/kvault*) den Prozess der Generierung von Tripels mittels Extraktoren.

Dabei greifen sie neben Mikroformaten und HTML-Strukturanalysen insbesondere auf die Techniken des NLP zurück. Zum Einsatz kommt u. a. das Verfahren „Named Entity Recognition", welches in einem Text gezielt nach Elementen (bspw. Personennamen) sucht und diese dann entsprechenden vordefinierten Klassen (bspw. Name) zuordnet. Im Folgenden werden die extrahierten Tripel mit statistischen Verfahren wie neuronalen Netzen kombiniert, um verborgene Verbindungen zu entdecken und mit Wahrscheinlichkeiten hinsichtlich der Güte zu belegen. Das Ergebnis dieser Verfahren ist eine Wissensdatenbank, gefüllt durch 16 unterschiedliche Extraktionssysteme, welche den Umfang bisheriger Datenbanken um den Faktor 38 übertrifft sowie eine detaillierte Einschätzung über den Unsicherheitsfaktor (Irrtum) beinhaltet. Generell gilt, dass eine steigende Anzahl an Extraktionssystemen schnell zu einer hohen Prognosesicherheit hinsichtlich des Wahrheitsgehaltes der Tripel führt. Der Einsatz von vier Extraktionssystemen liefert bereits eine Zuverlässigkeit von ca. 90 %. Gleichzeitig steigt die Prognosegenauigkeit der Tripel mit der Anzahl unterschiedlicher Quellen, welche die Information auf der Website in Tripel bereitstellen (siehe Abbildung 18).

Wie durch Tripelextraktion neues Wissen entstehen kann

-stark vereinfachte Prinzipdarstellung-

	USA	Kenya	N.Amer.
W_1	1	0	-
W_2	1	-	0
W_3	1	-	0
W_4	1	0	-
W_5	-	1	-
W_6	0	1	-
W_7	-	.07	-
W_8	-	0	-
$p(V_d\|\hat{C}_d)$	.995	.004	0

Abb. 21: Extraktionsgüte sowie Einschätzung des wahren Wertes *(Quelle: Dong, Xin Luna et al.; a. a. O.)*

Knowledge-Based Trust – wer lügt, der fliegt

Die bereits angesprochene Veröffentlichung zum Knowledge-Based Trust bedient sich der Tripel und bewertet mittels statistischer Methoden deren Wahrheitsgehalt. Websites werden nicht bestraft, wenn nur wenige extrahierbare Fakten bereitgestellt werden, solange diese korrekt sind. Der Prozess der Extraktion ist fehleranfällig, wodurch es eines mehrschichtigen Verifizierungsprozesses unter Berücksichtigung der notwendigen Granularität von Webseiten innerhalb einer Website bedarf. Die größte Herausforderung in der Bewertung ist die Anzahl zur Verfügung stehender Tripel – entweder es sind zu wenige verfügbar oder es stehen zu viele zur Verfügung (rechnerischer Engpass), um die Entität der Seite zu bestimmen. Die Autoren des Forschungsberichtes erwähnen übrigens, dass auf über einer Milliarde Webseiten wegen der Contentarmut noch nicht einmal ein einziges Tripel extrahiert werden konnte.

Letztendlich steht am Ende ein Trust-Wert für eine Seite bzw. eine Domain zur Verfügung, welcher diese Limitationen berücksichtigt. Der Prozess folgt den nachstehenden Prinzipien und wird in der wissenschaftlichen Publikation anhand der Frage „Welche Nationalität hat Barack Obama?“ exemplarisch dargestellt (siehe Abbildung 19).

Abbildung 19 zeigt exemplarisch, wie acht verschiedene Webseiten (W1-W8) dahin gehend geprüft werden, ob diese die Nationalität des amerikanischen Präsidenten Barack Obama kommunizieren und mit welchem Inhalt (Spalte „Value“). Die ersten vier Webseiten liefern das richtige Ergebnis, W5-W6 kommunizieren ein inhaltlich falsches Ergebnis, W7 und W8 stellen keine Informationen zur Nationalität zur Verfügung. E1 bis E5 sind Extraktoren, die aus der Webseite das Tripel (A, Nationalität, B) ermitteln. Extraktor E1 extrahiert alle Informationen korrekt, d. h., die kommunizierte Information der Webseite („Obama hat Nationalität USA bzw. Kenya“) ist identisch mit dem Inhalt des Tripels (Obama, Nationalität, USA/Kenya). Extraktor E2 extrahiert einzelne Tripel korrekt, jedoch nicht bei allen Webseiten. E3 extrahiert zwar die Tripel, jedoch auch für Webseite 7, obwohl diese Seite keine Aussage zur Nationalität von Obama trifft. Die Extraktoren E4 und E5 haben eine geringe Qualität durch fehlende bzw. falsche Werte.

