Bild: KI Midjourney | Bearbeitung: c't

Basis für tausend Suchmaschinen

Die EU will bis 2025 einen öffentlichen Web-Index aufbauen

Muss es denn immer Google oder Bing sein? Mit dem Projekt OpenWebSearch will die EU ihre Souveränität im Internet schützen. Das Ziel ist ein frei zugängliches Webverzeichnis, das diverse Suchmaschinen und Sprachmodelle füttert und einen Boom neuer Webdienste auslösen soll.

Von Arne Grävemeyer

Wer im Web etwas sucht, der googelt in den allermeisten Fällen, seltener befragt er Microsofts Suchmaschine Bing. Einige kleinere Suchmaschinen generieren ihre Ergebnislisten zu einem wesentlichen Teil mithilfe der Web-Indexe von Google (Startpage) oder von Bing (DuckDuckGo, Ecosia, MetaGer, Neeva, Qwant, You.com). Nachdem Yahoo in Bing aufgegangen ist, gibt es weltweit neben den beiden amerikanischen Suchindexen nur noch zwei weitere Angebote in einer vergleichbaren Größenordnung. Das sind Yandex aus Russland sowie Baidu aus China, und beide haben in der westlichen Welt eine deutlich geringere Relevanz.

Wie groß die Abhängigkeit von Google und Bing mittlerweile ist, hat Microsoft den Suchmaschinenbetreibern gerade vorgeführt. Im Februar hat der Konzern seine Abomodelle für das Bing-API deutlich verteuert, je nach Nutzungsmodell auf das Drei- bis Zehnfache der bisherigen Kosten. Zudem warnen Wissenschaftler wie der amerikanische Psychologe Robert Epstein bereits seit 2015 vor dem Search Engine Manipulation Effect, wonach marktbeherrschende Suchmaschinen

durch ihr Ergebnis-Ranking in der Lage sind, die Meinungsbildung in der Demokratie zu beeinflussen [1].

Was könnte man aber mit einem großen Web-Index alles machen, wenn dieser öffentlich frei verfügbar wäre? Man könnte alternative Suchmaschinen aufbauen oder spezialisierte Suchdienste nach ausgewählten Themen. Anwender hätten die freie Wahl und könnten ihre privaten Nutzerprofile besser schützen. Sprachforscher könnten in dem Datenpool eines großen Web-Index verfolgen, wie sich unsere Sprache entwickelt, und Soziologen könnten beobachten, wie wir in den sozialen Medien miteinander umgehen. Webdienste könnten darin Hinweise auf beginnende Pandemien oder andere Katastrophenfälle suchen und damit ein Frühwarnsystem aufbauen.

Datenpool beflügelt KI-Innovationen

Zudem eröffnet ein umfangreicher, auf europäische Quellen fokussierter Index die Chance, neue Sprachmodelle zu entwickeln. Davon könnten insbesondere kleinere Sprachen oder ausgewählte Sprachkombinationen wie etwa Tschechisch und Slowakisch profitieren, die bei den global ausgerichteten US-Diensten leicht durchs Raster fallen. Tatsächlich sind die Möglichkeiten, die ein solcher multilingualer Web-Index eröffnet, noch gar nicht vollständig ausgelotet. Die Wissenschaftler sehen aber gute Chancen, dass er innovative Services sowie neue Forschungsthemen nach sich zieht und einen Boom rund um neue Sprachmodelle auslöst.

Das sieht mittlerweile auch die Europäische Kommission so, die zudem mehr digitale Souveränität anstrebt. Im September 2022 startete das EU-Projekt OpenWebSearch (Link siehe ct.de/y6sw). Ziel ist der Aufbau eines öffentlich zugänglichen Open Web Index (OWI) und einer nachhaltigen, also auf Dauerbetrieb angelegten Infrastruktur. In drei Jahren soll ein Web-Index entstehen, der immerhin die Hälfte aller im Internet veröffentlichten Texte verzeichnet. Die beteiligten Partner rechnen dafür mit einem Speicherbedarf von etwa fünf Petabyte (fünf Millionen Gigabyte). Im Vergleich zu den Indexen von Google oder Bing wäre das zunächst ein vergleichsweise kleiner Datenpool. Die etablierte Konkurrenz kommt mit Texten, Bilddateien, Multimedia, Nutzungsdaten und Logfiles aus dem Internet auf jeweils Hunderte von Petabyte.

