OLAF KOPP

ENTITÄTEN-BASIERTE SUCHE: SO FUNKTIONIERT DER GOOGLE KNOWLEDGE GRAPH

DER AUTOR

Olaf Kopp ist Co-Founder und Head of SEO der Aufgesang GmbH. Er ist Moderator des Podcasts Content Kompass und Mitveranstalter des SEAcamps. Seine Kernthemen sind semantische SEO, Content-Marketing- und Online-Marketing-Strategien entlang der Customer-Journey.

Die Einführung des Knowledge Graph im Jahr 2012 und des Hummingbird-Algorithmus im Jahr 2013 waren neben Panda und Pinguin die einschneidendsten Veränderungen für die Suche seit Bestehen von Google. Durch den Knowledge Graph und Hummingbird geht es nicht mehr nur um Keywords und Dokumente, sondern mit wachsendem Einfluss um Entitäten und deren semantische Beziehung zueinander.

Dieser wirklich umfassende Beitrag öffnet die Tore zu einer bisher wenig erforschten und daher geheimnisvollen Welt des Knowledge Graph und die Bedeutung von Entitäten für die Google-Suche.

Foto: iStaNtsz_ill / gettyimages.de

Der Hummingbird-Algorithmus ist die aktuelle Ranking-Algorithmus-Version von Google, auf der alle Updates der letzten Jahre basieren. Hummingbird wurde 2013 zur semantischen Interpretation von Suchanfragen, Dokumenten, Domains, Apps … eingeführt.

Dabei stützte sich Hummingbird neben dem klassischen Suchindex auf den Knowledge Graph als zugrunde liegende Datenbank. Während im klassischen Suchindex Dokumente, Bilder, Videos … gespeichert werden, erfasst der Knowledge Graph Entitäten, Entitätstypen, deren Attribute und Beziehungen zueinander.

Zur Einführung des Hummingbird-Updates gab es eine Pressekonferenz. Hier einige Statements daraus:

Q&A. Hummingbird biggest revamp? Indeed, „it's not easy to build a new algorithm that's so good"

How's it different? People asking more complicated questions. So how keep results so relevant in light of these. Hummingbird impact all types of queries we get but far more effective on these long queries we get now.

Gave us an opportunity, hummigbird did, to take synoyms and knowledge graph and other things Google has been doing to understand meaning to rethink how we can use the power of all these things to combine meaning and predict how to match your query to the document in terms of what the query is really wanting and are the connections available in the doucuments and not just random coincidence that could be the case in early search engines.

Hummingbird ist im Nachhinein auch als Vorbereitung seitens Google auf die zunehmende Anzahl komplexer Fragestellungen via Voice Search über mobile Endgeräte und digitale Assistenten zu interpretieren. Weg von Keywords, hin zu Entitäten. Von der Suchmaschine zur Antwortmaschine.

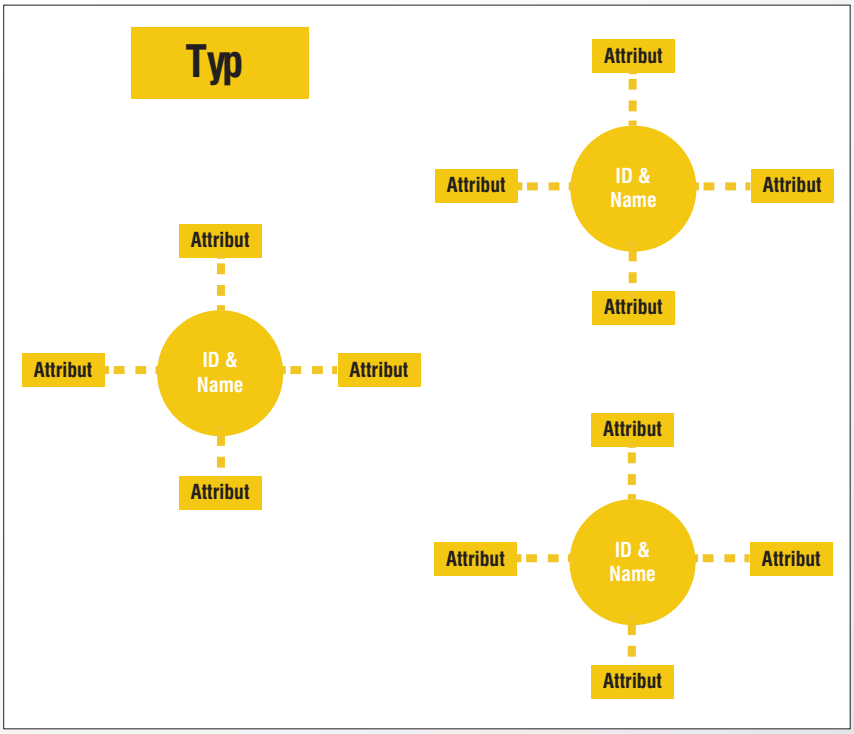

Abb.1: Aufbau einer Entität

Entitäten für die semantische Interpretation von Begriffen

Entitäten spielen für Google eine zentrale Rolle bei der Interpretation von Suchanfragen über Rankbrain, aber auch bei der Interpretation kompletter Inhalte, Sätze bzw. einzelner Aussagen.

Eine Entität ist ein Begriff aus der Philosophie, Semantik und Informatik. Eine Entität beschreibt das Wesen bzw. die Identität eines konkreten oder abstrakten Gegenstands des Seins. Entitäten sind eindeutig identifizierbar und damit einzigartig.

Grundsätzlich kann zwischen „Named Entities", zu Deutsch benannten Entitäten, und Konzepten unterschieden werden. Benannte Entitäten sind Objekte aus der echten Welt wie z. B. Personen, Orte, Organisationen, Produkte, Events … Konzepte sind abstrakte Entitäten physikalischer, psychologischer oder sozialer Natur wie z. B. Entfernung, Quantität, Emotionen, Menschenrechte. Frieden …

Sowohl bei der Nutzung von Suchmaschinen als auch bei der Suchmaschinenoptimierung haben die benannten Entitäten einen größeren Einfluss als Konzepte, da benannte Entitäten in Form der Knowledge Panels den prominentesten Platz einnehmen. Es ist dennoch wichtig, sich generell der Bedeutung von Entitäten bewusst zu sein.

In einem Interview von 2009 sagte Ori Allon, damaliger technischer Leiter des Google Search Quality Teams, in einem Interview mit IDG:

We're working really hard at search quality to have a better understanding of the context of the query, of what is the query. The query isn't the sum of all the terms. The query has a meaning behind it. For simple queries like ‚Britney Spears' and ‚Barack Obama' it's pretty easy for us to rank the pages. But when the query is ‚What medication should I take after my eye surgery?', that's much

harder. We need to understand the meaning ...

Im Kern möchte Google die Bedeutung und darüber die Nutzerintention bzw. Suchintention einer Suchanfrage identifizieren und dementsprechende Inhalte ausliefern. Das ist essenziell für eine positive Nutzererfahrung bei der Nutzung einer Suchmaschine.

Dazu muss Google den Kontext ermitteln. Hierbei sind der suchanfragenbezogene Kontext sowie der Nutzerkontext wie Standort des Suchenden und Suchhistorie wichtig. Beim Nutzerkontext geht es um Personalisierung, die Google laut eigener Aussage fast ausschließlich nur noch auf Standort und Art des Endgeräts als Einflussfaktor beschränkt.

Das wichtigste Kriterium für die Ermittlung des thematischen Kontexts und der Suchintention ist der Suchterm selbst. Hier hat die Einführung von Rankbrain im Jahr 2015 Google einen großen Schritt weitergebracht.

Wörter, die in Suchanfragen oder Inhalten vorkommen, können oft nur im semantischen Kontext verstanden werden. Erst dieser Kontext verleiht Wörtern und Sätzen die Bedeutung.

Betrachten wir zum Beispiel diese beiden Sätze: 1) „Der Jaguar ist aus dem Zoo ausgebrochen." 2) „Der Jaguar des Nachbarn ist kaputt." Das Wort „Jaguar" unterscheidet sich in diesen beiden Sätzen je nach Kontext. Vernünftigerweise sollte man zwei verschiedene Vektorräume des Worts „Jaguar" basierend auf ihren zwei verschiedenen Bedeutungen nutzen.

Es ist daher sinnvoll, einen Algorithmus so zu programmieren, dass bereits vor dem Zuführen der Trainingsdaten ein Text in mögliche semantische Kontexte eingeordnet werden kann. Dann kann der Begriff bzw. die Suchanfrage oder das Dokument in den passenden semantischen Vektorraum eingeordnet und in Beziehung zu anderen Begriffen im gleichen thematischen Kontext gesetzt werden. Dadurch kann dann auch ein bisher unbekannter Begriff gedeutet werden.

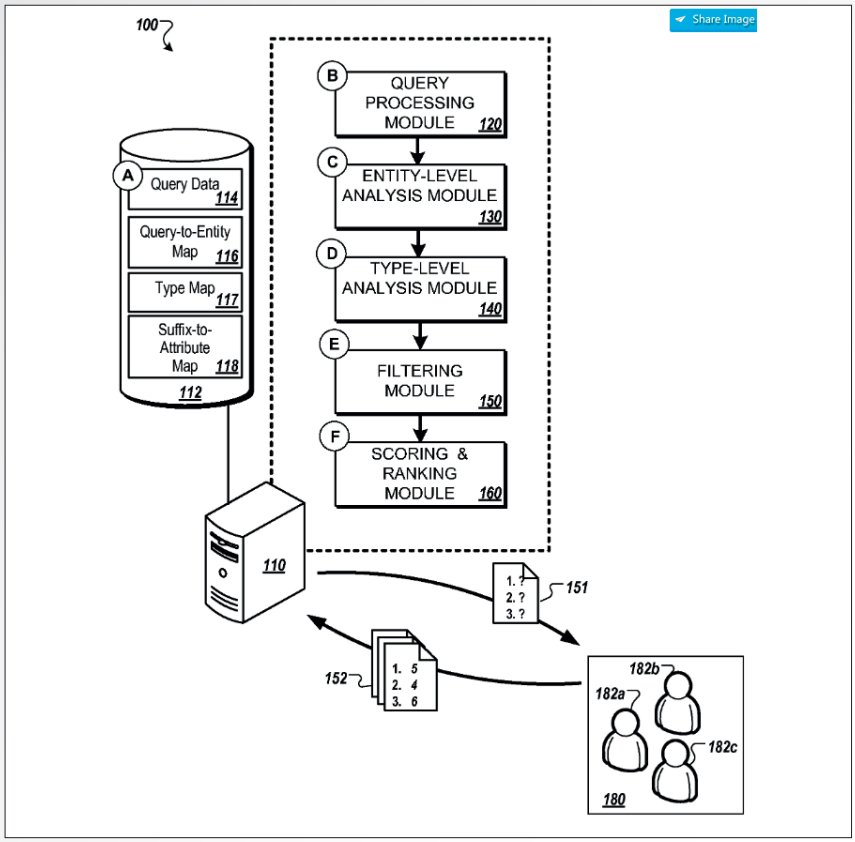

Abb.1: Aufbau einer Entität *(Quelle: © Olaf Kopp, Aufgesang GmbH)*

Durch Rankbrain ist Google seitdem in der Lage, Natural Language Processing (NLP) und Word Embeddings bzw. Vektorraumanalysen automatisiert und skalierbar für die Interpretation von Suchanfragen einzusetzen. Den Weg dorthin ebnen selbstlernende Algorithmen bzw. Systeme (Machine Learning), die es ermöglichen, komplexe Prozesse auch hinsichtlich der Geschwindigkeit bzw. Performance umzusetzen. Durch die Einführung von Rankbrain konnte Google das Spannungsfeld zwischen Skalierung und der Nutzung von NLP kombiniert mit Vektorraumanalysen für ein besseres semantisches Verständnis von Suchanfragen beseitigen.

