So funktionieren Suchmaschinen

inform@21

Worum geht es?

Damit man Informationen möglichst schnell durchsuchen kann, werden diese häufig sortiert

– alphabetisch

– nach Namen

– nach Grösse

– …

Worum geht es?

Computer speichern laufend eine riesige Menge an Informationen.

Sie müssen deshalb in der Lage sein, diese Informationen schnell zu durchsuchen.

Noch umfangreicher ist die Aufgabe für Suchmaschinen im Internet.

Suchmaschinen müssen bei einer Suchanfrage in kürzester Zeit die gesuchten Informationen aus Millionen von Web-seiten weltweit herausfiltern.

Wie machen Computer das?

Es gibt in der Informatik verschiedene Möglichkeiten, nach Informationen zu suchen.

Auch ein Computer sucht nicht immer nach dem gleichen Prinzip.

Die unterschiedlichen Suchmöglichkeiten nennt man Suchstrategien.

Bonbon-Spiel

Mit einem Spiel wollen wir nun zwei Suchstrategien kennen lernen:

- Lineare Suche
- Binäre Suche

Lineare Suche: Beispiel

Wir suchen in einem Lexikon, dem Duden oder einem Telefonbuch nach einem bestimmten Wort.

Bei der linearen Suche fangen wir vorne im Buch an und blättern Seite um Seite durch, bis wir das gesuchte Wort gefunden haben.

Lineare Suche: Fragen

1. Wie viele Begriffe müssen wir im schlechtesten Fall anschauen, wenn das Buch 10’000 Einträge hat?

2. Wie viele Begriffe müssen wir im besten Fall anschauen?

3. Wie viele Begriffe müssen wir durchschnittlich anschauen?

Lineare Suche: Lösungen

1. Wie viele Begriffe müssen wir im schlechtesten Fall anschauen, wenn das Buch 10’000 Einträge hat?
   Alle 10’000 Einträge, wenn der gesuchte Eintrag der letzte ist.

2. Wie viele Begriffe müssen wir im besten Fall anschauen?
   Genau einen, wenn der gesuchte Eintrag gleich der erste ist.

3. Wie viele Begriffe müssen wir durchschnittlich anschauen?
   Genau halb so viele, also 5’000 Einträge.

Lineare Suche

- Kennt ihr weitere Beispiele für eine lineare Suche?

- Was kommt euch in den Sinn?

Lineare Suche

Bei einer linearen Suche wird vom Anfang bis zum Ende der Reihe nach gesucht.

Computer können Informationen zwar extrem schnell durchforsten und verarbeiten. Bei einer Suchanfrage wäre es jedoch unmöglich, immer alle Informationen linear, vom Anfang bis zum Ende, zu durchsuchen.

Eine solche lineare Suche würde auch mit dem schnellsten Computer der Welt viel zu lange dauern.

Lineare Suche

Stell dir vor:
Ein Supermarkt hat 10’000 verschiedene Produkte im Sortiment.

Du willst ein Produkt kaufen und gehst damit an die Kasse.
Die Computerkasse liest den Strichcode ein. Er müsste somit alle 10’000 Produkte überprüfen, um den richtigen Preis zu finden.

Lineare Suche

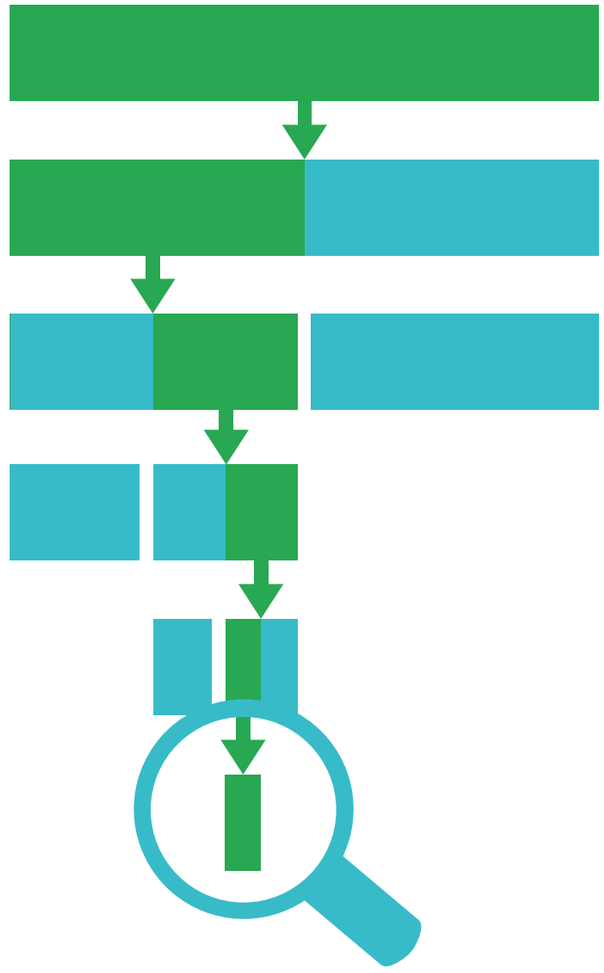
Obwohl der Computer mehrere Hundert Produkteinträge pro Sekunde abfragen kann, würde die Suche im ganzen Katalog ca. 10 Sekunden pro Artikel benötigen.

Wie lange würde das wohl für einen Familieneinkauf dauern, wenn ihr 50 Artikel einkaufen würdet?

Binäre Suche: Beispiel

Wir suchen wieder in einem Lexikon, dem Duden oder einem Telefonbuch nach einem bestimmten Wort, z. B. «Mathematik».

