



Dirk Lewandowski

# Understanding Search Engines

 Springer

---


# Understanding Search Engines

---

Dirk Lewandowski

# Understanding Search Engines

 Springer

Dirk Lewandowski   
Department of Information  
Hamburg University of Applied Sciences  
Hamburg, Germany

ISBN 978-3-031-22788-2      ISBN 978-3-031-22789-9 (eBook)  
<https://doi.org/10.1007/978-3-031-22789-9>

Translation from the German language edition: “Suchmaschinen verstehen” by Dirk Lewandowski,  
© Springer 2021. Published by Springer Vieweg, Berlin, Heidelberg. All Rights Reserved.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

---

## Preface

This book aims to give a comprehensive introduction to search engines. The particularity of this book is that it looks at the subject from different angles. These are, in particular, technology, use, Internet-based research, economics, and societal significance. In this way, I want to reflect the complexity of the search engines and Web search as a whole. I am convinced that only such a comprehensive view does justice to the topic and enables a real understanding.

A German-language version of this book has been available for several years. This English edition follows the third German edition of 2021. I am pleased that the publisher has made this international edition possible.

In this translation, care has been taken to adapt to the international context where necessary. However, for many examples, it does not matter in which country a search was carried out or a screenshot was taken. However, the references cited in the text were adapted where English-language sources were available. The further reading sections at the end of the chapters have also been adapted.

I would like to thank all those who have asked and encouraged me over the years to produce an English edition. So, here it is, and I hope it will be as valuable to many readers as the German editions have been.

Hamburg, Germany  
September 2022

Dirk Lewandowski

---

# Contents

<b>1</b>	<b>Introduction</b> . . . . .	1
1.1	The Importance of Search Engines . . . . .	2
1.2	A Book About Google? . . . . .	5
1.3	Objective of This Book . . . . .	6
1.4	Talking About Search Engines . . . . .	7
1.5	Structure of This Book . . . . .	7
1.6	Structure of the Chapters and Markings in the Text . . . . .	9
1.7	Summary . . . . .	9
	References . . . . .	10
<b>2</b>	<b>Ways of Searching the Web</b> . . . . .	11
2.1	Searching for a Website vs. Searching for Information on a Topic . . . . .	11
2.2	What Is a Document? . . . . .	13
2.3	Where Do People Search? . . . . .	13
2.4	Different Pathways to Information on the World Wide Web . . . . .	14
2.4.1	Search Engines . . . . .	14
2.4.2	Vertical Search Engines . . . . .	17
2.4.3	Metasearch Engines . . . . .	19
2.4.4	Web Directories . . . . .	20
2.4.5	Social Bookmarking Sites . . . . .	21
2.4.6	Question-Answering Sites . . . . .	21
2.4.7	Social Networking Sites . . . . .	22
2.5	Summary . . . . .	22
	References . . . . .	23
<b>3</b>	<b>How Search Engines Capture and Process Content from the Web</b> . . . . .	25
3.1	The World Wide Web and How Search Engines Acquire Its Contents . . . . .	28
3.2	Content Acquisition . . . . .	31
3.3	Web Crawling: Finding Documents on the Web . . . . .	33
3.3.1	Guiding and Excluding Search Engines . . . . .	37
3.3.2	Content Exclusion by Search Engine Providers . . . . .	39

---

3.3.3	Building the Database and Crawling for Vertical Collections . . . . .	40
3.4	The Indexer: Preprocessing Documents for Searching . . . . .	42
3.4.1	Indexing Images, Audio, and Video Files . . . . .	47
3.4.2	The Representation of Web Documents in Search Engines . . . . .	48
3.5	The Searcher: Understanding Queries . . . . .	51
3.6	Summary . . . . .	54
	References . . . . .	56
<b>4</b>	<b>User Interaction with Search Engines . . . . .</b>	<b>59</b>
4.1	The Search Process . . . . .	59
4.2	Collecting Usage Data . . . . .	60
4.3	Query Types . . . . .	61
4.4	Sessions . . . . .	64
4.5	Queries . . . . .	65
4.5.1	Entering Queries . . . . .	66
4.5.2	Autocomplete Suggestions . . . . .	68
4.5.3	Query Formulation . . . . .	69
4.5.4	Query Length . . . . .	71
4.5.5	Distribution of Queries by Frequency . . . . .	73
4.5.6	Query Trends . . . . .	74
4.5.7	Using Operators and Commands for Specific Searches . . . . .	76
4.6	Search Topics . . . . .	77
4.7	Summary . . . . .	78
	References . . . . .	79
<b>5</b>	<b>Ranking Search Results . . . . .</b>	<b>83</b>
5.1	Groups of Ranking Factors . . . . .	85
5.2	Text Statistics . . . . .	86
5.2.1	Identifying Potentially Relevant Documents . . . . .	86
5.2.2	Calculating Frequencies . . . . .	88
5.2.3	Considering the Structural Elements of Documents . . . . .	89
5.3	Popularity . . . . .	91
5.3.1	Link-Based Rankings . . . . .	92
5.3.2	Usage Statistics . . . . .	97
5.4	Freshness . . . . .	105
5.5	Locality . . . . .	106
5.6	Personalization . . . . .	111
5.7	Technical Ranking Factors . . . . .	113
5.8	Ranking and Spam . . . . .	114
5.9	Summary . . . . .	116
	References . . . . .	117

<b>6</b>	<b>Vertical Search</b> . . . . .	119
6.1	Vertical Search Engines as the Basis of Universal Search . . . . .	121
6.2	Types of Vertical Search Engines . . . . .	123
6.3	Collections . . . . .	125
6.3.1	News . . . . .	125
6.3.2	Scholarly Content . . . . .	129
6.3.3	Images . . . . .	132
6.3.4	Videos . . . . .	133
6.4	Integrating Vertical Search Engines into Universal Search . . . . .	133
6.5	Summary . . . . .	134
	References . . . . .	135
<b>7</b>	<b>Search Result Presentation</b> . . . . .	137
7.1	The Influence of Device Types and Screen Resolutions . . . . .	138
7.2	The Structure of Search Engine Result Pages . . . . .	139
7.3	Elements on Search Engine Result Pages . . . . .	146
7.3.1	Organic Results . . . . .	146
7.3.2	Advertising . . . . .	147
7.3.3	Universal Search Results . . . . .	148
7.3.4	Knowledge Graph Results . . . . .	150
7.3.5	Direct Answers . . . . .	150
7.3.6	Integration of Transactions . . . . .	152
7.3.7	Navigation Elements . . . . .	153
7.3.8	Support for Query Modification . . . . .	154
7.3.9	Search Options on the Result Page . . . . .	155
7.4	The Structure of Snippets . . . . .	156
7.5	Options Related to Single Results . . . . .	159
7.6	Selection of Suitable Results . . . . .	160
7.7	Summary . . . . .	161
	References . . . . .	162
<b>8</b>	<b>The Search Engine Market</b> . . . . .	165
8.1	Search Engines' Business Model . . . . .	165
8.2	The Importance of Search Engines for Online Advertising . . . . .	166
8.3	Search Engine Market Shares . . . . .	167
8.4	Important Search Engines . . . . .	169
8.5	Partnerships in the Search Engine Market . . . . .	170
8.6	Summary . . . . .	171
	References . . . . .	173
<b>9</b>	<b>Search Engine Optimization (SEO)</b> . . . . .	175
9.1	The Importance of Search Engine Optimization . . . . .	176
9.2	Fundamentals of Search Engine Optimization . . . . .	178
9.2.1	Content . . . . .	180
9.2.2	Architecture . . . . .	181
9.2.3	HTML . . . . .	181