Über Vektorraumanalysen lassen sich über Word Embeddings Suchanfragen, aber auch Sätze, explizite Fragestellungen oder komplette Inhalte analysieren. Die enthaltenen Wörter können durch ihren Kontext, also die umliegenden Wörter und Entitäten, besser verstanden werden. Durch Word Embeddings lassen sich fehlende Begriffe ergänzen bzw. umschreiben, um einen Satz oder einen Begriff verständlicher zu machen.

Bekannte Modelle für Word Embedding bzw. Vektorraumanalysen für die Anwendung von NLP sind zum Beispiel Word2vec in den zwei verschiedenen Anwendungen CBOW oder Skipgram und das darauf aufbauende von Facebook entwickelte Fasttext Embedding sowie die daraus entwickelten Contextual Embeddings wie z. B. ULM-Fit, Elmo und BERT. Doch das Problem bei diesen Modellen ist der Fokus auf die Begrifflichkeiten.

Aufbau, Struktur und Beziehungen von Entitäten

Entitäten können eindeutig zu einer Unique Identifier, i. d. R. eine eindeutige Zahlenreihe, identifiziert werden. Durch den Kontext der Eigenschaften bzw. Attribute und die Beziehungen zu anderen Entitäten kann jeder Entität eine eindeutige Bedeutung zugeschrieben

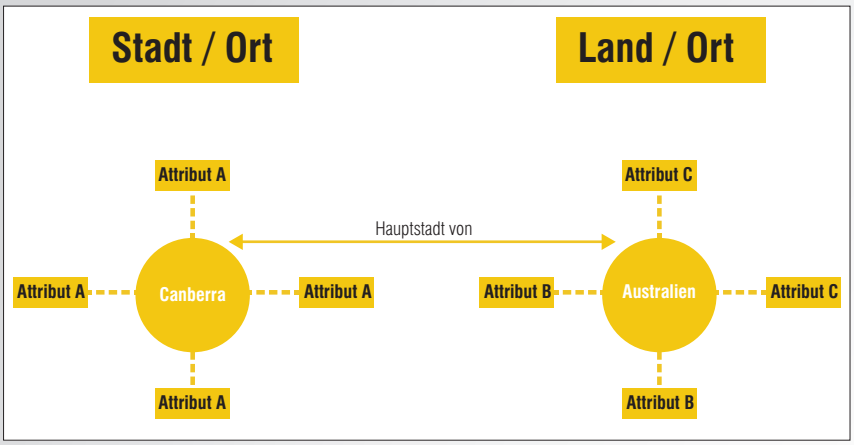

Abb.3: Beispiel für die Kennzeichnung von Entitäten mit Attributen und Beziehungsart

werden, auch wenn der Name der Entität mehrdeutig ist.

Das ist gerade bei mehrdeutigen Entitätsnamen oder der Identifikation von Synonymen wichtig. So kann ein Jaguar sowohl ein Tier als auch eine Automarke oder ein Panzer-Modell sein.

Durch die unterschiedlichen Eigenschaften dieser Entitäten können sie den verschiedenen thematischen Bereichen zugeordnet und voneinander abgegrenzt werden.

Während die Entität Jaguar (Tier) eher mit Eigenschaften wie Fell, Körperbau, Schwanz … in Verbindung gebracht wird und in Beziehung zu anderen Raubkatzen wie Puma oder Leopard steht, wird Jaguar (Automarke) eher mit Attributen wie PS, Motor, km/h, Hubraum … und zusammen mit anderen Automarken wie Porsche, Bentley oder dem britischen Königshaus genannt.

Dadurch können die gleichbenannten, aber unterschiedlichen Entitäten bezüglich ihrer Bedeutung klar voneinander abgegrenzt werden.

Attribute bzw. Eigenschaften und die Art der Beziehung zu anderen Entitäten sind die wichtigsten Classifier, die Google nutzen kann, um die Bedeutung von Entitäten zu verstehen. In verschiedenen wissenschaftlichen Publikationen werden diese Classifier auch als Fakten bezeichnet. Sie dienen neben der Interpretation auch der Zuordnung von Entitäten in Klassen von Entitätstypen.

In verschiedenen Google-Patenten findet man die Begriffe Entitätstypen und Entitätsklassen. Bestimmte Entitätstypen und Entitätsklassen haben eine ähnliche Zusammenstellung von Attributen und bilden damit eine Gruppe. Zum Beispiel können der Entitätsklasse „Person“ oder „Mensch“ immer Attribute wie Geburtsort, Wohnort, Geburtsdatum … zugeordnet werden. Dadurch ist der Entitätstyp klar definiert.

Entitätstypen beschreiben Gruppen von Entitäten, die aufgrund gleicher oder ähnlicher Attribute in Klassen zusammengefasst werden können.

Dabei gibt es bei den Entitätstypen unterschiedliche Hierarchieebenen, die in Beziehung zueinander stehen können. So sind die Entitätstypen Stadt und Land Sub-Klassen des Entitätstyps Ort. In der Regel haben Entitätstypen, die einer gleichen Hauptklasse angehören, mehrere gleiche Standard-Attribute. In der Gänze ihrer Attribute sind sie aber unterschiedlich.

Im dem Buch „Entity Oriented Search“ von Krisztian Balog findet man folgende Beschreibung für Entitätstypen:

Entities may be categorized into multiple entity types (or types for short). Types can also be thought of as containers (semantic categories) that group together entities with similar properties. An analogy can be made to object oriented programming, whereby an entity of a type is like an instance of a class.

Über eine Gewichtung der Attribute je Entität kann Google zum einen feststellen, wie relevant ein bestimmtes Attribut für eine Entität ist. Zum anderen könnte Google darüber auch die Relevanz der Entität für eine gestellte Suchanfrage nach diesem Attribut ermitteln.

Das Google-Patent Identifying and ranking attributes of entities (*http://einfach.st/gpat43*) zeigt einen Ansatz, wie so etwas funktionieren könnte.

Laut diesem Patent können über die Eingaben bestimmter Suchterm-Kombinationen Attribute Entitäten zugeordnet und gewichtet werden.

One innovative aspect of the subject matter described in this specification is embodied in methods that include the actions of: identifying queries in query data; determining, in each of the queries, (i) an entity-descriptive portion that refers to an entity and (ii) a suffix; determining a count of a number of times the one or more queries were submitted; estimating, based on the count, an entity-level count of query submissions that include the particular suffix and are considered to refer to a first entity; determining that the entity is a particular type of entity; determining a type-level count of the query submissions that include the first suffix and are estimated to refer to entities of the particular type of entity; and assigning a score to the particular suffix based on the entity-level count and the type-level count.

Über diese Methode könnte Google festlegen, welche Informationen zu Entitäten eines bestimmten Entitätstyps im Knowledge Panel angezeigt werden. Des Weiteren ließe sich bei mehrdeutigen Aussagen feststellen, welches Attribut das relevanteste ist, bezogen auf das Beispiel von oben.

Hier ein Beispiel:

Larry Page ist als Unternehmer, Informatiker und Ingenieur tätig. Welche dieser drei Aussagen ist die relevanteste bzw. zutreffendste?

Je mehr Menschen nach „Larry Page Unternehmer“ suchen, desto zutreffender ist das Attribut „Unternehmer“.

Anders ausgedrückt. Der Entitäts-

typ Unternehmer liegt im Vektorraum näher am Entitäten-Vektor Larry Page als die Entitätstypen Informatiker und Ingenieur.

Der Knowledge Graph: Googles semantische Datenbank

Der Knowledge Graph ist Googles semantische Datenbank. Hier werden Entitäten in Beziehung zueinander gestellt, mit Attributen versehen und in thematischen Kontext bzw. Ontologien gebracht.

Die grundsätzliche Struktur von Graphen besteht aus sogenannten Knoten und Kanten. Bezogen auf den Knowledge Graph sind die Knoten die Entitäten und die Kanten beschreiben die Art der Beziehung zwischen diesen Entitäten. Entitäten werden beschrieben durch eine Bezeichnung bzw. einen Namen und verschiedene Attribute bzw. Eigenschaften.

In einem Knowledge Graph werden alle Knoten, also Entitäten, mit Attributen versehen und nach Entitätstypen klassifiziert. Zudem werden die Kanten zwischen den Entitäten mit der Beziehungsart kommentiert.

Diese Struktur erlaubt es, Antworten auf Fragen zu geben, in denen ein Thema oder eine Entität gesucht wird, die in der Frage nicht genannt wird.

Im folgenden Beispiel sind „Australien“ und „Canberra“ die Entitäten und der Wert „Hauptstadt“ beschreibt die Art der Beziehung.

Diese Grafik sagt nichts anderes aus als: „Canberra ist die Hauptstadt von Australien.“ Damit kann Google auf die Frage „Welche Stadt ist die Hauptstadt von Australien?“ die richtige Antwort geben. Dabei ist es nicht wichtig, ob man explizit fragt oder implizit die Frage über den Suchterm „hauptstadt australien“ stellt. Das Ergebnis ist das gleiche.

Man kann diesen Zusammenhang auch grammatikalisch so darstellen: „Canberra“ ist das Subjekt, „Australien“ das Objekt und „(ist die) Hauptstadt“ ist das Prädikat bzw. die Prädikatsphrase.

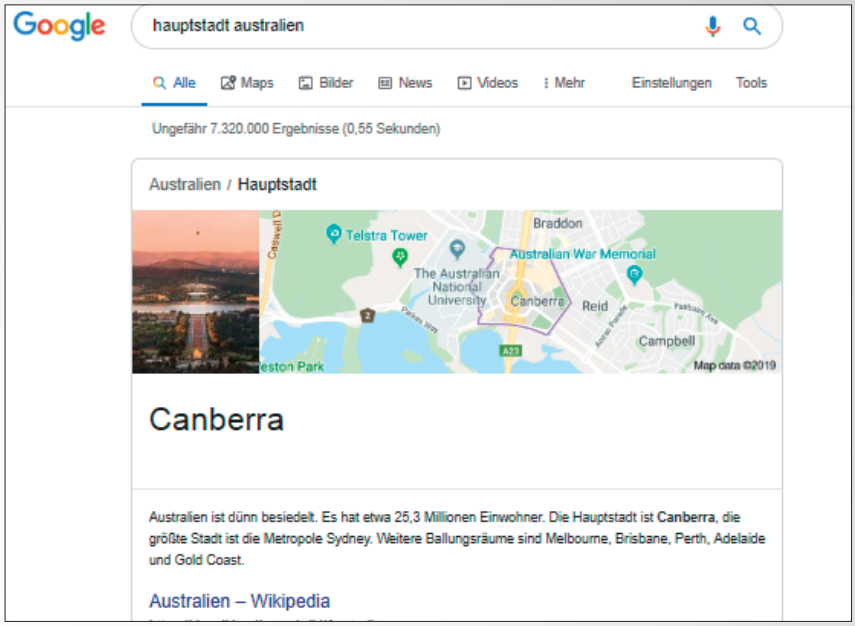

Abb. 4: Antwort auf die Frage nach der Hauptstadt von Australien bei Google

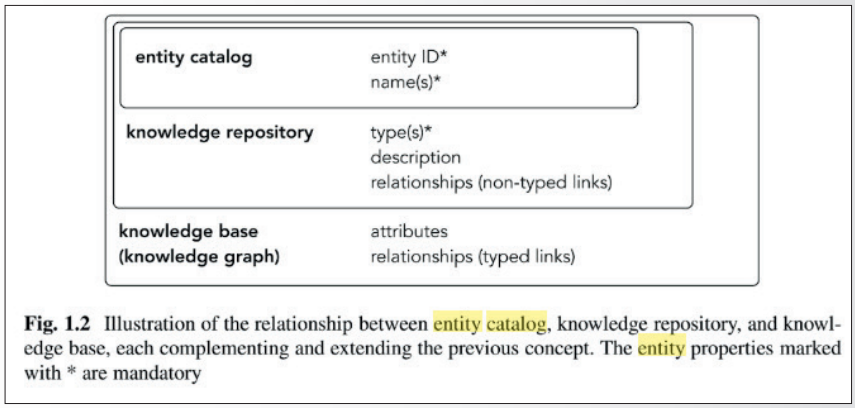

Abb. 5: Die drei Ebenen eines Knowledge Graph *(Quelle: Entity-Oriented Search – Krisztian Balog)*

Die strukturierten Daten kann Google über das Resource Description Framework, kurz RDF, erfassen. Eine Entität ist eine Zusammenfassung verschiedener RDF-Statements nach dem Muster Objekt – Prädikat – Subjekt. Ein Statement wäre z. B.: „Canberra ist die Hauptstadt von Australien.“ Man kann diesen Zusammenhang auch grammatikalisch so darstellen: „Canberra“ ist das Subjekt, „Australien“ das Objekt und „(ist die) Hauptstadt“ das Prädikat. Die Beziehungsart kann aber auch durch ein Verb beschrieben werden wie: „Thomas Müller spielt für Bayern München.“ Objekt und Subjekt sind demnach immer Entitäten. Das Prädikat kann ein Entitätstyp oder eine Entitätsklasse, ein Attribut, ein Verb oder eine Kombination aus allen sein. Nomen werden im Natural Language Processing prinzipiell immer als potenzielle Entitäten gesehen.