„Wir sind kein europäisches Google“, sagt Michael Granitzer, Inhaber des Lehrstuhls für Data Science an der Universität Passau, der das OpenWebSearch-Projekt koordiniert. Es gehe bei dem Projekt nicht um den Aufbau einer großen Suchmaschine, sondern viel grundlegender darum, eine Infrastruktur zu etablieren, mit der später Suchmaschinen und andere Dienste arbeiten können. Googles Größe ist am Anfang sicher unerreichbar. „Es wird eher wie bei Wikipedia sein, die im Vergleich zu großen Verlagen zunächst mit einem kleinen Kern startete und dann kontinuierlich wuchs.“

14 Projektpartner entwickeln derzeit Crawling-Techniken, wählen Metadaten aus, die der Index zusätzlich aufnehmen soll, und konzipieren eine dezentrale Aufteilung des OWI auf verschiedene Server und Standorte in Europa. So beteiligen sich etwa Infrastrukturpartner wie das Leibniz-Rechenzentrum in München, das CSC in Espoo, Finnland, das Europas größten Supercomputer betreibt, das tschechische National Supercomputing Center IT4Innovations sowie das CERN bei Genf.

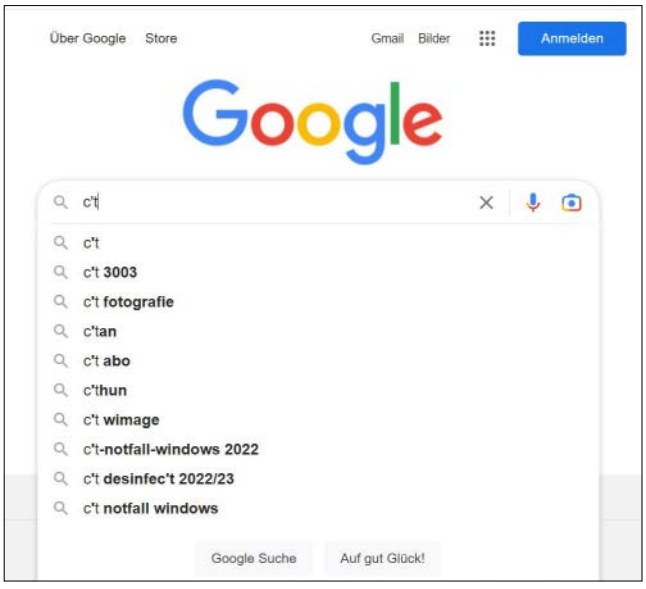
Höfliche Crawler

Die eingesetzten Webcrawler sollen „höflich“ vorgehen. „Im Web verursacht Crawling etwa 30 bis 40 Prozent der gesamten Netzlast, wenn man mal vom Bereich Streaming absieht“, sagt Stefan Voigt, Vorstand der deutschen Open Search Foundation, die den Anstoß zum OpenWebSearch-Projekt gegeben hat. Das ist für Webhoster ein erheblicher Kostenfaktor. Hinzu kommt, dass man eine Website leicht überlasten kann, wenn man alle HTML-Seiten mit mehreren Servern parallel abruft. Aufgrund eines Implementierungsfehlers passierte dies tatsächlich einmal, als Forscher aus dem Projekt die Seiten der Bank of America crawlten. Die Bank missverstand dies als eine Denial-of-Service-Attacke und blockte die beteiligten Server. Mit sogenannten Crawling Politeness Rules will man solche Missverständnisse prinzipiell vermeiden und zum Beispiel niemals mehr als eine Anfrage pro Sekunde starten sowie selbstverständlich die Seitenbeschränkungen der robots.txt-Datei beim Website-Betreiber einhalten. Damit können Website-Betreiber Bots dazu auffordern, bestimmte Bereiche ihrer Site nicht aufzusuchen.