9.2.4	Trust . . . . .	182
9.2.5	Links . . . . .	182
9.2.6	User-Related Factors . . . . .	183
9.2.7	Toxins . . . . .	183
9.2.8	Vertical Search Engines . . . . .	184
9.3	Search Engine Optimization and Spam . . . . .	185
9.4	The Role of Ranking Updates . . . . .	185
9.5	Search Engine Optimization for Special Collections . . . . .	186
9.6	The Position of Search Engine Providers . . . . .	187
9.7	Summary . . . . .	188
	References . . . . .	189
<b>10</b>	<b>Search Engine Advertising (SEA)</b> . . . . .	<b>191</b>
10.1	Specifics of Search Engine Advertising . . . . .	194
10.2	Functionality and Ranking . . . . .	195
10.3	Distinguishing Between Ads and Organic Results . . . . .	197
10.4	Advertising in Universal Search Results . . . . .	198
10.5	Summary . . . . .	199
	References . . . . .	200
<b>11</b>	<b>Alternatives to Google</b> . . . . .	<b>203</b>
11.1	Overlap Between Results from Different Search Engines . . . . .	204
11.2	Why Should One Use a Search Engine Other Than Google? . . . . .	204
11.2.1	Obtaining a “Second Opinion” . . . . .	205
11.2.2	More or Additional Results . . . . .	205
11.2.3	Different Results . . . . .	206
11.2.4	Better Results . . . . .	207
11.2.5	Different Result Presentation . . . . .	207
11.2.6	Different User Guidance . . . . .	207
11.2.7	Avoiding the Creation of User Profiles . . . . .	208
11.2.8	Alternative Search Options . . . . .	208
11.3	When Should One Use a Search Engine Other Than Google? . . . . .	208
11.4	Particularities of Google due to Its Market Dominance . . . . .	210
11.5	Summary . . . . .	212
	References . . . . .	213
<b>12</b>	<b>Search Skills</b> . . . . .	<b>215</b>
12.1	Source Selection . . . . .	217
12.2	Selecting the Right Keywords . . . . .	218
12.3	Boolean Operators . . . . .	218
12.4	Connecting Queries with Boolean Operators . . . . .	222
12.5	Advanced Search Forms . . . . .	223
12.6	Commands . . . . .	225
12.7	Complex Searches . . . . .	228
12.8	Summary . . . . .	228
	References . . . . .	229

<b>13</b>	<b>Search Result Quality</b> . . . . .	231
13.1	Criteria for Evaluating Texts on the Web . . . . .	231
13.2	Human vs. Machine Inspection of Quality . . . . .	232
13.3	Scientific Evaluation of Search Result Quality . . . . .	235
13.3.1	Standard Test Design of Retrieval Effectiveness Studies . . . . .	237
13.3.2	Measuring Retrieval Effectiveness Using Click-Through Data . . . . .	240
13.3.3	Evaluation in Interactive Information Retrieval . . . . .	241
13.4	Summary . . . . .	243
	References . . . . .	243
<b>14</b>	<b>The Deep Web</b> . . . . .	247
14.1	The Content of the Deep Web . . . . .	249
14.2	Sources vs. Content from Sources, Accessibility of Content via the Web . . . . .	251
14.3	The Size of the Deep Web . . . . .	254
14.4	Areas of the Deep Web . . . . .	255
14.5	Social Media as Deep Web Content . . . . .	256
14.6	What Role Does the Deep Web Play Today? . . . . .	258
14.7	Summary . . . . .	258
	References . . . . .	259
<b>15</b>	<b>Search Engines Between Bias and Neutrality</b> . . . . .	261
15.1	The Interests of Search Engine Providers . . . . .	262
15.2	Search Engine Bias . . . . .	263
15.3	The Effect of Search Engine Bias on Search Results . . . . .	265
15.4	Interest-Driven Presentation of Search Results . . . . .	267
15.5	What Would “Search Neutrality” Mean? . . . . .	270
15.6	Summary . . . . .	271
	References . . . . .	272
<b>16</b>	<b>The Future of Search</b> . . . . .	275
16.1	Search as a Basic Technology . . . . .	276
16.2	Changes in Queries and Documents . . . . .	277
16.3	Better Understanding of Documents and Queries . . . . .	278
16.4	The Economic Future of Search Engines . . . . .	278
16.5	The Societal Future of Search Engines . . . . .	279
16.6	Summary . . . . .	281
	References . . . . .	282
	<b>Glossary</b> . . . . .	283
	<b>Index</b> . . . . .	293



This book is about better understanding the search tools we use daily. Only when we have a basic understanding of how search engines are constructed and how they work can we use them effectively in our research.

However, not only the use of existing search engines is relevant here but also what we can learn from well-known search engines like Google when we want to build our own search systems. The starting point is that Web search engines are currently the leading systems in terms of technology, setting the standards in terms of both the search process and user behavior. Therefore, if we want to build our own search systems, we must comply with the habits shaped by Web search engines, whether we like it or not.

This book is an attempt to deal with the subject of search engines comprehensively in the sense of looking at it from different angles:

1. Technology: First of all, search engines are technical systems. This involves the gathering of the Web's content as well as ranking and presenting the search results.
2. Use: Search engines are not only shaped by their developers but also by their users. Since the data generated during use is incorporated into the ranking of the search results and the design of the user interface, usage significantly influences how search engines are designed.
3. Web-based research: Although, in most cases, search engines are used in a relatively simple way – and often not much more is needed for a successful search – search engines are also tools for professional information research. The fact that search engines are easy to use for everyone does not mean that every search task can be easily solved using them.
4. Economy: Search engines are of great importance for content producers who want to get their content on the market. Because they are central nodes in the Web, they also play an important economic role. Here, the main focus is on search engine visibility, which can be achieved through various online marketing measures (such as search engine optimization and placing advertisements).