Aber der Knowledge Graph ist mehr als eine Darstellung der Beziehung zwischen Entitäten. Er ist eine riesige Datenbank, in der Google das Wissen rund um Entitäten sammelt. Deswegen gibt es noch weitere Informationen, die im Knowledge Graph erfasst werden:

- Attribute (Eigenschaften) von Entitäten
- Relevanz-Scoring der Attribute, also wie nah die Attribute im Vektorraum zu den Entitäten stehen
- Entitätstypen

Als Grundlage für den Knowledge Graph dienen drei Ebenen:

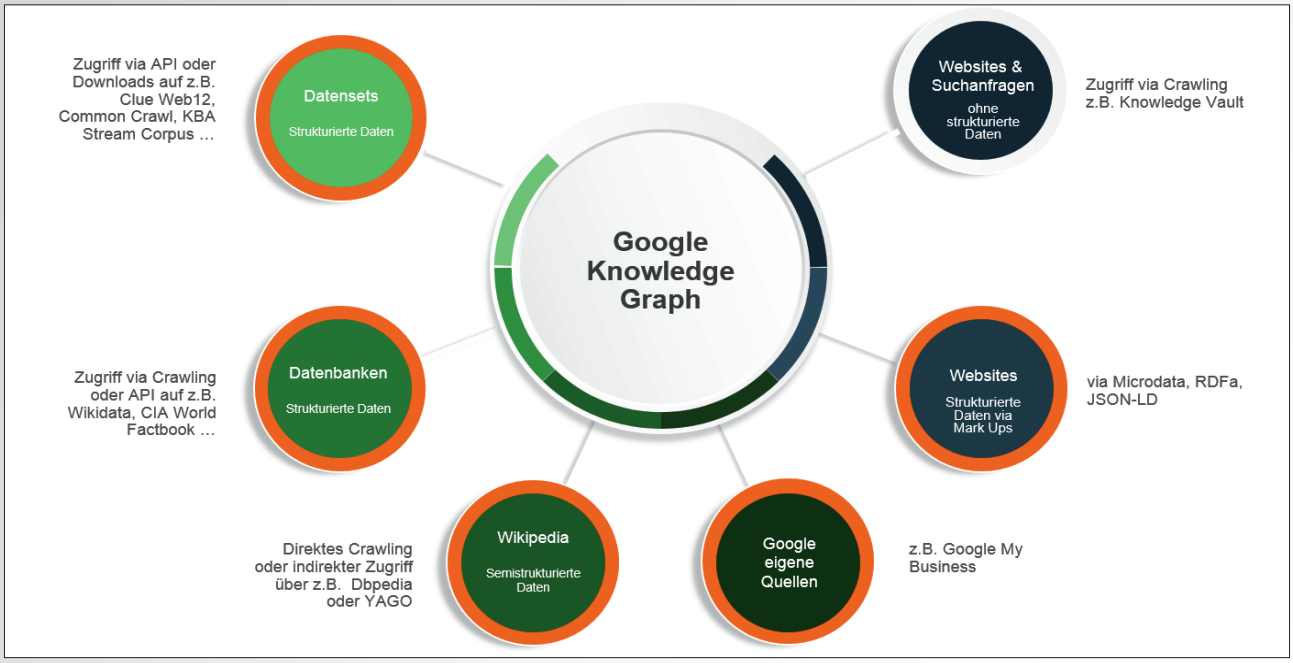

Abb. 6: Übersicht strukturierte, semistrukturierte und unstrukturierte Datenquellen für den Google Knowledge Graph *(Quelle: © Olaf Kopp, Aufgesang GmbH)*

» **Entitäten-Katalog:** Hier werden alle Entitäten gespeichert, die mit der Zeit identifiziert wurden.
» **Knowledge Respository:** Die Entitäten werden in einem Wissens-Depot (Knowledge Repository) mit den Informationen bzw. Attributen aus den verschiedenen Quellen zusammengeführt. Im Knowledge Repository geht es in erster Linie um die Zusammenführung und Speicherung von Beschreibungen und die Bildung semantischer Klassen bzw. Gruppen in Form von Entitätstypen. Googles Knowledge Repository ist aktuell der Knowledge Vault.
» **Knowledge Graph:** Im Knowledge Graph werden die Entitäten mit Attributen ergänzt und Beziehungen zwischen den Entitäten hergestellt.

Wie erkennt Google Entitäten und wie funktioniert das Data Mining für den Knowledge Graph?

Googles größte Herausforderung bezüglich des Knowledge Graph ist das Data Mining von Informationen bzw. Attributen bezüglich Entitäten sowie Entitätstypen und -klassen. Dabei gibt es ein großes Spannungsfeld zwischen der Bereitstellung einer möglichst vollständigen Wissens-Datenbank, die nahezu alle Informationen über Entitäten bereitstellt, und der Richtigkeit der Informationen.

Als Informationsquellen für den Knowledge Graph stehen Google eine Vielzahl an Datenbanken, manuell gepflegten Lexika und frei verfügbare Online-Quellen wie soziale Medien und Websites zur Verfügung. Grundsätzlich muss hier zwischen Datenquellen mit strukturierten, semistrukturierten und unstrukturierten Inhalten bzw. Informationen unterschieden werden.

Die einfachste Art für Google, um an Informationen zu Entitäten zu kommen, sind Quellen, über die strukturierte Daten bereitgestellt werden.

Strukturierte Datenquellen sind regelmäßig aktualisierte Datensets, auf die via API und Datenbanken mittels Crawling zugegriffen werden kann. Hier sind die Informationen bereits für Maschinen lesbar angelegt und können ohne große Aufarbeitung in den Knowledge Graph übernommen werden. Eine weitere Quelle für strukturierte Daten sind Inhalte von Websites, die vom Website-Betreiber mittels schema.org oder Json-Ld ausgezeichnet wurden.

Die meisten strukturierten Datenbanken stellen die Informationen im maschinenlesbaren RDF-Format zur Verfügung bzw. lassen eine Übersetzung in dieses Format zu. Google greift auf Datenbanken zu, in die sie Vertrauen haben, wie z. B. **Wikidata, CIA World Factbook** ..., strukturierte Datensets oder Übersetzungs-Datenbanken wie z. B. **DBpedia** oder **YAGO**, die die Informationen der Wikipedia in maschinenlesbare Daten übersetzen.

Da diese strukturierten Datenbanken und Datensets nur sehr langsam wachsen und aktualisiert werden, wundert es nicht, dass Google Webmaster immer wieder dazu animiert, mit strukturierten Daten in ihren Websites zu arbeiten. Je mehr strukturierte Daten Google sammelt und verarbeitet, desto näher kommen sie dem Ziel, auch unstrukturierte Daten verarbeiten zu können. Die strukturierten Daten funktionieren als Trainingsdaten für das maschinelle Lernen.

Die Auszeichnung von Inhalten durch **strukturierte Daten** auf der eigenen Website scheint ein probates Mittel zur Snippet-Optimierung zu sein, allerdings findet man nur sehr wenige Beispiele, bei denen eine Auszeichnung

mit strukturierten Daten zu einer Veränderung von Inhalten in Knowledge Panels führte.

Kürzlich gab Google erst bekannt, dass die Auszeichnung der sozialen Profile mit Markups nicht mehr notwendig ist, um diese im Knowledge Panel anzuzeigen. Das war einer der wenigen Anwendungsfälle, in denen man das Knowledge Panel über eigens mit Markups ausgezeichnete Inhalte beeinflussen konnte. Ein weiteres Indiz dafür, dass Google Markups nur als Zwischenlösung nutzt.

Hinsichtlich strukturierter Daten scheint aktuell Wikidata die wichtigste Quelle zu sein, zumindest was die Inhalte der Knowledge Panels und anderer Knowledge-Graph-Boxen angeht. Wikidata ist eine der größten öffentlichen Sammlungen allgemeiner Informationen, bestehend aus mehr als 400 Millionen Aussagen, über 200 Millionen Labels und Aliasse sowie über 1,2 Milliarden Kurzbeschreibungen in mehreren Hundert Sprachen zu mehr als 45 Millionen Entitäten.

Gerade für die Anreicherung von Entitäten und Entitätstypen-Klassen mit Standard-Attributen spielen strukturierte Daten eine große Rolle, aber nur als menschlich verifizierte Trainingsdaten und wahrscheinlich nur so lange, bis Google sie nicht mehr benötigt, da der Algorithmus genug Daten zum Lernen für die eigenen Modelle generiert hat *(http://einfach.st/semstruc)*.

Wikipedia spielt als semistrukturierte Datenquelle aktuell die größte Rolle für den Knowledge Graph.

Die durch die Wikipedianer sehr zuverlässig gepflegten Daten bieten einen hohen Richtigkeitsgrad der Informationen und die standardisierte Form der Wikipedia-Beiträge eine grundlegende Struktur. Zudem stellen Datenbanken wie DBpedia oder YAGO die Inhalte der Wikipedia in strukturierter Form kostenlos zur Verfügung.