In Summe wird die Nationalität aus zwölf Quellen (Paar aus Extraktor/Webseite) mit Kenya ermittelt und zwölfmal mit USA, d. h., es könnte angenommen werden, dass beide gleich wahrscheinlich sind. Eine nähere Betrachtung liefert zwei Ergebnisse: Es kann ein Fehler in den Extraktoren vorliegen (bspw. dass bei einer Website der Präsident einer Firma extrahiert wurde und nicht der Präsident der USA) und es kann ein Fehler bei der Quelle vorliegen (suboptimal recherchierter Content). Die Indikation, ob der Extraktor (e) das Triple (d mit Wert v) von der Website (w) liefert, geschieht durch ein iteratives Verfahren über mehrere Ebenen (Abbildung 20). Es werden die Verlässlichkeit/Qualität des Inhaltes der Website (AW), die Präzision der Extraktoren (Pe, Re) sowie eine Beurteilung des wahren Werts des Datenelements (Vd) berücksichtigt bzw. vorgenommen. Sofern von einer Webseite eine zu geringe Anzahl Tripel extrahiert wird, könnten lt. Google die Daten mit anderen Seiten der Site aggregiert betrachtet werden.

Dieses Verfahren liefert eine Aussage über die Güte der Extraktoren je Webseite sowie der Einschätzung über den wahren Wert des Elements „Barack Obama, Nationalität“. Mit sehr hoher Wahrscheinlichkeit liefert der Rechenalgorithmus bzw. das statistsche Verfahren die USA als Nationalität von Obama.

Maschinelles Lernen – das Ende der „Updates“

Paart man die statistische Validierung der Extraktoren sowie der Einschätzung der Qualität der Webseite mit der oben erwähnten Logik des Helene-Fischer-Beispiels (Kombination), so ist die Generierung einer Vielzahl weiterer Tripel/Fakten möglich. Die Maschine lernt dadurch exponentiell schneller und deutlich spezifischer. Die Güte der Suchergebnisse wird weiter zunehmen, eine bessere semantische Zuordnung der Datenquellen zur Suchintention vorgenommen und Konversation zwischen Mensch und Maschine effizienter

ermöglicht werden. Klassische Updates könnten durch permanentes Lernen ersetzt und Alternativen zu aktuellen Rankingfaktoren wie bspw. Backlinks ermöglicht werden.

Knowledge-Based Trust als Alternative zu Links?

In einem Test wurde bei 2.000 zufällig ausgewählte Websites der Knowledge-Based Trust (KBT) mit dem Google PageRank (Maß für externe Verlinkung) verglichen. Das Ergebnis wurde anschließend manuell verifiziert. Das Resultat lässt auf eine hohe Zuverlässigkeit des Kriteriums KBT schließen (siehe Abbildung 22). Sehr guter Content „erntet" gute bzw. viele Links („Links folgen Qualität"), wohingegen optimierungswürdiger Content wenige Links nach sich zieht (hart gesagt „Shit in – Shit out"). Spannend sind die beiden diagonalen Quadranten. Eine manuelle Analyse der PageRank-starken Seiten (links oben in Abbildung 22) ergab, dass die Inhalte als nicht vertrauenswürdig einzustufen sind, aber durch die Popularität („gossip" bzw. Forum) viele Backlinks aufweisen. Hier könnte ein Ranking durch die Vielzahl an Backlinks positiv beeinflusst sein, obwohl die Inhalte nur bedingt korrekt sind (bekannte „Halbwahrheiten"). Das andere Extrem (rechts unten in Abbildung 22) sind sehr gut recherchierte Nischenseiten mit hohem Problemlösungspotenzial („liefern sehr gute Antworten zu einer Suchanfrage"), welche jedoch eine geringe Verlinkung haben. Dies könnte den geringen Suchanfragen der Nische geschuldet sein oder aber auch dem geringen Alter der Domain. Offensichtlich könnte man Google hier Handlungsbedarf attestieren.