5. Society: Since search engines are the preferred means of searching for information and are used massively every day, they also have an enormous significance for knowledge acquisition in society. Among other things, this raises the question of whether search results are credible and whether search engines play a role in spreading misinformation and disinformation, often treated under the label of “fake news.”

My fundamental thesis is that one is impossible without the other: we cannot understand search engines as technical systems if we do not know their social significance. Nor can we understand their social impact if we do not know the underlying technology. Of course, one does not have to have the same detailed knowledge in all areas; but one should achieve a solid basis.

Of course, an introductory book cannot cover the topics mentioned comprehensively. My aim is instead to introduce the concepts central to discussing search engines and provide the basic knowledge that makes a well-founded discussion of search engines possible in the first place.

---

## 1.1 The Importance of Search Engines

In this book, I argue that search engines have an enormous social significance. This can be explained, on the one hand, by their mass use and, on the other hand, by the ranking and presentation of search results.

Search engines (like other services on the Internet) are used en masse. Their importance lies in the fact that we use them to search for information actively. Every time we enter a query, we reveal our interests. With every search engine result page (SERP) that a search engine returns to us, there is a (technically mediated) interpretation of both the query and the potentially relevant results. By performing these interpretations in a particular way, a search engine conveys a specific impression of the world of information found on the Web.

For every query, there is a result page that displays the results in a specific order. Although, in theory, we can select from all these results, we rely heavily on the order given by the search engine. De facto, we do not select from the possibly millions of results found but only from the few displayed first.

If we consider this, societal questions arise, such as how diverse the search engine market is: Is it okay to use only one search engine and have only one of many possible views of the information universe for each query?

The importance of search engines has already been put into punchy titles such as “Search Engine Society” (Halavais, 2018), “Society of the Query” (the title of a conference series and a book; König & Rasch, 2014), and “The Googlization of Everything” (Vaidhyanathan, 2011). Perhaps it is not necessary to go so far as to proclaim Google, search engines, or queries as the determining factor of our society; however, the enormous importance of search engines for our knowledge acquisition can no longer be denied.

If we look at the hard numbers, we see that search engines are the most popular service on the Internet. We regard the Internet as a collection of protocols and services, including e-mail, chat, and the File Transfer Protocol (FTP). It may seem surprising that the use of search engines is at the top of the list when users are asked about their activities on the Internet. Search engines are even more popular than writing and reading e-mails. For instance, 76% of all Germans use a search engine at least once a week, but “only” 65% read or write at least one e-mail during this time. This data comes from the ARD/ZDF-Onlinestudie (Beisch & Schäfer, 2020), which surveys the use of the Internet among the German population every year. Comparable studies confirm the high frequency of search engine use: the Eurobarometer study (European Commission, 2016) shows that 85% of all Internet users in Germany use a search engine at least once a week; the figure for daily use is still 48%. Germany is below the averages of the EU countries (88% and 57%, respectively).

Let’s look at the ARD/ZDF-Onlinestudie to see which other Internet services are used particularly often. We find that, in addition to e-mail and search engines, messengers (probably WhatsApp in particular) are the most popular. On the other hand, social media services only reach 36%.

A second way of looking at this is to look at the most popular websites (Alexa.com, 2021). Google is in the first and third place ([google.com](http://google.com) and [google.de](http://google.de)), followed by YouTube (second place), Amazon, and eBay. It is striking that not only Google is in the first place but eBay and Amazon are two major e-commerce companies that not only offer numerous opportunities for browsing but also play a major role in (product) searches.

The fact that search engines are a mass phenomenon can also be seen in the number of daily queries. Market research companies estimate the number of queries sent to Google alone at around 3.3 trillion in 2016 (Internet Live Stats & Statistic Brain Research Institute, 2017) – that’s more than a million queries per second!

An additional level of consideration arises when we look at how users access information on the World Wide Web. While there are theoretically many access points to information on the Web, search engines are the most prevalent. On the one hand, Web pages can, of course, be accessed directly by typing the address (Uniform Resource Locator; URL) into the browser bar. Then there are other services, such as social media services, which also lead us to websites. But none of these services has achieved a level of importance comparable to that of search engines for accessing information on the Web, nor is this situation likely to change in the foreseeable future.

Last but not least, search engines are also significant because of the online advertising market. The sale of ads in search engines (ads in response to a query) accounts for 40% of the market (Zenith, 2021); in Germany alone, search engine advertising generated sales of 4.1 billion euros in 2019 (Statista, 2021).

This form of advertising is particularly attractive because, with each search query, users reveal what they want to find and thus also whether and what they might want to buy. This makes it easy for advertisers to decide when they want to offer their product to a user. Scatter losses, i.e., the proportion of users who see an

advertisement but have no interest in it at that moment, can be significantly reduced or even avoided altogether in this way.

Search engine providers, like other companies, have to earn money. So far, the only model search engines have used to make money is the insertion of advertising in the form of text ads around the search results. Other revenue models have not caught on. This also means that search engine providers do not, as is often claimed, align their search engines solely with the demands and needs of users but also with their own profit intentions and those of their advertising customers.

For companies, however, the importance of search engines is not only a result of being able to use search engines as an advertising platform but also because of being found by users in the organic search results. The procedures that serve to increase the probability of being found are subsumed under the title of search engine optimization.

Already at this point, we see that if we consider search engines not only as technical systems but also as socially relevant, we are dealing with at least four stakeholder groups or actor groups (see Röhle, 2010, p. 14):

1. Search engine providers: On the one hand, search engine providers are interested in satisfying their users. This involves both the quality of the search results and the user experience. On the other hand, search engine providers' second major (or even more significant?) interest is to offer their advertisers an attractive environment and earn as much money as possible from advertising.
2. Users: The users' interest is to obtain satisfactory search results with little effort and not to be disturbed too much in their search process, for example, by intrusive advertising.
3. Content producers: Anyone who offers content on the Web also wants to be found by (potential) users. However, another interest of many content producers is to earn money with their content. This, in turn, means that it is not necessarily in their interest to make their content fully available to search engines.
4. Search engine optimizers: Search engine optimizers work on behalf of content producers to ensure that their offerings can be found on the Web, primarily in search engines. Their knowledge of the search engines' ranking procedures and their exploitation of these procedures to place "their" websites influence the search engine providers, who attempt to protect themselves against manipulation.

