



Mathias Erwin Glatz, BSc

Access to Knowledge on the Web

DIPLOMA THESIS

to achieve the university degree of

Magister der Naturwissenschaften

Teacher training programme: Computer Science and Computer Science Management

submitted to

Graz University of Technology

Supervisor

Em.Univ.-Prof. Dr.phil. Hermann Maurer

Institute for Interactive Systems and Data Science

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present diploma thesis.

Date

Signature

Abstract

The size of the web is incredibly huge and still growing massively. Without search engines, users would be hopelessly lost and the web would not be that popular. Despite the fact that search engines have improved over the years there are still many drawbacks and problems concerning the presentation of results. This work tries to point out some massive disadvantages of Google's structure of a search engine result page. The main contribution of this thesis is a new approach for a search engine - a so-called search assistant. A search assistant tries to get rid of the issues of current search engines. Furthermore, more convenience for users is provided by the automatic generation of a report concerning the submitted search query. This approach is proposed and discussed.

Further on, an implementation of a very early version of a search assistant is introduced and analysed. It is concluded that this approach actually works. However, some challenges remain to be mastered in future work.

Abstrakt

Das World Wide Web ist von unvorstellbarer Größe und wächst ständig weiter. Ohne Suchmaschinen wären Nutzer hoffnungslos verloren und das Internet wäre nicht so populär, wie es heute ist. Obwohl die Technik der Suchmaschinen immer weiter voranschreitet, gibt es immer noch massive Probleme in der Darstellung von Suchergebnissen. In dieser Arbeit wurde versucht, anhand der Ergebnisseite von Google diese Nachteile hervorzuheben und zu beschreiben. Der Hauptteil der Diplomarbeit war das Vorstellen und Analysieren einer neuen Art von Suchmaschine, einem sogenannten Suchassistenten. Der dargestellte Ansatz versucht die angesprochenen Probleme zu lösen und dem Nutzer zusätzliche Bequemlichkeit zu verschaffen, indem automatisiert Berichte über die eingegebenen Suchbegriffe generiert werden.

Neben der Diskussion der Vor- und Nachteile dieses Ansatzes, wird auch eine Implementierung einer ersten Version eines Suchassistenten besprochen und analysiert. Aufgrund dessen wird schlussgefolgert, dass die vorgestellte Idee eines Suchassistenten im kleinen Rahmen durchführbar ist, obwohl es noch diverse Probleme und Herausforderungen zu meistern gilt.

Contents

Abstract	iii
1. Introduction	1
2. World Wide Web and Search Engines	5
2.1. World Wide Web	5
2.2. Crawling and Indexing	6
2.3. PageRank – First Sophisticated Ranking Technology	8
2.3.1. Example Calculation for PageRank	9
2.3.2. Googlebombing	11
3. Today’s Ranking Factors on Google	12
3.1. List of Ranking Factors on Google	12
3.2. RankBrain – Analysing Queries using Machine Learning Technologies	16
3.3. Personalisation	16
4. Search Engine Result Page (SERP)	18
4.1. Structure	20
4.2. Implications, Drawbacks and Suggested Improvements	35
4.2.1. Ambiguous Search Terms	35
4.2.2. Knowledge Graph	39
4.2.3. People Also Ask (PAA)	47
5. The Future: Search Assistance	48
5.1. Main Idea of a Search Assistant	49
5.2. Search Assistant Applied on a Small Domain	52
5.2.1. Motivation for the Domain Chosen	53
5.2.2. Example of a Conventional Search on Domain Chosen	53

Contents

5.2.3. Search Assistance on Domain Chosen	55
5.3. Discussion	58
5.3.1. Improvements in Search	60
5.3.2. Challenges and Problems	63
6. Discussion of a First Version of a Search Assistant as Implemented – Sistant	66
6.1. SERP in Sistant	67
6.2. Server Configuration in Sistant	73
6.3. Summary	73
7. Summary and Future Work	76
A. Implementation Details of Sistant	79
A.1. Libraries Used	79
A.2. JSON Files	80
A.2.1. File Handling	80
A.2.2. Regions.json	81
A.2.3. Interests.json	82
A.2.4. City Files	83
A.3. Fetching and Parsing of HTML	86
A.4. Supported Countries and Cities	87
A.5. Supported Sources per Category	88
Bibliography	89

List of Figures

2.1. PageRank Example Graph	9
4.1. Type of Search on Google	21
4.2. Google Instant Example	21
4.3. Search settings on Google	21
4.4. Advanced Search on Google	22
4.5. Search Restrictions on Google	23
4.6. List of results for "Computer Science" on Google	24
4.7. List of results for "Climate Change" on Google	25
4.8. PAA for "Healthy Food" on Google	26
4.9. Snippet for "Computer Science" on Google	27
4.10. Knowledge Graph for "Arnold Schwarzenegger" on Google	29
4.11. Knowledge Graph for "Vladimir Putin" on Google	30
4.12. Knowledge Graph for "Twitter" on Google	31
4.13. Knowledge Graph for "Granola" on Google	32
4.14. List of movies with Arnold Schwarzenegger	33
4.15. Related searches for "Computer Science" on Google	34
4.16. Suggested location by Google	34
4.17. Help, Feedback on Google	34
4.18. SERP for "Galaxy" on Google	37
4.19. Knowledge Graph for "Graz University of Technology" on Google	40
4.20. Knowledge Graph for "London" on Google	42
4.21. Population for "London" on Wolfram-Alpha	43
4.22. List of Austrian Nobel laureates on Google	43
4.23. List of Austrian Nobel Prize winners on Wolfram-Alpha	44
4.24. Knowledge Graph for "Suck Me Shakespeer 2" on Google	46
4.25. Area of "United Kingdom" on Global-Geography.org	47