Schon bei der Analyse der Crawls offenbart sich ein weiterer Nachteil gegenüber dem marktbeherrschenden Index von Google: Das robuste Parsing von HTML-Dateien ist ein Problem, das sich nicht einfach technisch lösen lässt, sondern nur in Zusammenarbeit mit den Web-

c't kompakt

- Das EU-Projekt OpenWebSearch will einen öffentlichen großen Web-Index aufbauen.
- Dieser soll vielen neuen Suchmaschinen und Webdiensten offenstehen.
- Als Nebeneffekt bietet der neue Index einen Datenpool speziell für europäische Sprachmodelle.



Ob auf Smartphone, Tablet oder im Desktop-Browser: Die Google-Suchmaske ist mit weitem Abstand vorherrschend und hindert so kleinere Dienste, die Websuche mit neuen Ideen weiterzuentwickeln.

site-Betreibern. Google bietet ihnen dafür die Google Search Console, ein Analysetool, mit dem sie selbst eine Suchmaschinenoptimierung (Search Engine Optimization, SEO) betreiben und somit ihre Seiten speziell an den Google-Bot anpassen können. Dort sehen Webmaster auf einen Blick: Steht der Haupttext einer Webseite dort, wo Google ihn erwartet? Entsteht bei der Analyse ein vollständiger Parse-Baum? Welche Keywords erkennt der Google-Bot auf den Webseiten? Indem heute die meisten Betreiber ihre Websites für den Google-Index optimieren, hat der Marktführer seine Parsing-Probleme erfolgreich auf sie abgewälzt.

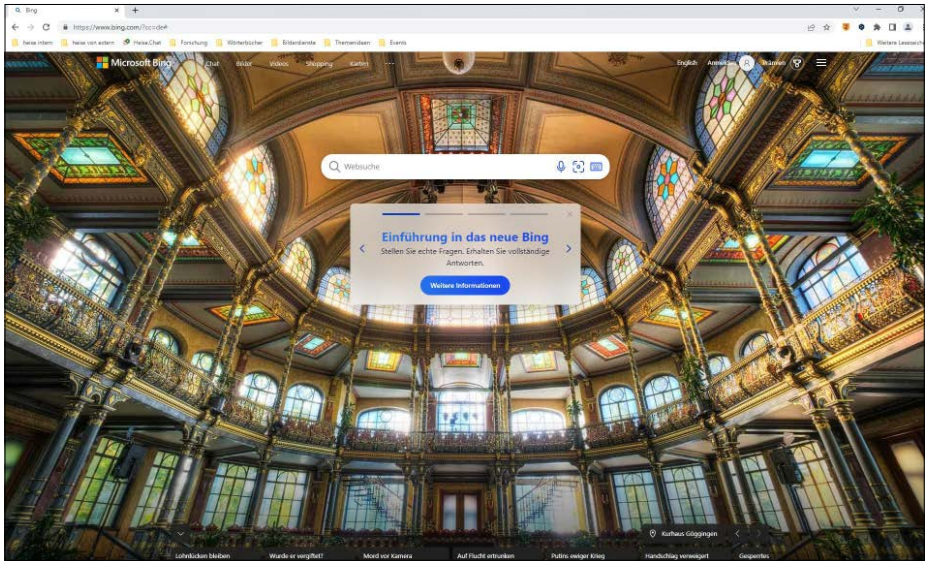
Ein neuer Web-Index, dessen Crawler sich nicht als Google-Bots ausgeben, und der auf andere Metadaten setzt, befindet sich demgegenüber zu Anfang klar im Nachteil. Daher ist vergleichbar zur Google Search Console eine Open Console für den OWI geplant. Um deren Akzeptanz bei den Website-Hostern zu erhöhen, weisen die Projektpartner darauf hin, dass sowohl der OWI als auch die Open Console alle Daten nach EU-Recht speichern und schützen. Für Hoster bedeutet das, dass sie jederzeit einsehen können, welche Daten über eine Website gesammelt worden sind. Damit können sie dann auch Rechtsansprüche auf Korrektur oder Löschung durchsetzen.