This brief explanation of the stakeholders already shows that this interplay can lead to conflicts. Search engine providers have to balance the interests of their users and their advertisers; search engine optimizers have to ensure the maximum visibility of their clients' offerings but must not exploit their knowledge of how search engines work to such an extent that they are penalized by search engine providers for manipulation.

Clearly, we are dealing with complex interactions in the search engine market. Only if we look at search engines from different perspectives are we able to classify these interactions and understand why search engines are designed the way they are.

Search engines have to meet the needs of different user groups; it is not enough for them to restrict their services to one of these groups.

When we talk about search engines and their importance for information access, we usually only consider the content initially produced for the Web. However, search engines have been trying to include content from the “real,” i.e., the physical world, in their search systems for years. Vaidhyathan (2011) distinguishes three types of content that search engines like Google capture:

1. Scan and link: External content is captured, aggregated, and made available for search (e.g., Web search).
2. Host and serve: Users’ content is collected and hosted on their own platform (e.g., YouTube).
3. Scan and serve: Things from the real world are transferred into the digital world by the search engine provider (e.g., Google Books, Google Street View).

Vaidhyathan (2011) summarizes this under “The Googlization of Everything” (which is also the title of his book) and thus illustrates not only that search engine content goes beyond the content of the Web (even if this continues to form the basis) but also that we are still at the beginning when it comes to the development of search engines: So far, only a small part of all the information that is of interest to search engines has been digitized and thus made available for search.

Furthermore, there is a second, largely taken-for-granted assumption, namely, that a search process must necessarily contain a query entered by the user. However, we see that search engines can increasingly generate queries by themselves by observing the behavior of a user and then offering information that is very likely to be useful to them. For example, suppose a user is walking through a city with their smartphone in their pocket. In that case, it is easy to predict their desire for a meal option at lunchtime and suggest a restaurant based on that user’s known past preferences and current location. To do this, a query (made up of the above information) is required, but the user does not have to enter it themselves. We will return to this in Chap. 4.

---

## 1.2 A Book About Google?

When we think of search engines, we primarily think of Google. We all use this search engine almost daily, usually for all kinds of search purposes. Here, again, the figures speak for themselves: in Germany, well over 90% of all queries to general search engines are directed to Google, while other search engines play only a minor role (Statcounter, 2021).

Therefore, this book is based on everyday experience with Google and tries to explain the structure and use of search engines using this well-known example. Nevertheless, this book aims to go further: to show which alternatives to Google there are and when it is worthwhile to use them. But this book will not describe all possible search engines; it is rather about introducing other search engines, utilizing

examples, and thus, first of all, getting the reader to think about whether Google is the best search engine for precisely their research before carrying out more complex searches.

To a certain extent, it can also be said that if you know one search engine, you will be better able to deal with all the others. We will learn about the basic structure of search engines and their most important functions by looking at Google, which we all already know, at least from the user side. The acquired knowledge can then easily be transferred to other search engines.

Most of the search examples and screenshots also come from Google. In most cases, however, the examples can be transferred to other search engines. Where this is not the case, this is indicated.

Regarding the similarity between the different search engines, we can generally say that Google's competitors are in a dilemma: Even if they offer innovative functions and try to do things differently from Google, they are fundamentally oriented toward Google's idea of how a search engine should look and work. This orientation toward Google cannot be blamed on the other search engine providers because, on the one hand, they can only win over users if those who are used to Google find their way around immediately; on the other hand, they have to distinguish themselves from Google to be able to offer any added value compared to this search engine.

---

### **1.3 Objective of This Book**

By its very nature, this book is restricted in its function as an introductory book and is intended as a general overview. This also means that many topics cannot be dealt with in detail, but we must remain "on the surface" instead. However, this does not mean that the contents must be superficial. On the contrary, I have tried to present the contents as simply as possible but without sacrificing the necessary accuracy. Some topics are explored via a specific example (such as a vertical search engine), which is explained in more detail, so that this information can then be applied to other topics.

This book is about transfer: what you learn from one or a few search engines should be transferable to others. Therefore, it does not matter that some of the contents in this book – especially when it comes to details of a particular search engine – may have already changed by the time this book is published. This is unavoidable, especially in rapidly evolving fields, but the goal is to convey basic knowledge about search engines that can then be applied to all search engines.

This book is not a substitute for introductory works on, for example, information retrieval or searching the Web, even though topics from these areas are covered. The relevant introductory literature on the respective topics is mentioned in the respective chapters. This book aims to provide an overview and a consideration of different perspectives on search engines, not an all-encompassing presentation of the individual topics.

Students in particular often fear that they will only be able to understand search engines if they delve into algorithms and technical details. In this book, the essential



---

procedures are described in a concise and understandable manner, but the main aim is to understand the ideas underlying the technical processes. This will enable us to assess why search engines work as well or as poorly as they do at present and what prospects there are for their further development.

It is only natural that in any attempt to look at a topic from different perspectives, one gravitates toward one's own subject and focuses on the interests of one's own discipline. Thus, my interest and the focus of my consideration naturally follow the subject area and the methods of information science, which always (also) considers technical information systems from the perspective of humans. In addition, however, I have made an effort to also consider the perspective of other subjects such as computer science and media and communication studies (including their literature).

---

## 1.4 Talking About Search Engines

To talk about an object, you need a consistent vocabulary. You must know that when you use certain terms, you are talking about the same thing. In order to avoid talking past each other, it is therefore necessary to agree on terminology. Since there is currently no single, unified terminology in the field of search engines, and search engine optimizers, information scientists, and communication scientists, for example, each speak a language of their own, this book is also intended to contribute to mutual understanding. At the end of the book, there is a glossary that lists and explains all important terms in alphabetical order. I have made an effort to include synonyms and related terms so that readers who have already gained some knowledge from the literature can find “their” terms and quickly get used to the terminology I have used.

---

## 1.5 Structure of This Book

Of course, you can read this book from cover to cover, which was my primary intention when writing it. However, if you only want to read about a specific topic, the chapter structure allows you to do so.

Following this introductory chapter, Chap. 2 covers different ways of searching the Web. Indeed, search engines like Google are not the only form of access to information on the Web, even if the form of the algorithmic universal search engine has become widely accepted. The various forms of search systems are briefly introduced, and their significance is discussed in the context of searching the Web.

Chapter 3 then explains the basic technical structure of algorithmic search engines. It explains how search engines obtain content from the Web, how this content is prepared so that it can be searched efficiently, and how users' queries can be interpreted and processed automatically.

After these two technical chapters, we consider the user side in Chap. 4: what is actually searched for in search engines, how are queries formulated, and how do users select the most suitable results?