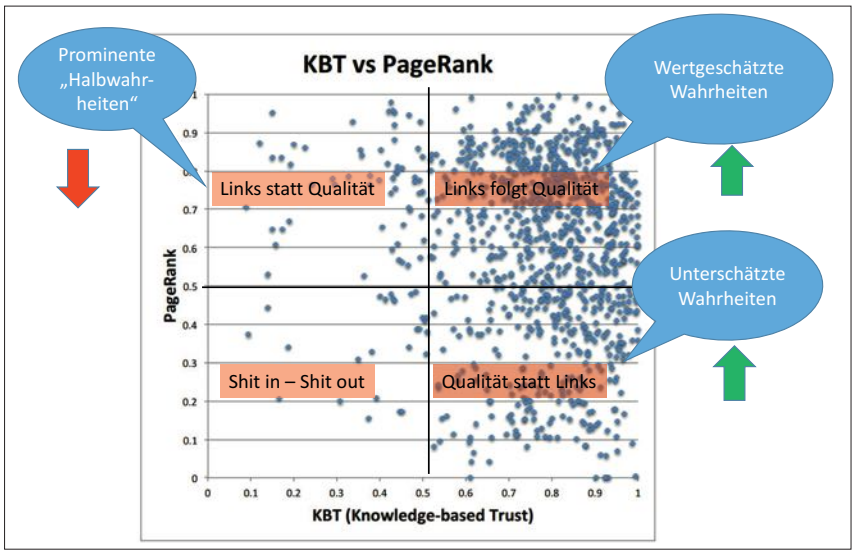


Abb. 22: Vergleich Knowledge-Based Trust und PageRank *(Quelle: Quelle: Dong, Xin Luna et al.; a. a. O. ergänzt mit Anmerkungen)*

INFO:

Die russische Suchmaschine Yandex kündigte Ende 2013 an, bei kommerziellen Suchergebnissen Backlinks bei der Bewertung außen vor zu lassen. Für Linkkauf werden dort von SEOs um die 200 Mio. Euro pro Jahr ausgegeben. Die Tendenz ist ist zwar fallend, aber Webmaster geben noch immer in unvorstellbarem Maß Geld für eigentlich unnütze Maßnahmen aus. In einem Interview mit Marcus Tandler kündigte Alexander Sadovsky, Head of Web Search bei Yandex, an, dass Links nun wieder in die Wertung mit aufgenommen werden – allerdings mit einem durchaus auch stark negativen Effekt. So werde es einen deutlichen Malus für linkkaufende Sites geben. Man darf gespannt sein, wie schnell diese Maßnahme dem Linkkauf den Garaus macht. Nachzulesen ist das Interview unter *http://einfach.st/ylinks*.

Fazit

Suchmaschinen lernen derzeit explosionsartig die Semantik von Texten. Behshad Behzadi von Google bestätigte erst im März dieses Jahres, dass in der hauseigenen Knowlede-Base 40 Mrd. Fakten und 570 Mio. Entitäten gespeichert sind – und es würden täglich mehr.

Die logischen Verbindungen, die Menschen automatisch aus ihren bisherigen Erfahrungen ziehen, können durch die Generierung und Verknüpfung von Tripeln nachgebildet werden. Spannend und für die Wissensexplosion verantwortlich ist die (Re-)Kombination des bestehenden Wissens, das beim Menschen unter Transferwissen subsummiert wird. Durch die fortschreitende Rechenleistung und nahezu unbegrenzte Speicherkapazität scheint es tatsächlich nur eine Frage der Zeit zu sein, bis Maschinen eine valide Qualitätseinschätzung von Webseiten vornehmen können. Durch sprachspezifische Besonderheiten, d. h. Komplexität von Grammatik und Interpunktion, kann der Prozess für nicht-englische Sprachen länger dauern – jedoch nicht aufgehalten werden. Aus der SEO-Perspektive könnten Suchmaschinen den Wert externer Signale wie Links zukünftig immer mehr abschwächen und den Inhalt einer Webseite bzw. dessen Wahrheitsgehalt deutlicher honorieren. Dadurch kann durchaus eine Win-win-Situation für Google und den Nutzer entstehen: Manuell aufgebaute Links würden unattraktiver und qualitativer Content würde belohnt werden – daraus folgend könnte weiterer Antrieb für Webverantwortliche entstehen, einen wirklich herausragenden Content zu produzieren.

Und wenn man sich nun noch ins Gedächtnis ruft, dass mittlerweile bereits selbstfahrende Autos und sich stabil bewegende Roboter miteinander vernetzt werden und das „Gelernte" dann durch die Bilderkennung geschleust wird, dann wird unmittelbar klar, dass die Wissenssammlung einzelne Webseiten als Quelle bereits verlassen hat. Reale Objekte, deren Aussehen und Verhalten lassen sich nämlich ebenfalls „vertripeln" … ¶