Lieber selbst crawlen

Mit einer weiteren Funktion könnte die Open Console den Website-Verantwortlichen entgegenkommen: Die sollen mit dem neuen OWI vereinbaren können, bei Änderungen ihre Seiten selbst zu crawlen und die Daten im WARC-Format (Web-ARChive) bereitzustellen. Auf diese Weise bestimmen sie selbst, zu welchen Zeitpunkten und in welchem Rhythmus der Crawling-Traffic auf ihren Servern läuft. Darüber hinaus steht zu hoffen, dass bei einem wachsenden OWI viele Spezial-Crawler mit ihren Abfragen nicht mehr die bereits erfassten Server der Websites belasten, sondern stattdessen einfach im öffentlich zugänglichen Web-Index suchen.

Zusätzlich sollen Hoster mit einer Website Registry Auskunft über Nutzungsrechte geben können. Sind Inhalte beispielsweise unter CC0-Lizenz frei verfügbar, darf man sie für kommerzielle Zwecke verwenden? Handelt es sich um einen fachlich strukturierten Text, enthält er persönliche Daten oder gibt er Meinungen wieder? Angaben zu solchen Fragen sind nicht nur interessant für Informationssuchende, sondern auch wertvoll für Entwickler von Sprachmodellen. Denn um die zu trainieren, benötigen sie eine gut gepflegte Datenbasis. Texte, die etwa tendenziöse Meinungen transportieren, führen im Endeffekt wieder zu Sprach-KIs mit Vorurteilen: dem schon häufig beobachteten, sogenannten Bias.

Langfristig könnte ein öffentlicher, transparent aufgebauter Web-Index die SEO-Landschaft komplett verändern. Heute versucht fast jeder Betreiber, seine Website für das Ranking in der Google-Suche zu optimieren. Wenn anstelle dieses Monopols eine Vielzahl von spezialisierten Suchdiensten treten sollte, würde sich die Situation grundlegend ändern. „Aus meiner Sicht sollten sich die Webseiten-Betreiber und die Web-User auf die Inhaltserstellung konzentrieren“, sagt Christian Gütl, Leiter des Cognitive and Digital Science Lab an der TU Graz. Wenn die Websuche nicht mehr durch einen Monopolisten beherrscht werde, komme es nur noch darauf an, Inhalte möglichst strukturiert aufzubereiten und mit aussagekräftigen Metadaten zu unterstützen.



Microsofts Suchmaschine Bing greift mit ChatGPT-Unterstützung den Marktführer Google an und hat gleichzeitig das Bing-API deutlich verteuert.

Keine Nutzeranalyse

Noch offen und intensiv diskutiert ist die Auswahl der Suchfaktoren, die der neue Web-Index aufnehmen soll. Außer den Nutzungsrechten für Inhalte diskutieren die Partner vor allem inhaltliche Faktoren wie die Inhaltsqualität, Genres und eine Pagerank-Gewichtung der Linkstrukturen. Hinzu kommen technische Faktoren wie etwa die Antwortzeiten. Spannende Forschungsprojekte betreffen die Idee der Georeferenzierung, die Webdienste für einzelne Städte oder Gemeinden ermöglicht. Allerdings erfordert die räumliche Einordnung von Weblinks zunächst eine semantische Analyse, denn schließlich ist Paris Hilton etwas anderes als das Hilton in Paris.

Ein sehr wichtiger Unterschied zu den heutigen großen Indexen: User-Clicks und weitere Nutzeranalysen finden keinen Eingang in den OWI. Können sie auch kaum, denn die Projektpartner planen nicht, selbst Suchmaschinen zu betreiben und damit Nutzerdaten einzusammeln.