Closely related are the ranking procedures, i.e., the arrangement of search results. Chapter 5 describes the basic procedures and explains their significance. Although it is often claimed that the ranking of search results is the big secret of every search engine, knowledge of the most important ranking factors can at least fundamentally explain the arrangement of search results, even if the concrete ranking depends on a multitude of weightings that cannot be traced in detail. This understanding, in turn, can help us both in our searches and in preparing our own content for search engines or even in creating our own information systems.

Chapter 6 then shows how search results from the general Web Index are extended by adding so-called vertical collections such as news, images, or videos. For this purpose, the well-known search engines have built and integrated numerous vertical search engines whose results are displayed on the search engine result pages.

Chapter 7 is devoted to the presentation of search results. For some years now, the well-known search engines have deviated from the usual list form of search result presentation and have instead established new forms of compiling search results with concepts such as universal search and knowledge graph. This has made the result pages more attractive and increased the choices available on these pages. With this type of result presentation, the search engines also deliberately guide the users' attention.

This brings us to the economic realities related to search engines. In Chap. 8, we deal with the search engine market and thus, among other things, with the question of how Google has succeeded in almost completely dominating the search engine market (at least in Europe). Of course, the question of whether such a situation is desirable and how it could be changed is also raised here.

Chapter 9 is devoted to the side of those who want to make their content best available via search engines and their helpers, the search engine optimizers. They use their knowledge of search engines' indexing and ranking procedures to make content easier to find and to bring traffic to their customers. Their techniques range from simple text modification to complex procedures that consider the Web's linking structure.

Chapter 10 provides a detailed description of the advertising displayed in search engines. On the one hand, it deals with the ads shown on search result pages as a type of search result, and on the other hand, with the question to what extent users can distinguish these ads from the organic search results.

Chapter 11 deals with alternatives to Google. First, it is important to answer the question of what makes a search engine an alternative search engine. Is it enough that it is simply a search engine other than Google? Then, based on fundamental considerations and concrete situations in the search process, we will explain in which cases it is worth switching to another search engine.

In Chap. 12, we change the perspective again and consider search engines as tools for advanced Internet research. In the chapter on user behavior, it became clear that most users put little effort into formulating their queries and sifting through the results. Therefore, we want to see what strategies and commands we can use to get the most out of search engines.

Another topic related to searching, but also to the general evaluation of search engines, is the question of the quality of the search results, which we will address in Chap. 13. The quality of search results can be viewed from two perspectives: One is about the user's result evaluation in the course of their search; the other is about scientific comparisons of the result quality of different search engines.

Chapter 14 deals with the contents of the Web that are not accessible to general search engines, the so-called Deep Web. An enormous treasure trove of information cannot be found with Google and similar search engines or at least to a limited extent. We will see why this content remains hidden from the search engines and with what methods we can nevertheless access it.

While the previous topics dealt with aspects from the areas of technology, use, and Web-based research, Chap. 15 deals with the societal role of search engines. What role do search engines play in knowledge acquisition, and what role should they play?

Finally, Chap. 16 focuses on the future of search. Of course, a book like this can only ever offer a snapshot, and 10 years ago, it would have provided a different picture than today. However, the "problem" of search has by no means been solved (and may never be solved), so it is worth looking at today's search engines not only in their evolution toward the current state but also to venture a look into the (near) future.

---

## 1.6 Structure of the Chapters and Markings in the Text

By their very nature, chapters on different topics must be structured differently. Nevertheless, the chapters in this book have certain similarities: At the beginning of each chapter, there is an introduction that defines the topic and briefly describes its significance for the book. Detailed explanations follow this. At the end of each chapter, there is a summary that reviews the most important points. Each chapter also contains a bibliography and, in a separate box, a list of recommendations for further reading for those who wish to delve deeper into the topic.

There are also examples in boxes throughout the text that illustrate and deepen what is said in the main text but are not essential for understanding the main text.

---

## 1.7 Summary

Search engines are essential tools for accessing the information on the World Wide Web. In this book, we will look at them from the perspectives of technology, usage, economic aspects, searching the Web, and society.

The significance of search engines results from their mass use and the fact that they are by far the preferred means of accessing information on the World Wide Web. However, one search engine, in particular, Google, is used for most searches.

We should not only consider search engines as technical systems. Due to the interactions of different stakeholders (search engine providers, users, content

producers, and search engine optimizers), there are numerous factors which influence search results and which the search engine providers do not exclusively direct.

In terms of content, search engines no longer only capture the Web content but also offer platforms on which users can create content themselves, which is then made searchable. Furthermore, search engine providers offer various vertical search engines in addition to Web search, whose results are included in the general search engine result pages. Finally, content from the physical world is transferred to the digital world and integrated into search.

---

## References

- Alexa.com. (2021). *Top sites in Germany*. <https://www.alexa.com/topsites/countries/DE>.
- Beisch, N., & Schäfer, C. (2020). Ergebnisse der ARD/ZDF-Onlinestudie 2020: Internetnutzung mit großer Dynamik: Medien, Kommunikatio, Social Median. *Media Perspektiven*, 51(9), 462–481.
- European Commission. (2016). *Special Eurobarometer 447 – Online Platforms*. European Commission. <https://doi.org/10.2759/937517>
- Halavais, A. (2018). Search engine society. Polity.
- Internet Live Stats, & Statistic Brain Research Institute. (2017). Anzahl der Suchanfragen bei Google weltweit in den Jahren 2000 bis 2016 (in Milliarden). In *Statista – Das Statistik-Portal*. <https://de.statista.com/statistik/daten/studie/71769/umfrage/anzahl-der-google-suchanfragen-pro-jahr/>.
- König, R., & Rasch, M. (Eds.). (2014). *Society of the query reader: Reflections on web search*. Institute of Network Cultures.
- Röhle, T. (2010). *Der Google-Komplex: Über Macht im Zeitalter des Internets*. Transcript. <https://doi.org/10.14361/transcript.9783839414781>
- Statcounter. (2021). *Search engine market share*. <https://gs.statcounter.com/search-engine-market-share>.
- Statista. (2021). Prognose der Umsätze mit Suchmaschinenwerbung in Deutschland in den Jahren 2017 bis 2025 (in Millionen Euro). In *Statista – Das Statistik-Portal*. <https://de.statista.com/prognosen/456188/umsaetze-mit-suchmaschinenwerbung-in-deutschland>.
- Vaidhyanathan, S. (2011). *The Googlization of everything (and why we should worry)*. University of California Press. <https://doi.org/10.1525/9780520948693>
- Zenith. (2021). Prognose zu den Investitionen in Internetwerbung weltweit in den Jahren 2018 bis 2022 nach Segmenten (in Milliarden US-Dollar). In *Statista – Das Statistik-Portal*. <https://de.statista.com/statistik/daten/studie/209291/umfrage/investitionen-in-internetwerbung-weltweit-nach-segmenten/>.