Bei der binären Suche schlagen wir das Buch in der Mitte auf und überlegen uns, ob das Wort «Mathematik» eher in der linken oder in der rechten Hälfte des Buches vorkommt.



Binäre Suche: Fragen

1. Wie viele Begriffe müssen wir im schlechtesten Fall anschauen, wenn das Buch 10’000 Einträge hat?

2. Wie viele Begriffe müssen wir im besten Fall anschauen?

3. Wie viele Begriffe müssen wir durchschnittlich anschauen?

Binäre Suche: Lösungen

1. Wie viele Begriffe müssen wir im schlechtesten Fall anschauen, wenn das Buch 10’000 Einträge hat?
   Angenommen wir suchen den ersten Begriff, dann schauen wir 5000, 2500, 1250, 625, 312, 156, 78, 39, 20, 10, 5, 3, 2 ,1 an. Also insgesamt 14 Einträge.

2. Wie viele Begriffe müssen wir im besten Fall anschauen?
   Genau einen, wenn es der Eintrag 5000 ist.

3. Wie viele Begriffe müssen wir durchschnittlich anschauen?
   Zwischen 1 und 14 Einträgen.

Binäre Suche: Hinweise

In der Praxis ist der Durchschnittswert wichtig, da wir im Telefonbuch ja nicht immer nur genau die erste Person suchen.

Mit der binären Suche müssen wir statt 5000 Einträgen also nur noch maximal 14 Einträge anschauen.

Damit das funktioniert, müssen die Daten aber sortiert sein.

Binäre Suche: Aufgaben

1. Wie viele Einträge müssten maximal angeschaut werden, wenn das Buch 20’000 Begriffe beinhaltet?

2. Wie viele Begriffe hat das Buch, wenn ihr maximal 20 Einträge anschauen müsst?

Binäre Suche: Lösungen

1. Wie viele Einträge müssten maximal angeschaut werden, wenn das Buch 20’000 Begriffe beinhaltet?
   Bei 20’000 Begriffen sind es maximal 15 Einträge (10’000, 5000, 2500, 1250, 625, 312, 156, 78, 39, 20, 10, 5, 3, 2 ,1).

2. Wie viele Begriffe hat das Buch, wenn ihr maximal 20 Einträge anschauen müsst?
   800’000 Begriffe (20 Einträge: 400’000, 200’000, 100’000, 50’000, 25’000, 12’500, 6250, 3125, 1563, 781, 390, 195, 97, 48, 25, 12, 6, 3, 2, 1)

Binäre Suche

Nehmen wir noch einmal das Beispiel aus dem Supermarkt.

<u>Stell dir vor:</u>
Ein Supermarkt hat 10’000 verschiedene Produkte im Sortiment.

Du willst ein Produkt kaufen und gehst damit an die Kasse.
Die Computerkasse liest den Strichcode ein.

Binäre Suche

Jeder Artikel hat eine Nummer erhalten. Alle Artikel werden der Reihe nach sortiert.

Beim Einlesen des Strichcodes wird nach der Nummer des Produktes gesucht.

Die Suchmaschine schaut sich zuerst die Artikelnummer genau in der Mitte der Liste an und entscheidet dann, in welcher Hälfte der Liste sie weitersuchen muss.

Binäre Suche

Dieses Vorgehen wird so lange fortgesetzt, bis der Suchartikel gefunden wurde.

Im ungünstigsten Fall würde es bei diesem Beispiel ebenfalls 14 Versuche benötigen.

Die Suche würde für den Computer nur knapp zwei Hundertstelsekunden dauern.

Binäre Suche

– Kennt ihr weitere Beispiele für eine binäre Suche?

– Was kommt euch in den Sinn?

Hashing-Suche

Das Internet besteht aus vielen Millionen von Webseiten.

Wie sollen all die vielen Webseiten auf der ganzen Welt von einer Suchmaschine durchsucht werden?

Hashing-Suche

Suchmaschinen setzen die Hashing-Suche ein.

Voraussetzung für eine Hashing-Suche ist eine geordnete und gekennzeichnete Datenmenge (z. B. eine sortierte und geordnete Liste der Suchergebnisse).

Dieses sortierte Anordnen von Informationen nennt man katalogisieren.

Einen sortierten Katalog (Liste, Tabelle etc.) nennt man Index.

Jede Suchmaschine benötigt als Grundlage für die Suche einen Index.

Der Index wird von sogenannten Webcrawlern (auch Spidern, Bots etc.) erstellt.

Hashing-Suche

Ein Webcrawler (auch Spider, Searchbot oder Robot) ist ein Computerprogramm, das automatisch das World Wide Web durchsucht und Webseiten analysiert.

Webcrawler sind eine spezielle Art von Bots, also Computerprogramme, die weitgehend automatisch sich wiederholenden Aufgaben nachgehen.

Hashing-Suche

Bei der Hashing-Suche wird der Suchbegriff so verändert, dass er eine Art «versteckte Zusatzinformation» enthält.

Diese gibt an, in welchem Abschnitt des Suchbereichs (z. B. Tabelle, Liste etc.) sich der Suchbegriff ungefähr befindet. Dadurch kann der gesuchte Begriff viel schneller gefunden werden.

Und so suchen Suchmaschinen

- Suchmaschinen durchsuchen das Internet nach Webseiten mit Bots.

- Die Webseiten werden analysiert und es werden Stichworte herausgeschrieben.

- Die Webseiten werden in ein Stichwortverzeichnis eingetragen (Index).

- Suchfragen werden mithilfe des aufgebauten Indexes beantwortet.

- Das heisst: Mittels Hashing-Suche wird gezielt, schnell und effizient gesucht.