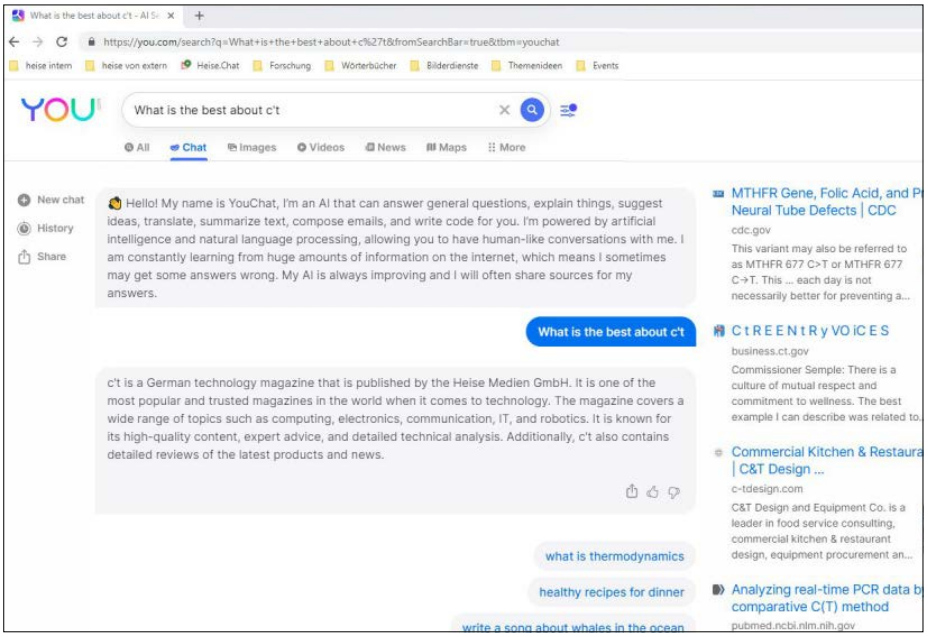
„Allerdings können Webdienste, die auf dem OWI aufbauen, jederzeit die Suchbegriffe und das Nutzerverhalten auf ihrer Plattform mitloggen“, erläutert Phil Höfer, technischer Beirat des gemeinnützigen Vereins Suma e. V., der die Metasuchmaschine MetaGer betreibt. Das sei in gewissem Rahmen sogar erforderlich, um beispielsweise Massenabfragen von Spammern abblocken zu können oder den eigenen Dienst zu analysieren und zu verbessern. Auf der anderen Seite werde der Anwender aber in einem vielfältigen Diensteangebot frei entscheiden können, wem er seine Nutzungsdaten anvertraut.

Sicher verteilt

Als eine für Europa kritische Infrastruktur ist es geplant, den OWI auf mehrere Rechenzentren aufzuteilen. So können sich ähnlich wie Nationalbibliotheken in ver-

schiedenen Ländern regionale Schwerpunkte ausbilden, etwa nach Herkunftssprachen. Zudem könnten verschiedene Organisationen die Pflege verschiedener Teilindexe übernehmen und diese hosten, sei es etwa ein Index für Geowissenschaften oder ein Index für Finanzmärkte. Das CERN beispielsweise zeigt sich schon interessiert, alle Informationen zu Partikelphysik aufzubereiten und zu verwalten.

Das Projekt soll nicht ausschließlich in wissenschaftlicher Hand bleiben. „Es gibt Projektmittel und wir wollen durchaus auch wirtschaftliche Akteure ins Boot holen“, sagt Höfer. Es brauche nicht nur Forschergeist, sondern auch unternehmerisches Gespür und Geschäftsideen, um beispielsweise neue Special-Interest-Suchmaschinen und -Dienste auf Basis des OWI zu entwickeln. Denn nur wenn sich die Infrastrukturkosten mittelfristig durch Lizenzeinnahmen decken, kann der Web-Index nachhaltig wachsen.



Das Suchportal You.com bot als eines der ersten einen KI-Chatbot, der Suchanfragen aufnimmt und Ergebnisse vorstellt. Fortgeschrittene Sprachmodelle sollen diese Technik verbessern.

Pool für Sprachmodelle

Bereits zum Projektstart und damit noch vor dem Hype um ChatGPT betrachteten die Partner den Open Web Index mit seinem Fokus auf europäische Inhalte und Sprachen als einen Datenpool für spezialisierte Sprachmodelle. Neue Suchmaschinen könnten diese Modelle auch gleich als Schnittstelle für Suchanfragen einsetzen. „Die Benutzer suchen in der Regel nicht Links, sondern Antworten auf ihre Fragen oder sogar Lösungsvorschläge“, sagt Gütl. Das spreche für den Einsatz von Chatbots.