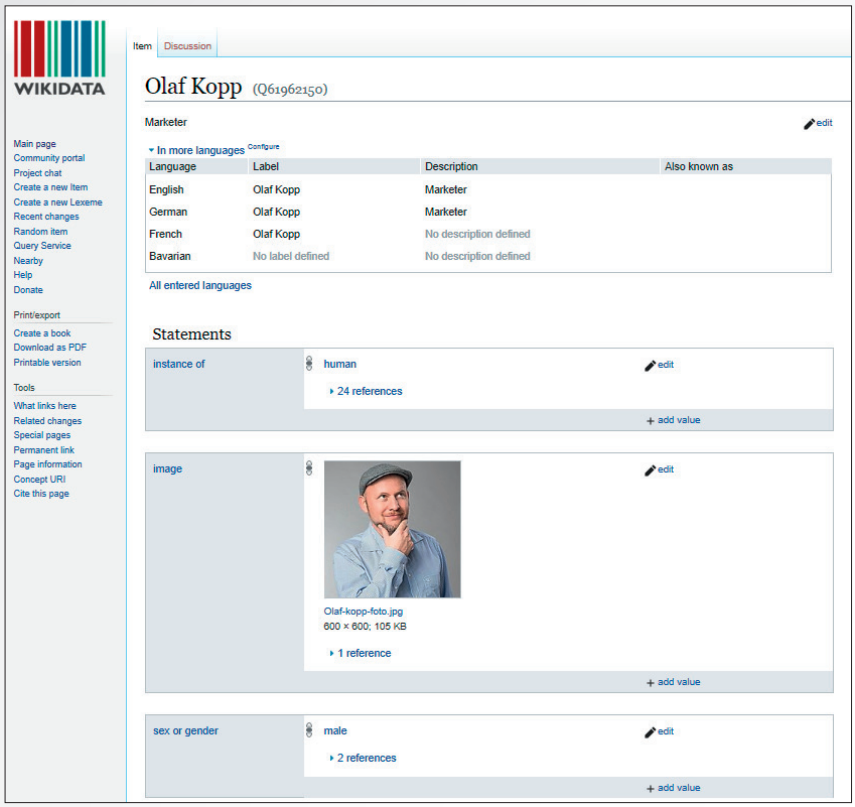

Abb. 7: Beispiel eines Wikidata-Eintrags für die Entität Olaf Kopp

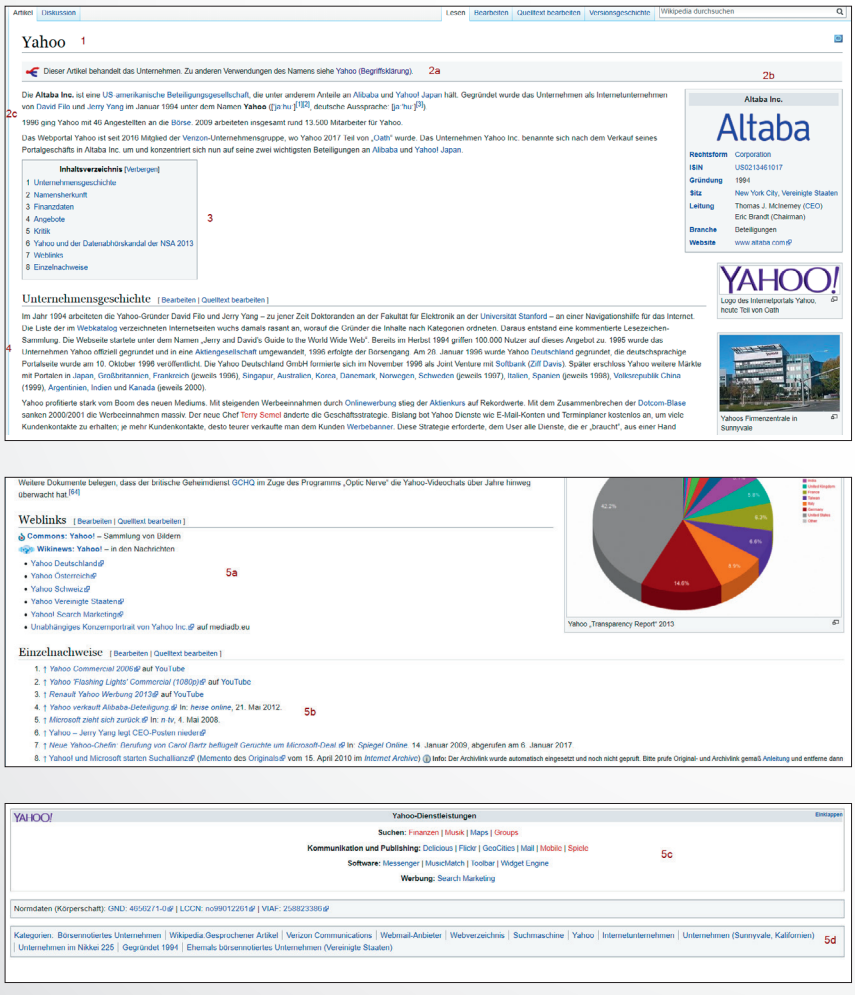

Abb. 8: Beispiel-Struktur eines Wikipedia-Eintrags


Artikel Diskussion

Yahoo (Begriffsklärung)

Yahoo bezeichnet:

- ein amerikanisches Internetunternehmen, siehe Yahoo
- Menschen im Roman Gullivers Reisen von Jonathan Swift
- Yahoo Falls, ein Wasserfall im McCreary County, Kentucky

Yahoo ist der Name folgender Personen:

- Yahoo Serious (* 1953; bürgerlich *Greg Pead*) australischer Entertainer

Siehe auch:

Wiktionary: yahoo – Bedeutungserklärungen, Wortherkunft, Synonyme, Übersetzungen

Dies ist eine **Begriffsklärungsseite** zur Unterscheidung mehrerer mit demselben Wort bezeichneter Begriffe.

Abb. 9: Beispiel für eine Begriffserklärungsseite von Yahoo bei Wikipedia

Semistrukturierte Daten sind Informationen, die nicht nach allgemeinen Auszeichnungsstandards wie z. B. nach RDF, schema.org … explizit ausgezeichnet sind, aber eine implizite Struktur aufweisen. Aus dieser impliziten Struktur lassen sich i. d. R. über Workarounds strukturierte Daten gewinnen.

Die Extrahierung der Informationen aus Datenquellen mit semistrukturierten Daten kann über einen Template-Based Extractor durchgeführt werden. Dieser kann aufgrund einer immer wiederkehrenden gleichen Struktur von Beiträgen Inhaltsabschnitte identifizieren und aus ihnen Informationen extrahieren.

Wikipedia basiert technisch auf dem MediaWiki-CMS. Dadurch sind die Inhalte mit rudimentären Markups versehen und können einfach via XML-, SQL-Dumps oder als HTML downgeloadet werden.

Die Struktur eines typischen Wikipedia-Beitrags ist eine Template-Vorlage für die Klassifizierung von Entitäten nach Kategorien, Identifikation von Attributen und das Extrahieren von Informationen für Featured Snippets und Knowledge Panels. Der sehr ähnliche bzw. identische Aufbau der einzelnen Wikipedia-Beiträge sieht z. B. so aus:

- Titel (1)
- Lead Section (2)
  - Einführung (2a)
  - Infobox (2b)
  - Einleitungstext (2c)
- Inhaltsverzeichnis (3)
- Haupttext (4)
- Ergänzungen (5)
  - Fußnoten, Weblinks und wissenschaftliche Arbeiten (5a)
  - Weiterführende Links (5b)
  - Kategorien (5d)

Der Titel eines jeden Wikipedia-Beitrags gibt den Entitäts-Namen wieder. Bei mehrdeutigen Titeln wird zur klareren Abgrenzung zu den anderen Entitäten mit gleichen Namen der Typ im Titel mit angehängt, wie beim Zeichner Michael Jordan. Dort lautet der Titel Michael Jordan (Zeichner), um ihn von der populäreren Entität des Basketballspielers Michael Jordan abzugrenzen.

Die Info-Box (2b) eines Wikipedia-Beitrags rechts oben stellt strukturierte Daten zur jeweiligen Entität bereit. Den Einleitungstext (2c) findet man häufig im Knowledge Panel zu der jeweiligen Entität wieder. Dazu weiter unten in diesem Beitrag mehr.

Die internen Verlinkungen innerhalb der Wikipedia geben Google einen Aufschluss darüber, welche weiterführenden Themen bzw. anderen Entitäten im semantischen Zusammenhang mit der jeweiligen Entität stehen. Deswegen nutzen wir in der Agentur bereits seit über vier Jahren ein eigenes Wikipedia-Script, welches die internen Verlinkungen relevanter Wikipedia-Beiträge analysiert.

Die Wikipedia bietet eine Reihe an Spezialseiten, die Google helfen können, Entitäten besser zu verstehen, zu gruppieren und zu klassifizieren.

- **Listen- & Kategorie-Seiten für die Einordnung nach Entitätstypen und -klassen:** Die Kategorien, denen eine Entität in der Wikipedia zugeordnet ist, findet man immer am Ende eines Beitrags (siehe 5d). Auf den Kategorie-Seiten selber findet man eine Übersicht aller Oberkategorien, Unterkategorien und Entitäten, die dieser Kategorie zugeordnet sind. Listen-Seiten (z. B. hier) geben ähnlich wie Kategorie-Seiten eine Übersicht aller Elemente, die mit dem Listen-Thema in Verbindung stehen. Über diese beiden Seitentypen könnte Google die jeweilige Entität den Entitätstypen und -klassen zuordnen. Wikipedia verfügt verglichen mit den anderen großen Wissens-Datenbanken über die meisten Typen-Klassen. *Quelle: Entity Oriented Search von Krisztian Balog*

Welche zentrale Rolle die Wikipedia bei der Identifikation von Entitäten und deren thematischem Kontext spielen kann, zeigt das wissenschaftliche Papier Using Encyclopedic Knowledge for Named Entity Disambiguation.

Beziehungen zwischen Entitäten könnte Google u. a. über Annotationen bzw. Verlinkungen innerhalb der Wikipedia herstellen.

An annotation is the linking of a mention to an entity. A tag is the annotation of a text with an entity which captures a topic (explicitly mentioned) in the input text.

- **Weiterleitende Spezialseiten** leiten Nutzer der Wikipedia weiter zum Hauptbegriff. In diesem Beispiel ist Internet-Marketing ein Synonym für

den Hauptbegriff Online-Marketing. Über diese weiterleitenden Seiten kann Google synonyme Bezeichnungen für eine Entität identifizieren und sie der Hauptbezeichnung zuordnen. Dies funktioniert ähnlich einem Canonical-Tag in der Suchmaschinenoptimierung.

» **Begriffserklärungs-Seiten** für die Erkennung mehrfacher Bedeutungen geben einen Überblick über alle Entitäten, die einen identischen oder sehr ähnlichen Namen haben. Dadurch bekommt Google einen Überblick, bei welchen Namen es mehrdeutige Entitäten gibt.

Die Verarbeitung strukturierter und semistrukturierter Daten z. B. aus der Wikipedia oder Wikidata stellt kein großes Problem für Google mehr dar. Aber strukturierte Daten sind nur bedingt verfügbar und sind abhängig von der manuellen Pflege, was einen skalierbaren vollständigen Aufbau von Wissen verhindert.

Data Mining über unstrukturierte Datenquellen via NLP

Der Knowledge Graph ist deswegen noch sehr lückenhaft, da die Informationen aus den genannten Datenquellen sehr unvollständig sind, was die Gesamtmenge aller Entitäten in der realen Welt betrifft.

Die Verarbeitung unstrukturierter Daten aus dem gesamten Fundus des Internets über Natural Language Processing (NLP) würde dieses Ziel erfüllen. Aber hier stellt gerade die Bewertung und Feststellung der Richtigkeit eine riesige Herausforderung dar. Hier scheint Google selbstlernende Algorithmen (Machine Learning) als Lösung entdeckt zu haben, was die viele Google-Patente hinsichtlich Data Mining via Machine bzw. Deep Learning aus den letzten Jahren zeigen.

Zur Gewährleistung eines hohen Grads an Vollständigkeit stellen unstrukturierte Datenquellen eine wichtige, aber auch herausfordernde Datenquelle für den Knowledge Graph dar. Für das Data Mining unstrukturierter Datenquellen hat Google 2014 das Projekt Knowledge Vault der Öffentlichkeit vorgestellt. Der Knowledge Vault soll in der Lage sein, unstrukturiertes Wissen aus dem gesamten Internet wie z. B. Suchanfragen, soziale Medien und Websites jeglicher Art zu generieren, zu verifizieren und strukturiert für den Knowledge Graph aufzuarbeiten. Hier spielt Natural Language Processing (NLP) eine wichtige Rolle.

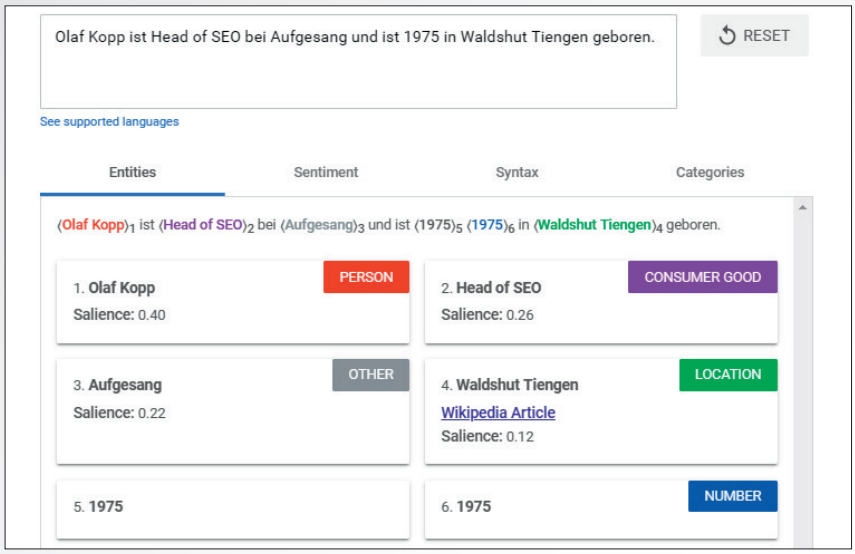

Abb. 10: Beispiel einer Entitäten-Analyse über die Natural Language Processing API von Google

Über Natural Language Processing (NLP) ist Google bereits in der Lage, für Entitäten aus unstrukturierten Texten heraus folgende Analysen und Maßnahmen vollautomatisiert durchzuführen:

» **Kennzeichnung von Wörtern nach Wortarten**: Wortartenkennzeichnung klassifiziert Wörter nach Wortarten wie z. B. Subjekt, Objekt, Prädikat, Adjektiv …

» **Analyse und Extraktion von benannten Entitäten**: Dieser Aspekt sollte uns aus den vorangegangenen Beiträgen bekannt sein. Damit wird versucht, Wörter mit einer „bekannten“ Bedeutung zu identifizieren und Klassen von Entitätstypen zuzuordnen. Im Allgemeinen sind benannte Entitäten Menschen, Orte und Dinge (Substantive). Entitäten können auch Produktnamen enthalten. Dies sind im Allgemeinen die Wörter, die ein Knowledge Panel auslösen. Aber auch Begriffe, die kein eigenes Knowledge Panel auslösen, können Entitäten sein.