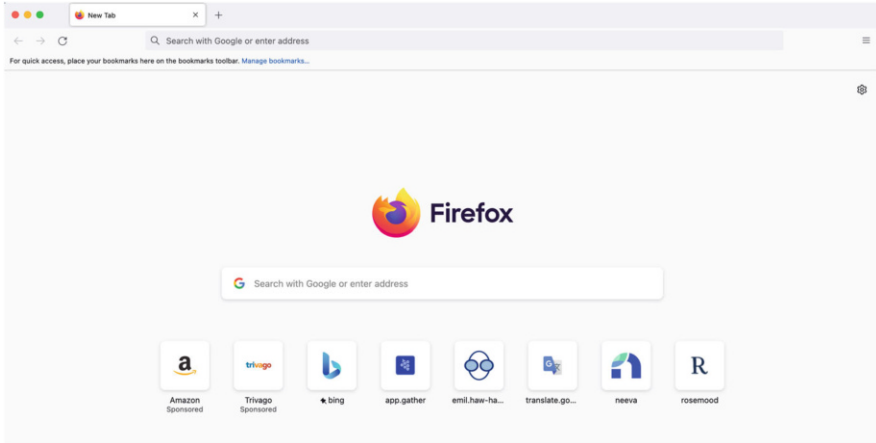


At first glance, searching the Web may seem trivial: we enter a query and receive a search engine result page (SERP) on which we select a result. But this is only one of the many ways to access information on the Web. In this chapter, we introduce the different ways of accessing the information on the Web and explain why access via search engines has become dominant.

## 2.1 Searching for a Website vs. Searching for Information on a Topic

First, we need to ask what we want or can achieve with a search. For now, it is sufficient to distinguish three cases. We will explain these cases with the help of the Firefox browser starting page shown in Fig. 2.1:

1. A user wants to go to a specific website they already know. To do this, they enter the URL into their browser's address bar (Fig. 2.1, top line). Then, on the website, they either read something directly, conduct a search, or click on further documents. This process has little to do with our intuitive understanding of searching, but it is a means of getting to the information we are looking for. For example, a user interested in news on a current topic can go directly to a news website and either read relevant articles directly on the front page, click on them there, or search for articles using the internal search function of the news website.
2. A user wants to go to a specific website they either already know or do not yet know about and searches for it via the address bar (combined URL and search bar) (Fig. 2.1, top line) or the search field (Fig. 2.1, field placed in the middle). The search is carried out in the previously set search engine (for settings, see Chap. 8). Such a search may be for a known website – in which case searching is merely a “shortcut” to direct entry in the address bar (e.g., entering “ny times” in the search field instead of “[www.nytimes.com](http://www.nytimes.com)” in the address bar) or it may be to help if the user can no longer remember the exact address of the website they are



**Fig. 2.1** Start page of the Firefox browser with the address bar and Google search as the default start page (August 26, 2022)

looking for (e.g., if they no longer know whether a website ends with .com or .org). When searching directly for a website that is not yet known, the user at least assumes that such a website exists and searches accordingly.

3. The third type is a user looking for information not yet known to them. This type differs fundamentally from the previous two as the user is not looking for a specific website but for information on a topic. Here, it is impossible to predict with certainty whether this information can be found on a particular website or whether the information from a single website is sufficient to satisfy the information need.

We will discuss the subdivision of search queries according to intentions or information needs in more detail in Sect. 4.3. For the time being, it is sufficient to distinguish between a search for known websites and a search for unknown information. To be able to assess different ways of accessing information on the Web, it is essential that we can already distinguish between these cases.

In our example, we have already seen that search queries can be entered in different places. We will return to the significance of the search engine preset in a browser's search box or address bar and the preset start page in the chapter on the search engine market (Chap. 8).

### **Is Searching the Web Like Looking for a Needle in a Haystack?**

Searching the Web is often compared to looking for a needle in a haystack. This picture is meant to illustrate that it is difficult to find the right thing (the needle) because of the vast amount of information available (the haystack).

(continued)

But this image is skewed. In the case of the haystack, we know what the needle looks like, and there is only one needle, so we can tell when our search is finished.

However, in the case of searching for previously unknown information, we do not always have such a clearly defined idea of what we want to find. There could be several needles that might also serve our purpose differently. And it could also be that we are only satisfied when we have found several needles that complement or confirm each other.

---

## 2.2 What Is a Document?

By its very design, the Web is multimedial and contains much more than just text. In this respect, search engines are not only there to find text on the Web but also other types of information – even if search today is (still) primarily text-based. But regardless of whether it is a text, an image, or a video, we will speak of a document or, alternatively, an information object.

So what is a document? When we think of documents, we might first think of official documents, such as those issued by public authorities with a stamp and signature. In information science, however, the term is defined much more broadly: a document is a record of information, regardless of whether it is in written form (text document) or, for example, in pictorial form (image document). Concerning search engines, this means that every piece of content they display (text, images, videos, etc.) is a document. Therefore, we sometimes speak of information objects instead to make it clear that we are not only talking about textual documents.

---

## 2.3 Where Do People Search?

The times when search engines were used mainly in the same context, namely, on desktop computers or laptops, are long gone. People now search on a wide range of devices, ranging from smartphones and tablets to wearables and purely voice-activated devices. Especially in the case of tablets and smartphones, we must distinguish between Web search and search within specific applications: For searches within apps, only a limited amount of data has to be searched, whereas Web search is about “the whole thing,” i.e., the most complete representation of the Web possible. No matter which device we use for searching: (Web) search is a central part of our Internet use. However, as we will see, user behavior differs depending on the context (e.g., mobile vs. at home) and device (e.g., large screen on a laptop vs. small screen on a smartphone). Search engines are adapted to deliver adjusted results and result displays on different devices and in different contexts (see Sect. 7.1 for more details).

## 2.4 Different Pathways to Information on the World Wide Web

Search engines are by no means the only way to access the information on the Web. In the following, we will present the different types of access and related systems. We will then put them into relation to search engines, which will again be the exclusive focus of the subsequent chapters. We will start with Web search engines themselves, as they are our starting point, and we will then compare the advantages and disadvantages of the other systems with them.

In general, a distinction can be made between search engines and other systems:

- Search engines include general search engines, vertical search engines, hybrid search engines, and metasearch engines.
- Other systems include Web directories, social bookmarking sites, question answering sites, and social networks.