Die Bing-Suchmaschine mit ihrer ChatGPT-Schnittstelle ist offenbar gerade dabei, diesen Forschungsansatz rechts zu überholen. Projektkoordinator Granitzer sieht daneben vor allem die Suchmaschine You.com als ein gelungenes Beispiel, wie ein KI-Chatbot Suchanfragen aufnehmen und Ergebnis-Links beschreiben kann. Allerdings könne man sich auf die Antworten der KI derzeit nicht verlassen. Zudem seien die Kosten und die Skalierbarkeit solcher Systeme noch zu klären. Trotzdem denkt Granitzer, dass Suchmaschinen in Zukunft Sprachmodelle als Schnittstelle integrieren müssen. Denn diese Technik bietet eine Möglichkeit, Anfragen in natürlicher Sprache oder in Form eines Gesprächs mit einem Chatbot zu verarbeiten und Ergebnisse zusammenzufassen. „Damit sind sie das bessere Interface, wenn es zum Beispiel darum geht, sich erstmal einen Überblick zu verschaffen.“

Neuartige Webdienste

Das OpenWebSearch-Projekt sucht nach weiteren Projektpartnern, die ihre Expertise einbringen könnten. So gibt es bereits Gespräche mit spezialisierten Diensten wie etwa einem Wortschatzlexikon, das mithilfe eines aktuellen, frei zugänglichen Web-Index automatisiert und sozusagen live erkennen soll, wenn neue Wortkreationen, Metaphern und Redewendungen entstehen. Ein anderes Projekt namens Europe Media Monitor beschäftigt sich damit, Trendthemen, aktuelle Ereignisse und Entwicklungen aus dem Web zu fischen.

Anstatt wie Google und seine Nachahmer einfach nur Suchrankings auszugeben, könnten neue Dienste stattdessen etwa Pro- und Contra-Argumente zu Streitfragen auflisten, wie das beispielsweise schon heute die Website Args.me vorführt.

Zudem könnte der OWI die Basis für fachliche und themenspezifische Recherchedienste bilden. Die Einschränkung des Suchindexes könnte sogar so weit gehen, dass spezialisierte Suchmaschinen fürs Handy entstehen. Damit kann der Anwender dann lokal ohne Webzugriff und folglich auch ohne Preisgabe seiner Nutzerdaten ausschließlich auf seinem Mobilgerät recherchieren.

Granitzer kann sich sogar vorstellen, dass zum OWI einmal ein Search Engine Hub entsteht. Der Anwender wählt seine Interessengebiete und einige Funktionen wie etwa Volltextsuche. Er beschränkt die Anfrage beziehungsweise die Ergebnisse auf besonders glaubwürdige oder auf besonders populäre Quellen. Und am Ende schickt er sein Formular ab und erhält eine nach seinen Wünschen konfigurierte Suchmaschine.

Kritische Infrastruktur

Zunächst läuft das EU-Projekt OpenWebSearch noch bis September 2025. Bis dahin wollen die Partner den fünf Petabyte großen Basisindex aufbauen. Im Anschluss wird es dann um eine nachhaltige Finanzierung dieser Infrastruktur gehen, voraussichtlich durch weitere EU-Mittel.

Im Sinne der digitalen Souveränität Europas kann man den Open Web Index sicherlich als kritische Infrastruktur ansehen. Die Projektpartner hoffen, dass damit transparente Strukturen im Web entstehen. Der angestrebte europäische Web-Index verspricht mehr Pluralität und nützt hoffentlich vor allem denjenigen, die auf ihren Websites einfach die besten und verlässlichsten Informationen liefern.

(agr@ct.de) c't

Literatur

[1] Robert Epstein, Ronald Robertson, The search engine manipulation effect (SEME) and itspossible impact on the outcomes of elections, Proceedings of the National Academy of Sciences (PNAS), 4. August 2015, pnas.org/doi/full/10.1073/pnas.1419828112

Links zu OpenWebSearch: ***ct.de/y6sw***