» **Wortabhängigkeiten**: Wortabhängigkeiten schaffen Beziehungen zwischen den Wörtern basierend auf Grammatikregeln. Dieser Prozess bildet auch „Sprünge“ zwischen Wörtern ab.

» **Parsing Labels**: Die Kennzeichnung klassifiziert die Abhängigkeit oder die Art der Beziehung zwischen zwei Wörtern, die über eine Abhängigkeit verbunden sind.

» **Salience-Scoring**: Salience bestimmt, wie intensiv ein Text sich mit einem Thema beschäftigt. Dies wird in NLP basierend auf den sogenannten Indikatorwörtern bestimmt. Im Allgemeinen wird der Bekanntheitsgrad durch das Mitzitieren von Wörtern im Web und die Beziehungen zwischen Entitäten in Datenbanken wie Wikipedia und Freebase bestimmt. Google wendet dieses Verknüpfungsdiagramm wahrscheinlich auch auf die Entitätsextraktion in Dokumenten an, um diese Wortbeziehungen zu bestimmen. Ein ähnliches Vorgehen kennen erfahrene SEOs von der TF-IDF-Analyse.

» **Sentiment-Analysen**: Kurz gesagt ist dies eine Bewertung der in einem

```
{
      „@type“: „EntitySearchResult“,
      „result“: {
        „@id“: „kg:/m/012x34“,
        „name“: „Jaguar Cars“,
        „@type“: [
          „Organization“, „Corporation“, „Thing“
        ],
        „description“: „Automobilhersteller“,
        „detailedDescription“: {
          „articleBody“: „Jaguar ist die Luxusfahrzeugmarke von
                         Jaguar Land Rover, einem britischen multina-
                         tionalen Automobilhersteller mit Hauptsitz
                         in Whitley, Coventry, England.“,
          „url“: „https://en.wikipedia.org/wiki/Jaguar_Cars“,
          „license“: „https://en.wikipedia.org/wiki/Wikipedia:Text_
                     of_Creative_Commons_Attribution-ShareA-
                     like_3.0_Unported_License“
        },
        „image“: {
          „contentUrl“: „http://t0.gstatic.com/images?q=tbn:ANd-
                        9GcQOjSYSg4NkqZeKvwHez-VjQ9l7Ic“,
          „url“: „https://commons.wikimedia.org/wiki/File:Jaguar_
                 XK8_Convertible_-_Flickr_-_The_Car_Spy_(22).jpg“
        },
        „url“: „http://www.jaguar.co.uk/“
      },
      „resultScore“: 139.01976
},
```


Code 1: Ergebnis einer Anfrage zum gesuchten Begriff „jaguar“

Artikel zum Ausdruck gebrachten Meinung (Ansicht oder Haltung) zu im Text behandelten Entitäten. Diese Sentiment-Analysen lassen sich auch auf Entitäten anwenden.

» **Fachkategorisierung**: Auf Makroebene klassifiziert NLP Text in Betreffkategorien. Die Kategorisierung von Themen hilft dabei, allgemein zu bestimmen, worum es in dem Text geht.
» **Textklassifizierung & Funktion**: NLP kann noch weiter gehen und die beabsichtigte Funktion bzw. den Zweck des Inhalts bestimmen.
» **Extrahierung von Content-Typen**: Google kann mithilfe von Strukturmustern bzw. des Kontexts den Inhaltstyp eines bestimmten Texts ohne die Ausweisung mit strukturierten Daten bestimmen. Das HTML, die Formatierung des Texts und der Datentyp des Texts (Datum, Ort, URL usw.) können verwendet werden, um den Text ohne zusätzliches Markup zu verstehen Mithilfe dieses Prozesses kann Google ermitteln, ob es sich bei Text um ein Ereignis, ein Rezept, ein Produkt oder einen anderen Inhaltstyp handelt, ohne dass Markups verwendet werden müssen.
» **Identifikation einer impliziten Bedeutung aufgrund der Struktur**: Die Formatierung eines Textkörpers kann seine implizite Bedeutung ändern. Überschriften, Zeilenumbrüche, Listen und Nähe vermitteln ein sekundäres Verständnis des Textes. Wenn beispielsweise Text in einer HTML-sortierten Liste oder in einer Reihe von Überschriften mit Zahlen davor angezeigt wird, handelt es sich wahrscheinlich um einen Vorgang oder eine Rangfolge. Die Struktur wird nicht nur durch HTML-Tags definiert, sondern auch durch die visuelle Schriftgröße/-stärke und -nähe beim Rendern.

Natural Language Processing bietet gerade in Kombination mit Machine Learning viele Möglichkeiten für automatisierte Analysen weit über die Anwendung auf den Knowledge Graph hinaus. Über **Natural Language Processing** können aus Texten jeglicher Art Entitäten identifiziert, extrahiert und mit Attributen angereichert werden. Zudem können Beziehungen zwischen Entitäten ermittelt werden. Also alle Aufgaben, die man für den Aufbau eines Knowledge Graph hinsichtlich eines vollständigen Informationsbestands braucht, können erledigt werden.

Je besser die selbstlernenden Algorithmen funktionieren, desto besser können Inhalte, Suchanfragen und Entitäten hinsichtlich Bedeutung und Relevanz analysiert und interpretiert werden. Die Vollständigkeit einer Wissens-Datenbank wie des Knowledge Graph wäre garantiert.

Die große Herausforderung ist die Sicherstellung der Richtigkeit der Informationen. Hier scheint Google noch kein Mittel gefunden zu haben, damit der Knowledge Vault zur vollen Entfaltung kommen kann.

Ein Blick in die Struktur über die Knowledge Graph API

Die Knowledge Graph API ist eine Schnittstelle, über die Entwickler Informationen aus dem Google Knowledge Graph transferieren können. Die Informationen können über diese Schnittstelle nur gelesen, nicht verändert oder ergänzt werden. Sie kann u. a. genutzt werden für:

» Erhalten einer Rangliste der bekanntesten Entitäten, die bestimmten Kriterien entsprechen
» Objekte in einem Suchfeld vorausschauend vervollständigen
» Annotieren/Organisieren von Inhalten mithilfe der Knowledge-Graph-Entitäten

Ein Blick über die ausgegebenen Attribute gibt einen Aufschluss darüber, wie die Daten im Knowledge Graph organisiert sind. Folgende Ausgabe-Attribute gibt es:

» **@id**: Der kanonische URI für die Entität
» **@type**: Entitätstypen bzw. die Liste der unterstützten schema.org-Typen, die der Entität entsprechen
» **Description**: Eine kurze Beschreibung der Entität
» **Bild-URL**: Ein Bild zur Identifizierung der Entität
» **Detailed Description**: Eine detaillierte Beschreibung der Entität
» **URL**: Die offizielle Website-URL der Entität, falls verfügbar
» **resultScore**: Ein Indikator dafür, wie gut die Entität mit den Anforderungseinschränkungen übereinstimmt

In Code 1 sehen Sie eine Abfrage zum gesuchten Begriff „jaguar“ über die Knowledge Graph API.

Das Knowledge Panel und weitere SERP-Boxen als Fenster im Knowledge Graph

Sobald Google erkennt, dass es sich bei einer Suchanfrage um eine entitätenbasierte Suche handelt, werden verschiedene Boxen-Varianten in den SERPs ausgeliefert.

Hierbei spielt es eine entscheidende Rolle, ob nach einem Attribut einer Entität, einer Beziehungsart zwischen Entitäten oder direkt nach einer Entität gesucht wird. Auch der Entitätstyp ist entscheidend dafür, welche Boxen-Varianten ausgespielt werden. Wenn man z. B. nach Entitäten der Entitätstypen Personen, Orte, Bauwerke oder Unternehmen sucht, bekommt man das klassische Knowledge Panel auf der rechten Seite der SERPs ausgeliefert.

Dieses Knowledge Panel kann man auch als Entitäten-Box bezeichnen und es wird für nahezu alle Entitätstypen ausgeliefert.

Das klassische Knowledge Panel erkennt man an dem Teilen-Button im oberen Bereich des Panels. Nicht zu verwechseln mit den My-Business-Boxen. Diese basieren nicht auf dem Knowledge Graph, sondern auf einem Eintrag bei Google My Business. Inwiefern die Daten aus My Business auch im Knowledge Graph berücksichtigt werden, ist nicht klar, aber es ist nicht unwahrscheinlich.

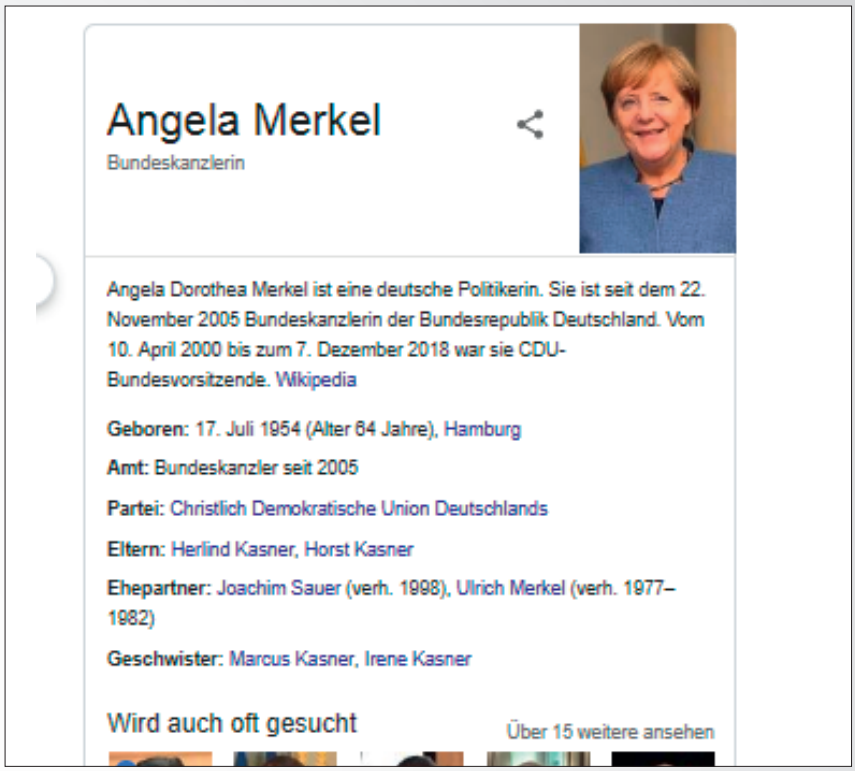

Abb. 11: Beispiel Knowledge Panel für die Suchanfrage „angela merkel“

Bei bestimmten Entitätstypen werden zusätzliche Boxen, auch Knowledge Cards genannt, angezeigt.