To understand the idea of search engines, it is essential to realize that the different ways of accessing Web content also have different objectives. For example, it would be unfair to compare the scope of the databases of search engines and Web directories, as they have very different requirements regarding the comprehensiveness of their databases. It is also relevant whether a system aims to support ad hoc searches (i.e., searches based on the input of a search query) or whether the system is to support browsing of content or monitoring specific sources. For example, the latter is the case with social networks, where users “follow” people or accounts by subscribing to their new messages. This means that content from these accounts is displayed regularly without having to repeatedly conduct a new search.

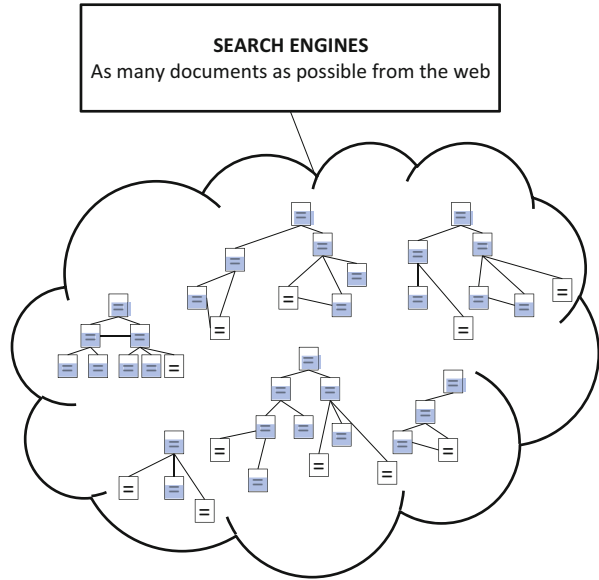
### 2.4.1 Search Engines

When we speak of search engines, we usually mean Web search engines (also known as general, universal, or algorithmic search engines). These engines claim to cover the content of the Web as completely as possible and, if necessary, to enrich it with additional content (see Sect. 3.2). Figure 2.2 schematically shows which contents of the Web search engines cover. The cloud represents the universe of the Web. It contains a multitude of documents stored within websites (illustrated by the hierarchical structure of documents). The content that is captured by the search engine is highlighted. Although search engines aim to capture the total content of the Web, this objective is not achieved and cannot be reached either. We will look at the reasons for this in more detail in Chap. 3.

Nevertheless, search engines achieve greater coverage of the Web than any other type of search system. This is due, on the one hand, to their universal claim and, on the other hand, to the fact that they capture the content automatically. This process is described in detail in Chap. 3; at this point, it should suffice to say that search engines can capture a huge number of documents on the Web and make them searchable.



**Fig. 2.2** The contents of search engines



**Fig. 2.3** Start page of the AltaVista search engine (1996); <https://web.archive.org/web/19961023234631/http://altavista.digital.com/>

### The Concept of the Algorithmic Search Engine in the 1990s

The idea of the search engine as we know it already evolved in the early days of the Web. Early search engines such as Lycos and WebCrawler already worked on the same principle as Google and other search engines do today: They gather the pages available on the Web by following links and return ranked lists of results in response to search queries. This process is fully automatic.

(continued)

Perhaps the best way to illustrate the similarity between earlier and today's search engines is to look at the homepage of AltaVista, the leading search engine at the time, in 1996 (Fig. 2.3).

Firstly, the similarity with today's search engines like Google is striking: There is a centrally placed search field, next to which is a button that can be used to submit the search. In principle, users can enter whatever they want without having to learn a specific query language. Whether single words, whole sentences, or questions: it is up to the automatic processing of the search engine to deliver results that match the search queries.

Secondly, the AltaVista homepage contains information about the size of its database. It states 30 million documents – a large number at the time, considering that the Web was still in its infancy. In the meantime, the Web has grown many times over, but the challenge of capturing its content in a complete and up-to-date manner and making it available for search has remained (see Chap. 3).

Thirdly, it should be pointed out that AltaVista made its search results available via other portals, including Yahoo, as early as 1996. Even then, many providers did not build their own search engines but used the results of one of the big search engines in cooperation. We will return to such cooperation in Chap. 8 and see its influence on the current search engine market.

However, the differences between then and now should not be concealed. Already above the search box, some things are different from today's search engines: With AltaVista, you could choose between different search modes directly on the start page, in this case, between the preset simple search with only one search field and an advanced search. Already in the simple search, you could choose the collection to be searched (default: Web) and the format in which the search results would be displayed. In later chapters, we will get to know the advanced search and different result presentations.

The texts on the AltaVista home page around the search box are also illuminating. On the one hand, a link to a mirror site is offered; such "mirrors" are nothing more than copies of websites available in another geographical location, in this case, Australia. Internet connections in 1996 were much less developed than they are today, and one often had to wait quite a long time for responses from remote Web servers. Mirrors were created to shorten these waiting times. Today, search engines have data centers spread around the world that distribute the search engine's database and the processing of search queries. However, users no longer have to select one of these data centers explicitly, but both the index and the processing of search queries are distributed automatically.

## 2.4.2 Vertical Search Engines

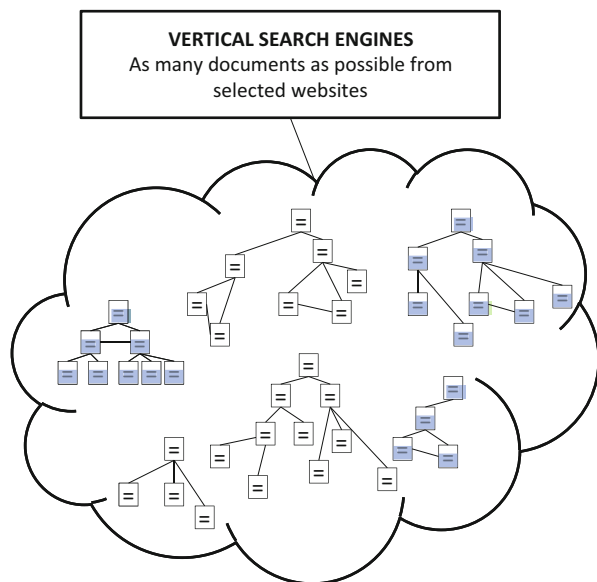
There is a distinction to be made between general search engines and vertical search engines. Vertical search engines aim to capture as many documents as possible from selected websites. The term “vertical search engines” is often used instead of “special search engines”; in this terminology, universal search engines are referred to as horizontal search engines.