Bei Suchen nach Entitätstypen wie z. B. Ereignissen bekommt man neben dem klassischen Knowledge Panel zusätzlich eine Knowledge Card mit dem Attribut Datum des Ereignisses angezeigt, aber nur, wenn das Ereignis in der Zukunft liegt.

Aktuelle Ereignisse kann Google zum einen über eine plötzlich steigende Anzahl an Suchanfragen zu einer Entität und/oder einer Kombination aus Entitäten und Event-Trigger-Begriffen erkennen. Oder Google erkennt ein Ereignis über Kookkurrenzen dieser Merkmale in aktuellen Berichten in Nachrichten-Magazinen, Blogs …

Ereignisse selbst sind Entitäten und stehen oft mit anderen Entitäten wie z. B. Schauspieler, Musiker, Bands oder andere prominente Teilnehmer von Veranstaltungen im Zusammenhang. Ereignisse ohne Entitäten sind ungewöhnlich.

Bei der Suche nach Filmen können u. a. Plattformen und Kinovorstellungen angezeigt werden.

Wenn eine Suchanfrage nicht eindeutig nur mit einer Entität beantwortet werden kann, wird oft ein Karussell oberhalb der SERPs angezeigt. Im folgenden Beispiel beinhaltet die Suchanfrage „schauspieler in“ als Prädikatsphrase und die Entität „avengers“ als Objekt. Das gesuchte Subjekt kann nicht nur durch eine einzige Entität beantwortet werden. So liefert Google ein Karussell aller Entitäten aus, die passen. Durch einen Klick auf eine der angezeigten Entitäten lässt sich die Suche spezifizieren.

Je nach Entitätstyp gibt es ein „Standard-Set“ an Attributen. So gehören zu Filmen Attribute wie z. B.:

» Erscheinungsdatum
» Regisseur
» Titel

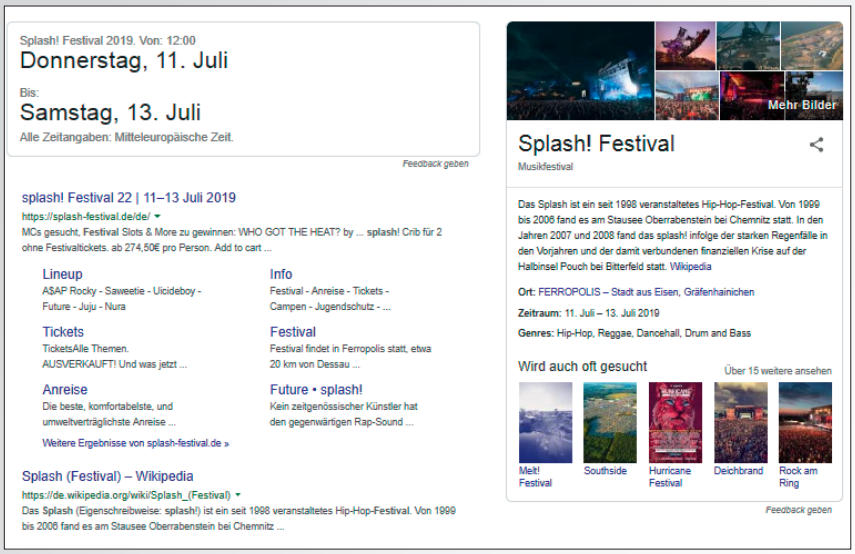

Abb. 12: Beispiel Knowledge Panel und Knowledge Card für die Suchanfrage „splash“

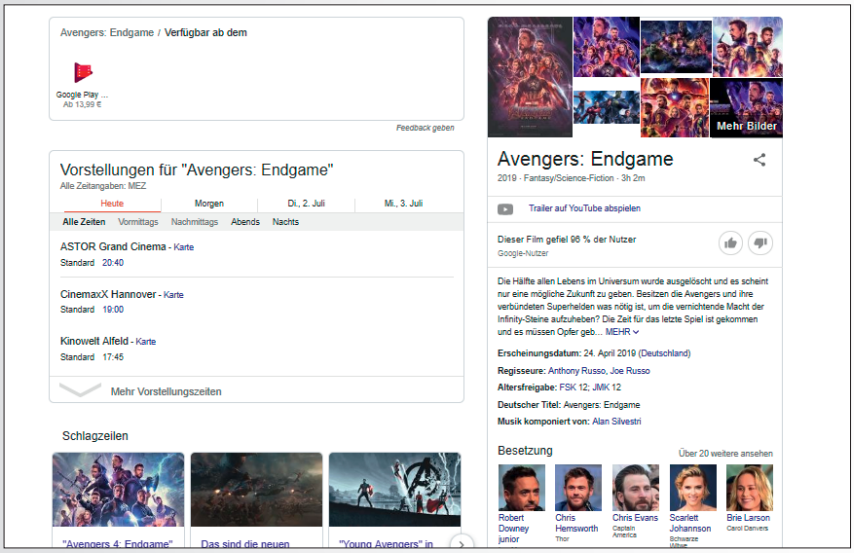

Abb. 13: Beispiel Knowledge Panel und Knowledge Card für die Entität „Avengers: Endgame“

» Besetzung
» Rollen
» Altersfreigabe

Bei Bauwerken oder Sehenswürdigkeiten sind es eher Attribute wie:

» Adresse
» Architektonische Höhe
» Baubeginn
» Öffnungszeiten
» Baukosten

Auch bei Suchanfragen bezüglich Attributen von Entitäten werden Boxen ausgeliefert, insofern die gesuchten Attribute im Knowledge Graph erfasst sind. So wird bei der Suchanfrage „wie hoch ist der eiffelturm“ oder „höhe eiffelturm“ eine Box wie in Abbildung 15 zu sehen ist ausgeliefert.

Nicht zu allen Entitäten werden Knowledge Panels ausgeliefert. Dazu aber an anderer Stelle mehr.

Frage-Antwort-Boxen wie z. B. bei der Suchanfrage „Angela Merkel“ sind mehr oder weniger prominent bei sehr vielen Suchtermen auf der ersten SERP platziert. Inwiefern die Frage-Antwort-Boxen auf Informationen aus dem Knowledge Graph basieren, ist nicht ganz klar, da sie nicht immer einen Entitäten-Bezug haben.

Während Frage-Antwort-Boxen sowohl bei benannten Entitäten als auch bei Themen bzw. Konzepten erscheinen können, findet man Featured Snippets in erster Linie, wenn nach Letzterem gesucht wird. Featured Snippets sind sehr selten in Bezug auf benannte Entitäten zu sehen.

Dennoch können bei Suchanfragen zu Themen und Konzepten sowohl Featured Snippets als auch Knowledge Panels erscheinen. Es kommt sogar vor, dass beide ausgeliefert werden, wie hier am Beispiel für die Suchanfrage „umwelt“.

Der Einfluss von Wikipedia auf Knowledge Panels ist größer als auf Featured Snippets. Bei Featured Snippets findet man häufig auch Informationen aus anderen Quellen. Doch für beide ist Wikipedia bisher die vertrauenswürdigste Datenquelle.

Auslieferung des Knowledge Panels bei gleichnamigen Entitäten

Eine interessante Metrik ist der Resultscore. Den Resultscore einer Entität zu einem bestimmten entitätsbezogenen Suchbegriff kann man über die Knowledge Graph API abfragen (s. o.).

Der Wert beschreibt, wie gut eine Entität z. B. zu einer entitätenbasierten Suche passt.

Dieser Score ist kein fester Wert. Er verändert sich laufend. Auf welcher Grundlage Google diesen Score ermittelt, ist unklar. Es kann sein, dass es mit der Popularität der Entitäten und/oder der Häufigkeit der Nennungen in den jeweiligen thematischen Kontexten bzw. den Kookkurrenzen zusammenhängt.

Der Automobilhersteller Jaguar hat mit 139.01976 einen deutlich höheren Resultscore als das Tier mit 70.045113. Das hat zur Folge, dass sowohl in den SERPs als auch im Knowledge Panel nahezu ausschließlich Suchergebnisse zur Automarke angezeigt werden.

Hier sieht man, dass Informationen aus dem Knowledge Graph durchaus auch einen Einfluss auf das Ranking der normalen Suchergebnisse haben können.

Die Entität mit dem höchsten Resultscore besetzt oft zuerst das klassische Knowledge Panel.

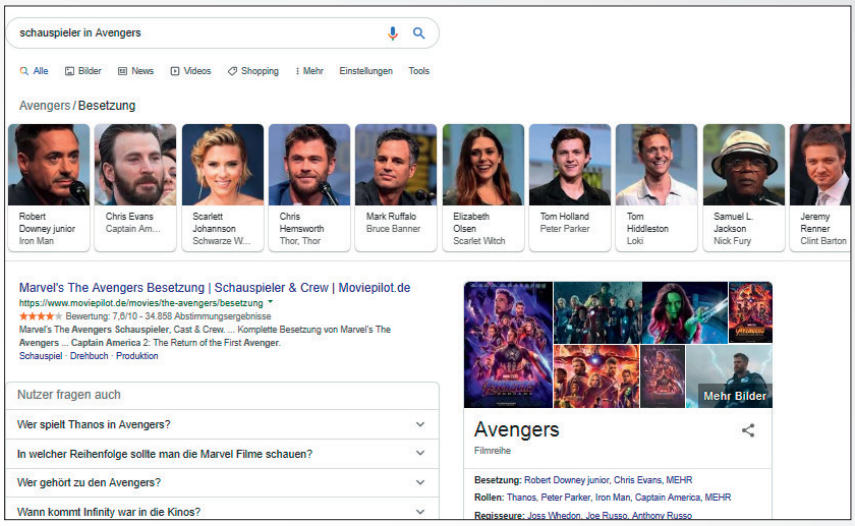

Abb. 14: Beispiel Entitäten-Karussell

Abb. 15: : Beispiel Antwort-Box für das Attribut „Architektonische Höhe“ für die Entität „Eiffelturm“

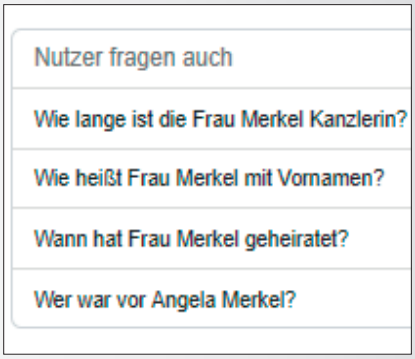

Abb. 16: Beispiel Frage-Antwort-Box für „angela merkel“

Doch nicht immer scheint der Resultscore entscheidend zu sein, für welche Entität Google das Knowledge Panel ausliefert. Im folgenden Beispiel ist der gesuchte mehrdeutige Begriff „heide“.

Im Knowledge Graph sind folgende Bedeutungen, Entitätstypen und Resultscores erfasst:

» Heidekräuter (Pflanze, Resultscore 80.790)
» Besenheide (Pflanze, Resultscore 73.955)
» Lüneburger Heide (Touristen-Attraktion, 53.314)
» Heidekräutergewächse (Pflanze, 52.968)
» Heide (Stadt, Ort, Verwaltungsbereich, 52.231)

Als Knowledge Panel wird trotz des geringeren Resultscores die Stadt ausgegeben. Es scheint neben dem Resultscore noch weitere Kriterien zu geben, für welche Entität ein Knowledge Panel ausgeliefert wird.

Verglichen mit den anderen Entitäten hat die Stadt Heide als einzige Entität eine offizielle URL. Das kann ein Grund sein. Zudem wäre es möglich, dass der Entitätstyp „Stadt“ wichtiger ist als der Entitätstyp „Pflanze“ oder „Touristen-Attraktion“. Generell scheint es so zu sein, dass bestimmte Entitätstypen wie Stadt, Orte … favorisiert als Knowledge Panel ausgegeben werden.

Wenn Google nicht sicher ist, welche Entität gesucht wird, werden zusätzlich zum Knowledge Panel kleine optionale „Spezifizierungsboxen“ mit weiteren Bedeutungen angezeigt. Darüber möchte Google herausfinden, ob man nach einer anderen Entität sucht.

Durch den Klick auf die Box spezifiziert man die Suchanfrage hinsichtlich Synonymen oder eines zusätzlichen Worts zur besseren kontextuellen Einordnung.

Welche Entitäten werden im Knowledge Graph aufgenommen und bekommen ein eigenes Knowledge Panel?

Eine der grundlegenden Fragen für SEOs ist, welche Entitäten in den Google Knowledge Graph aufgenommen werden. Im Knowledge Graph werden laut Google in erster Linie nur benannte Entitäten aus den Klassen der folgenden Entitätstypen erfasst.

» Bücher und Bücher-Serien
» Bildungseinrichtungen, Behörden, lokale Geschäfte, Unternehmen
» Ereignisse
» Filme und Film-Serien
» Musikgruppen und Alben
» Personen
» Orte
» Sport-Mannschaften
» TV-Serien
» Video-Spiele und -Serien
» Websites bzw. Domains

Für diese Entitätstypen liefert Google auch die Knowledge Panels aus. Nicht alle Entitäten aus diesen Klassen werden mit einem Knowledge Panel in den SERPs präsentiert. Die Entitäten müssen eine bestimmte gesellschaftliche Relevanz oder Autorität im jeweiligen Bereich besitzen.

Anhand welcher Kriterien Google diese Relevanz bewertet, ist nicht klar dokumentiert bzw. es fehlen konkrete Aussagen seitens Google. Aufgrund verschiedener Google-Patente und wissenschaftlicher Ausführungen sind einige Faktoren naheliegend.

Neben den benannten Entitäten können auch Themen oder Konzepte wie z. B. Online-Marketing, Biologie, Umwelt, Klima … als Entitäten im Knowledge Graph erfasst werden.

Der sicherste Weg, als Entität zu gelten, ist ein Eintrag bei Wikipedia oder Wikidata.

Wikidata ist eine Datenbank, in der strukturierte Informationen zu Entitäten erfasst werden. Wikidata ist das

Abb. 17: Beispiel Frage-Antwort-Box für „umwelt“

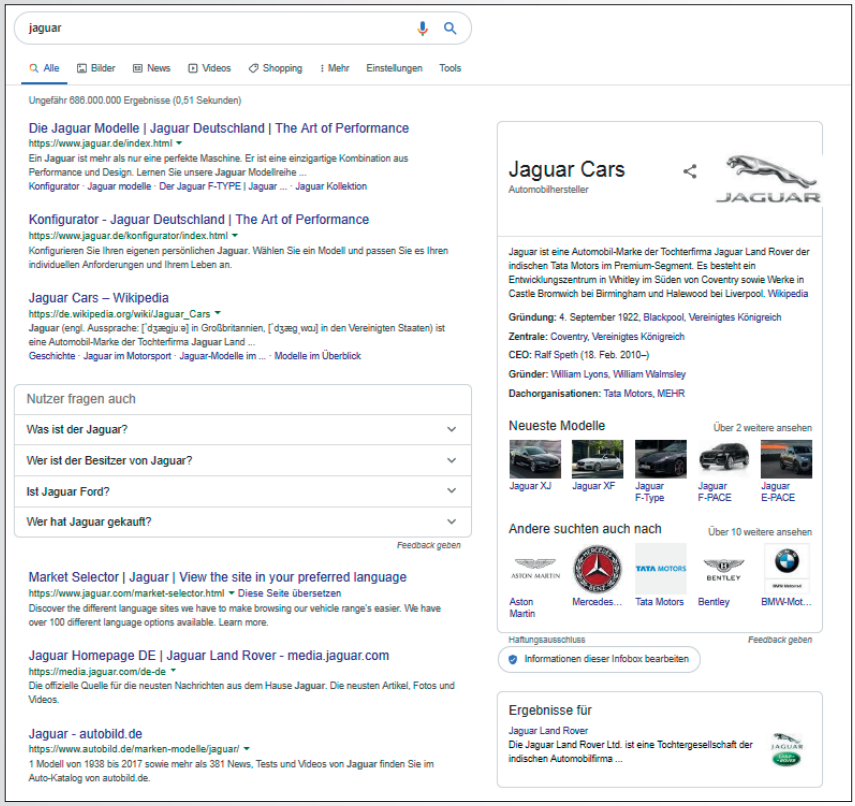

Abb. 18: Beispiel SERP für die Suchanfrage „jaguar“

Google-eigene Nachfolgeprojekt zu Freebase. Freebase wurde 2010 von Google gekauft. 2015 wurde ein Großteil der Freebase-Datensätze in Wikidata überführt.

Einen Wikidata-Eintrag kann jeder selbst anlegen. Allerdings ist dieser auch keine Garantie für ein Knowledge Panel. Nicht jeder Wikidata-Eintrag wird automatisch in den Knowledge Graph übernommen.

Zudem behalten sich die Moderatoren bei Wikidata vor, die Einträge zu prüfen und wieder aus der Datenbank zu löschen, wenn man nicht genug referenzierende Quellen im Wikidata-Eintrag angibt. Um Manipulation vorzubeugen, sollte ein Eintrag aus mindestens einer dritten Quelle verifiziert sein.

Viele SEOs schauen gerade auf Wikidata, wenn es um die Beeinflussung/Manipulation der Knowledge Panels geht. Dort lassen sich relativ einfach Einträge für Entitäten anlegen, die auch noch häufig den Weg in die SERPs über ein Knowledge Panel finden.

Hinsichtlich der Relevanz für den Knowledge Graph findet man bei Wikidata folgende Aussage (*http://einfach.st/wikid4*):

Während Freebase die offen zugängliche Grundlage für den Knowledge Graph war, gilt das nicht ebenso für Wikidata. Wikidata ist eine bestimmte Quelle des Knowledge Graph unter vielen, hat aber nicht die gleiche Stellung, wie Freebase hatte.

Wikidata scheint aber aktuell eine durchaus relevante Quelle zu sein, wie einige Tests von SEO-Kollegen zeigten, die über Wikidata relativ einfach Knowledge Panels erzeugen konnten.

Da die Manipulationsgefahr dadurch steigt, könnte Wikidata eine Zwischenlösung sein, bis der Knowledge Vault nahezu fehlerfrei funktioniert.

Während in Wikidata eher stichpunktartig Attribute einer Entität zugeordnet werden, beschreibt Wikipedia die Entität in einem ausführlichen Text. Ein Wikipedia-Beitrag ist damit die ausführliche Beschreibung einer Entität und stellt als externes Dokument eine wichtige Quelle für den Knowledge Graph dar.

Wikipedia-Beiträge spielen aktuell bei vielen Knowledge Panels und Featured Snippets eine übergeordnete Rolle als Quelle für Informationen und werden von Google als Beweis für die Relevanz einer Entität genutzt.

In einem wissenschaftlichen Projekt mit dem Titel „A Framework for Benchmarking Entity-Annotation Systems“, an dem auch ein Google-Mitarbeiter beteiligt war, wird eine Entität gleichgesetzt mit einem Wikipedia-Beitrag (*http://einfach.st/googresearch4*).

An entity (or concept, topic) is a Wikipedia article which is uniquely identified by its page-ID.

Ein Eintrag bei Wikipedia bleibt allerdings den meisten Unternehmen und Personen verwehrt, da ihnen die gesellschaftliche Relevanz in den Augen der Wikipedianer fehlt.

Dass Google in Sachen Knowledge Panel so restriktiv ist, zeigt, wie vorsichtig Google mit dieser exponierten Fläche in den SERPs umgeht. Massenhafte Manipulation der Knowledge Panels könnte die Nutzererfahrung deutlich

beeinträchtigen und dem Image von Google als Suchmaschine empfindlich schaden.

Hinsichtlich des Ziels einer nahezu vollständigen und fehlerfreien Wissens-Datenbank, die Informationen zu allen Entitäten der Welt beinhaltet, reicht dieser Ansatz allerdings nicht aus.

Inwiefern Google neben den über die Knowledge Panels ersichtlichen Entitäten weitere Entitäten im Knowledge Graph erfasst, ist unklar. Eine Vermutung ist, dass die bisher sichtbaren Knowledge Panels nur die Spitze des Eisbergs sind. Google ist aktuell über den Knowledge Vault via Natural Language Processing in der Lage, Entitäten, Attribute und Beziehungen aus unstrukturierten Informationen aus frei verfügbaren Online-Quellen wie Websites automatisiert zu extrahieren.

Sobald Google einen Weg gefunden hat, die Validierung der Informationen rund um weitere Entitäten mit einer vertretbaren Fehlertoleranz ohne manuelle Prüfung zu bewerkstelligen, wird es auch unabhängig von Wikipedia und Wikidata Knowledge Panels in den SERPs zu sehen geben.

Deswegen bleibt abzuwarten, inwieweit Google das Tor zum Knowledge Graph insbesondere für Unternehmen und Einzelpersonen zukünftig offener gestaltet.

Wie groß ist der Einfluss von Entitäten und Knowledge Graph auf die SERPs und das Ranking von Inhalten?

Der Einfluss des Knowledge Graph auf die SERPs ist seit 2013 stetig angestiegen. Immer mehr SERP-Elemente, die auf Informationen aus dem Knowledge Graph basieren, wie Knowledge Panel und Ergänzungs-Boxen, wurden über die Jahre in den SERPs ergänzt. Dadurch verlieren die klassischen Suchergebnisse immer weiter an Sichtbarkeit.

Abb. 19: Beispiel Spezifizierungs-Box für die Entität „bing“

Zudem spielen Knowledge-Graph-Elemente wie z. B. die Featured Snippets eine zentrale Rolle bei der Antwort-Ausgabe digitaler Assistenten.

Neben den augenscheinlichen Einfluss durch die neuen SERP-Elemente bleibt die Frage, inwieweit der Knowledge Graph und die dort erfassten Entitäten eine Rolle für die Indexierung und das Ranking von Inhalten spielen.

Stand heute scheint Google durch semantische Strukturen wie Knowledge Graph oder Machine Learning dem durch die ehemalige VP Marissa May formulierten Ziel sehr nahe zu sein: Weg von einer rein keywordbasierten Suchmaschine zu einer konzeptionell bzw. kontextbasierten Suchmaschine bzw. Antwortmaschine.

Right now, Google is really good with keywords and that's a limitation we think the search engine should be able to overcome with time. People should be able to ask questions and we should understand their meaning, or they should be able to talk about things at a conceptual level. We see a lot of concept-based questions — not about what words will appear on the page but more like „what is this about?" (http://einfach.st/infoworld2)

Das ist auch höchste Zeit – wenn man bedenkt, dass die Voice Search auf dem weltweiten Vormarsch ist und die Komplexität der Suchanfragen dadurch immer größer wird.

Wie groß der Einfluss von Entitäten auf den klassischen Such-Index und das Ranking aktuell ist, kann man nur schwer sagen.

Dazu gibt es unterschiedliche Perspektiven.

Die geschätzten amerikanische Kollegin Cindy Krum vermutet, dass der Mobile Index auf den Informationen aus dem Knowledge Graph beruht, weshalb sie ihm den Namen Entity First Index gibt. Hier sollen die Inhalte bzw. Dokumente und Entitäten, die mit der Haupt-Entität in Beziehung stehen, der Haupt-Entität untergeordnet und danach in eine Entitäten-Hierarchie gebracht werden. Die Beziehungen der Elemente untereinander werden nicht mehr basierend auf einem Link Graph, sondern basierend auf dem Knowledge Graph hergestellt. Der Link Graph in der bisherigen Form wäre aufgrund der steigenden Anzahl an Inhalten und verschiedenen Plattformen irgendwann nicht mehr skalierbar, so Krum *(http://einfach.st/mobilemoxi)*.

Cindys Theorie kann mit Blick auf die Zukunft ein durchaus realistisches Szenario sein. Aufgrund diverser Beobachtungen und Untersuchungen scheint aktuell der Knowledge Graph neben dem Such-Index noch parallel zu existieren.

In einigen Google-Patenten wird auch immer wieder von einer Entity-Database geschrieben, die neben einem Search-Index existiert. Diese Entity-Database ist in Bezug auf Google offensichtlich der Knowledge Graph.