Vertical search engines are restricted to a specific topic and thus make a more targeted search possible. The ranking can be specially adapted to the documents they index, as can the subject indexing of the documents. Finally, there are also advantages in the presentation of the results, which can be adapted to the individual purpose of the vertical search engines and the proficiency of the target audience. The fact that vertical search engines cannot be replaced by universal search engines results from the problems of the latter (for a detailed explanation, see Sect. 6.1):

1. Universal search engines have technical restrictions and (despite the label universal) cannot cover the entire Web.
2. There are financial hurdles that restrict the collection of content and its indexing.
3. Universal search engines are geared toward the average user.
4. They have to provide consistent indexing of all content so that everything is searchable together.<sup>1</sup>

Vertical search engines intentionally restrict themselves to a specific area of the Web (see Fig. 2.4). Usually, they are limited to certain sources, i.e., websites. These

**Fig. 2.4** Contents of vertical search engines



<sup>1</sup>Of course, universal search engines can carry out individual indexing for certain types of content or certain databases. However, this cannot be done for the mass of offerings to be indexed.

websites are typically selected by hand. For example, if one wants to build a vertical search engine for news, it makes sense first to compile the relevant news websites, which the search engine then continuously scans for new content (pages).

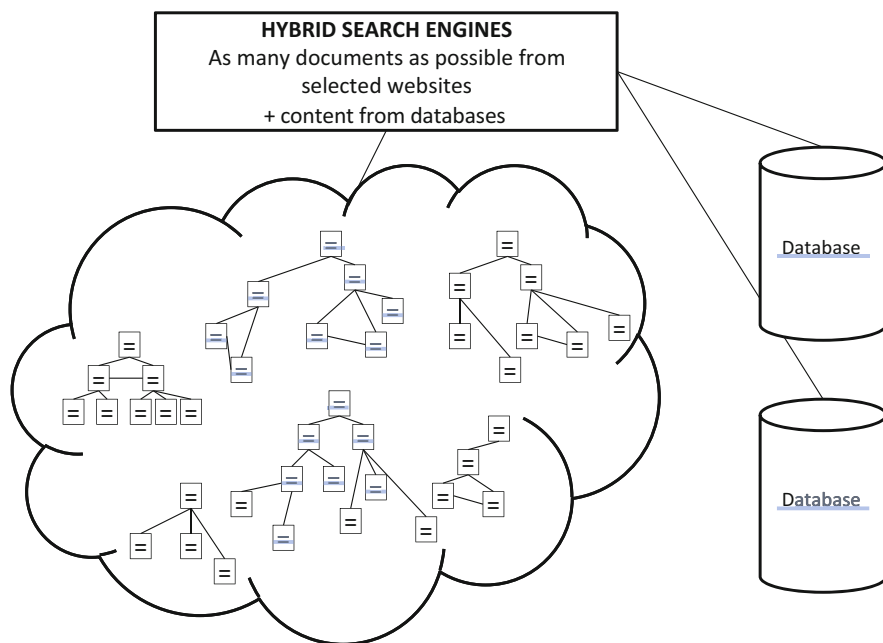
A website is a self-contained offering on the Web that can contain several Web pages. Differentiation is made via the domain (e.g., [nytimes.com](https://nytimes.com)), subdomains, or directories (e.g., [archive.nytimes.com](https://archive.nytimes.com) or [nytimes.com/section/world](https://nytimes.com/section/world)).

On the other hand, a Web page is a single document usually composed of text and associated media elements (images, videos, etc.).

### Examples of Vertical Search Engines

Vertical search engines can be restricted to very different topics. Examples include Google News (<https://news.google.com/>), which is restricted to news, and Swiggle (<https://swiggle.org.uk/>), which is restricted to content suitable for children.

Hybrid search engines are a particular type of vertical search engine. Like vertical search engines, they cover a selected part of the World Wide Web but add additional content from databases to the resulting inventory. This database content is not part of the WWW and, therefore, cannot be found through standard search engines (for technical details, see Chap. 14). Figure 2.5 illustrates the hybrid search engine model.



**Fig. 2.5** Contents of hybrid search engines

### 2.4.3 Metasearch Engines

At first glance, metasearch engines look like other search engines. They also provide the user with the same service, namely, potential access to all World Wide Web content. However, they differ from the “real” search engines in that they do not have their own index but, as soon as the user enters a query, they retrieve results from several other, “real” search engines, merge them, and display them in their own results display (see Fig. 2.6).

The idea behind metasearch engines is that no one search engine can cover the entire Web. Therefore, combining the results of several search engines that cover different areas of the Web would be worthwhile. A second advantage is supposed to lie in a better relevance ranking of the results since the best results are already fetched by each of the giving search engines, from which a ranking of the best is then created.

However, there is considerable criticism of the concept of metasearch engines, which is mainly directed at the fact that the supposed advantages of metasearch are claimed but not empirically proven (Thomas, 2012). It can also be argued that metasearch is an outdated idea, as today’s search engines no longer have the coverage problems that search engines had in the 1990s, when the concept of metasearch engines was born. At the very least, the benefits of better coverage only play a role in a few cases today.

Even the supposed advantage of ranking no longer exists today, at least not to the same extent as it did in the past: For one thing, the universal search engines have become far better in this respect, and for another, metasearch engines do not have access to all the documents of the providing search engines (see Fig. 2.6). Rather,

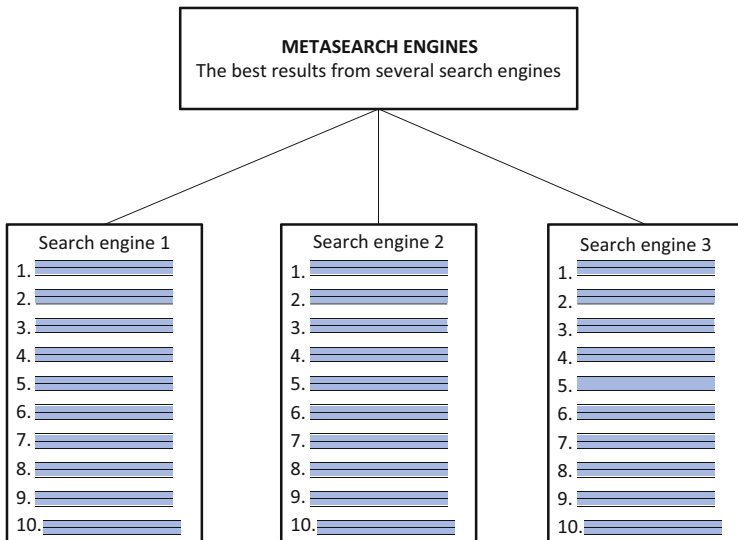


Fig. 2.6 Contents of metasearch engines