So heißt es im Google-Patent Entity database data aggregation *(http://einfach.st/gpat44)*:

an entity database storing an entity-relationship graph representing elements in the virtualization environment, wherein:

» *each of the elements is represented by an entity-type node in the entity-relationship graph,*
» *relationships between the elements are represented by edges between the nodes, and*
» *information regarding each of the entity-type nodes is accessible through a query interface.*

Vorstellen kann man sich eine Art Schnittstelle zwischen dem Knowledge Graph und dem Such-Index, über die wechselseitig Informationen zu Entitäten ausgetauscht bzw. in Beziehung gebracht werden.

In dieser Entitäten-Inhalts-Schnittstelle geht es darum, herauszufinden:

» Ob in einem Inhalt Entitäten vorkommen
» Ob es eine Hauptentität gibt, von der der Inhalt handelt
» Welcher Ontologie oder welchen Ontologien die Hauptentität zugeordnet werden kann
» Welchem Urheber bzw. welcher Entität der Inhalt zuzuordnen ist
» In welcher Beziehung die im Inhalt vorkommenden Entitäten zueinander stehen
» Welche Eigenschaften bzw. Attribute den Entitäten zuzuordnen sind

Zudem gibt es immer wieder Aussagen seitens Google, dass der Link Graph immer noch die zentrale Rolle beim Ranking spielt und nicht der Beziehungsgraph zwischen Entitäten, wie Cindy vermutet.

Dass es eine Verbindung zwischen dem Knowledge Graph und dem klassischen Such-Index geben muss, zeigt das bereits erwähnte Jaguar-Beispiel. Bei Suchanfragen bezüglich gleichnamiger Entitäten muss Google herausfinden, welches die am wahrscheinlichsten gesuchte Entität ist und was dementsprechend der passende Dokumenten-Korpus bzw. ggf. die passenden Korpusse sind. Ein Zusammenspiel zwischen Knowledge Graph und Such-Index ist unabdingbar.

TIPP

Wer noch tieferes Interesse am Thema Semantik, Natural Language Processing und Entitäten für SEO hat, findet weitere ausführliche Beiträge des Autors unter *http://einfach.st/koppseo.*

Mehr deutlich sichtbare Zusammenhänge sind bisher nicht erkennbar, aber da kann noch einiges passieren.

Die Auswirkungen von Knowledge Graph und Hummingbird auf das Ranking der klassischen Suchergebnisse sind nur schleichend wahrzunehmen, da Google nur langsam beim Verstehen der Bedeutung einzelner Entitäten vorwärtskommt. Das Verständnis von Entitäten geschieht top-down nach Relevanz. Und die relevantesten Entitäten sind in Wikidata bzw. Wikipedia erfasst.

Wikipedia und Wikidata bilden allerdings aufgrund der manuellen Pflege nur einen Bruchteil der Entitäten und zugehörigen Attribute ab.

Wenn Google das Spannungsfeld zwischen Vollständigkeit und Richtigkeit in den Griff bekommt, wird es wahrscheinlich zu einem Paradigmenwechsel auch beim Ranking kommen. Hier spielen Googles Fortschritte beim Machine Learning und Natural Language Processing eine entscheidende Rolle, um das automatisiert oder teilautomatisiert erledigen zu lassen.

Wenn es so weit ist, werden wir uns bei der Suchmaschinenoptimierung immer mehr mit Entitäten, Attributen und den Beziehungen dieser zueinander beschäftigen als mit Keywords und der Optimierung einzelner Dokumente.

In den in den letzten Jahren veröffentlichten wissenschaftlichen Papieren und Patenten von Google liest man deutlich häufiger über Entitäten als z. B. über Keywords. Es vergeht kein Monat, in dem kein neues Google-Patent in Bezug auf Entitäten das Licht der Welt erblickt. Das zeigt deutlich, in welche Richtung sich Google entwickelt.

Was bedeutet das für die Suchmaschinenoptimierung (SEO)?

Bisher gibt es relativ wenig Informationen seitens Google darüber, wie man hinsichtlich Entitäten und Knowledge Graph optimieren kann. Auch bei den Webmaster Hangouts mit John Müller spielen die Themen keine große Rolle. Die einzigen Quellen dazu sind bisher technische Anleitungen für die Nutzung der Natural Language API oder Knowledge Graph API, Google-Patente oder Dokumentationen wissenschaftlicher Arbeiten.

Aufgrund der Informationen aus den vielen Quellen der letzten Jahre kann man hinsichtlich SEO zu den folgenden Schlüssen kommen:

» Jeder SEO sollte sich mit den Themen Knowledge Graph, Entitäten, Natural Language Processing und Word Embeddings beschäftigen. Diese Themen werden zukünftig die wichtigsten Grundlagen für moderne Suchmaschinenoptimierung sein.
» Aktuell ist die einzige selbstbestimmte Möglichkeit, als Entität im Knowledge Graph erfasst zu werden, ein Wikidata- oder Wikipedia-Eintrag. Da die Relevanzkriterien für die Aufnahme in der Wikipedia höher sind als für Wikidata, kann man es zuerst dort probieren. Ich empfehle hier, gewissenhaft vorzugehen und sich kritisch zu hinterfragen, ob man wirklich so relevant ist, wie man denkt. Auch bei Wikidata werden die Einträge von Moderatoren überprüft und sie nehmen sich auch die Freiheit, diese wieder zu entfernen oder Accounts zu sperren. Ein Wikidata-Eintrag ist zudem keine Garantie für ein Knowledge Panel. Nur wer die Relevanz der angelegten Entität belegen kann, wird ein Knowledge Panel bekommen.
» Sorg dafür, dass Google dich als Entität erkennt. Dafür sollte man den Namen der Marke, des Unternehmens

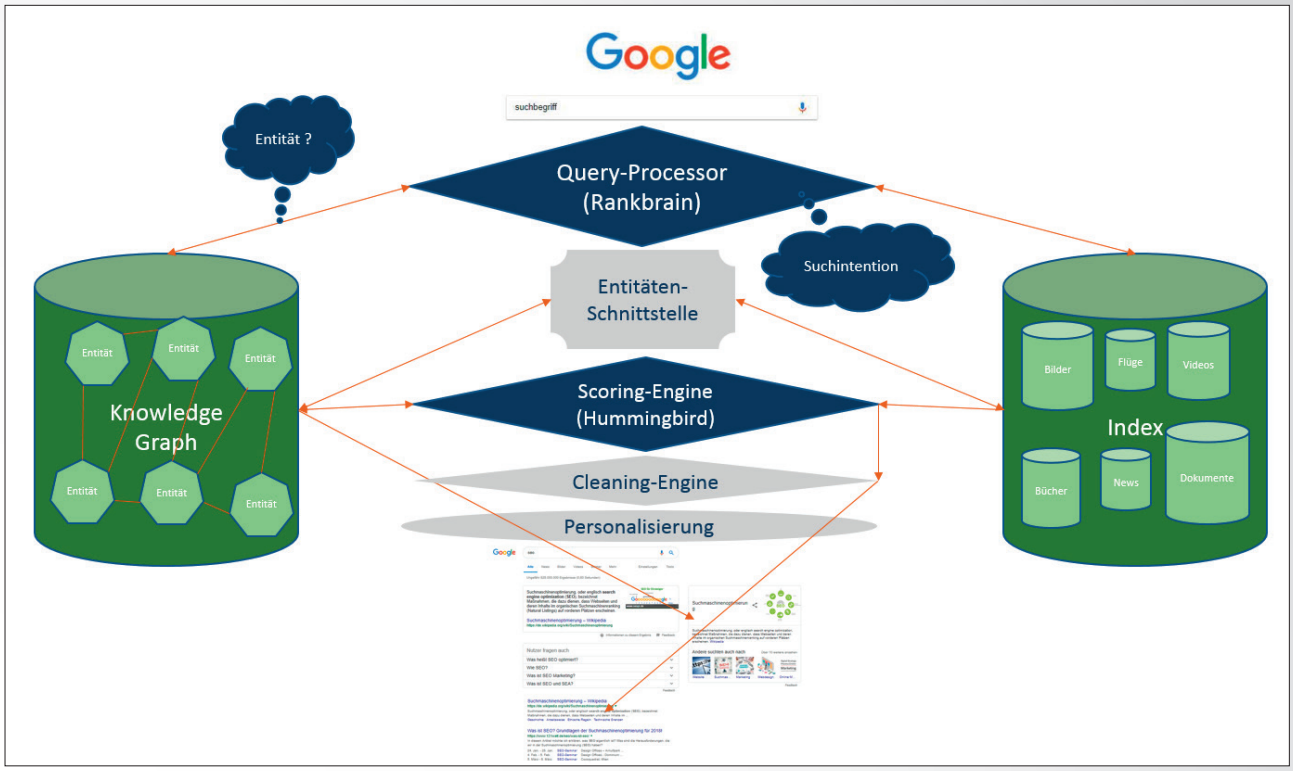

Abb. 20: Mögliche Funktionsweise von Google *(Quelle: © Olaf Kopp, Aufgesang GmbH)*

so häufig wie möglich klar ersichtlich als Subjekt in Sätzen verwenden. Dadurch fällt es leichter, den Namen als Entität via NLP zu identifizieren.

- Da man davon ausgehen kann, dass Google zukünftig immer mehr Natural Language Processing zur Interpretation von Texten nutzt, sollten einige Empfehlungen bei der Texterstellung berücksichtigt werden:
  - Bei der Erstellung von Texten sollte man grammatikalisch klar und einfach schreiben, also z. B. auf Schachtelsätze und viele Nebensätze verzichten. Damit tut man dem Leser einen Gefallen und die Bedeutung lässt sich besser über NLP interpretieren.
  - Bei der Erstellung von Texten sollte man möglichst auf Personalpronomen verzichten und die Entität beim Namen nennen, damit Google eindeutig versteht, was gemeint ist.
  - Bei der Erstellung sollten Adverbien und Adjektive nur genutzt werden, wenn sie auch wirklich wichtig für das Verständnis eines Satzes sind.
  - Zusammengefasst sollte auf Geschwafel und Bla, bla … verzichtet werden, um Google und Lesern einen Gefallen zu tun.
- Das semantisch passende Umfeld rund um die in einem Text beschriebenen Entitäten macht es Suchmaschinen möglich, diese besser zu erfassen und zu deuten. Darüber können Google & Co auch den Inhalt thematisch einordnen. Wenn viele Kookkurrenzen zwischen einer Hauptentität mit anderen Entitäten, Attributen … vorkommen, ist es eine gute Empfehlung, diese auch in den eigenen Inhalten zu verwenden. TF-IDF-Analysen sind hier ein probates Mittel, um Begrifflichkeiten zu identifizieren, die das semantische Umfeld einer Entität bzw. eines Themas beschreiben.
- Verknüpfe deine Repräsentanzen wie Domains, Apps, Kanäle auf YouTube und Social Media miteinander, um sie als digitale Abbilder deiner Entität zusammenzuführen. Als Person kann man auch auf die Autorenprofile auf anderen Medien verweisen (wenn vorhanden).
- Mit Blick auf die Zukunft sollten Personen und Unternehmen an der eignen Entitäten-Relevanz arbeiten. Das bedeutet, Marketing und PR mit Blick auch auf mögliche Signale für Suchmaschinen zu betreiben. Signale können sein:
  - Erwähnungen und Verlinkungen in branchenrelevanten Medien
  - Steigende Anzahl an Brand- und Navigational-Suchen
  - Steigende Klickrate auf den eigenen Suchergebnissen
  - Inhalte auf der eigenen Website, die klar signalisieren, für welche Themen man Experte ist, und die positives Feedback z. B. über Kommentare und soziale Medien bekommen

Dabei sollte, wenn möglich, immer auf Kookkurrenzen von Brand mit Thema geachtet werden. Dadurch kann Google zum einem die passenden Vektorräume für unsere Inhalte und die Entität identifizieren und zum anderen die Nähe bzw. Autorität und Relevanz zu bestimmten Themen und zugehörigen Keyword-Gruppen ermitteln. **Zusammengefasst: Werde eine Marke und eine Autorität!**