List of Figures

5.1. Upper SERP for "Vienna" on Google	54
5.2. Lower SERP for "Vienna" on Google	56
5.3. Upper SERP for "Ryan Reynolds" on Google	57
5.4. Possible structure of a SERP by a search assistant	59
6.1. Hint in Sistant	67
6.2. Header in Sistant	68
6.3. Selectable interests in Sistant	68
6.4. Image category for London in Sistant	69
6.5. Key facts category for London in Sistant	70
6.6. General Information category for London in Sistant	71
6.7. Points of interest grids for London in Sistant	72
6.8. Server configuration for London in Sistant	74

1. Introduction

The size of the world wide web is tremendous and still growing massively. Without search engines, users would not have any chance to find information on a specific topic without knowing a URL. Already in the early days of the web, the desire to "know" what is available on the web was there.

Following the timeline in (Wall, 2017): The first approach for a search engine was a program that downloaded lists of directories and files located on public anonymous FTP (File Transfer Protocol) sites and created a searchable database based on words which it would match with search queries submitted by users. It was created by Alan Emtage, a student at McGill University in Montreal in 1990.

Two years later, in June 1993 the first web robot was introduced by Matthew Gray. Originally, it was a crawler to obtain the size of the web. Later in 1993, the first web search engine emerged. It was a crawler that downloaded meta information of a page and created an index out of it, called ALIWEB. Webmasters could submit their websites to be indexed with an additional description. A drawback of this solution was that many users did not know how to and where to submit their sites.

In the same year, JumpStation was invented. It crawled title and header of a site and retrieved them using a linear-search approach. At the same time, WWW Worm had surfaced on the web. It indexed titles and URLs. Both approaches had the drawback that they list results in the order they were found. Thus, no ranking was available. Repository-Based Software Engineering (RBSE) did solve that problem. In April 1994, WebCrawler, a search engine that indexed whole webpages was released. It became popular to the extent that it could not be used during daytime. Since then, this approach has become standard for all major search engines. In parallel, many web directories were created, often edited by humans. A web directory

1. Introduction

is a list of websites, containing a name, description, and an URL. In many cases, the list is split into different categories to ease browsing for users. New entries may be added either for free or for a charge. Some providers of web directories offered a search functionality for their lists as well. An example is Yahoo which outsourced the searching technology for many years to third-parties. Within the following years, many search engines arose. For instance, Lycos, Yahoo, Infoseek, BackRub (former Google), Ask Jeeves, MSN, Baidu, and Altavista.

Today, Google has become the top search engine. Currently, it is the most used search engine with a market share of about 77.4% on desktop computers and incredible 90.4% on mobile devices (NetMarketShare, 2017).

Over the years, many improvements in the technology of searching have been realised. Still even today, presentation of results and interaction with the user are constantly evolving. By starting with a list of (possibly) useful URLs much more information is returned by a search engine nowadays. Besides a list of (relevant) URLs (and short snippets extracted), an explanation of the search term (a featured snippet), related objects or things (within a so-called knowledge panel), and related searches (usually at the bottom) are shown for specific search terms. For instance, these are famous persons, companies, well-known objects, and others. The main focus of a search engine is to retrieve desired search results, independent from the search query.

This work exploits the main tasks of a search engine. These are crawling, indexing, ranking, and finally presenting information to users. Special focus is laid on the current presentation of results on Google - a so-called search engine result page (SERP). Although a SERP on Google has made a very good progress, there are many issues and drawbacks still existing. In this work, some major critical aspects are pointed out and analysed. These are ambiguous search terms, the lack of sources stated for data presented, and the lack of definitions for data presented.

Imagine the search query "Galaxy". Obviously, this search term is ambiguous. A user may look for an astronomical system. On the other hand, the user might be a soccer fan. Then, a soccer team containing the term "Galaxy" in its name could be meant. Another way of interpretation is something related to a series of smartphones, also containing "Galaxy" in its name.

1. Introduction

Furthermore, almost every search query is ambiguous. For instance, searching for "electricity". By default, this search query does not reveal the user's intention. A user might want to have information about technology on how to store electricity. Alternatively, information about technology to distribute it or its impact on the environment is what a user wants. There are many more aspects. To point it out, a search engine does not know what users mean. To solve this issue, the search engine has to allow interaction with its users. It could ask users in which topics they are interested to know better what users want to see.

Further on, considering a search for a city, for instance, "London". On the current SERP, users get, besides a list of ranked results, images and some key facts for the city. For instance, area and population are stated as two numbers. Those data-items are displayed within a knowledge panel. What is surprising that no source for these data is stated, major point of criticism in this work. Further on, it is not specified how these data presented should be interpreted. Looking at the population, nowhere it is stated to which area it applies to. Possible choices are the metro area, city area, or an intermediate of both. To point it out, it is essential that the source of data has to be stated and how data should be interpreted.

The main contribution of this thesis is a new approach for a search engine - a so-called search assistant. It is to mention, that many of the ideas proposed emerged in discussion with the supervisor, Hermann Maurer. On the one hand, its purpose is to get rid of the problems mentioned before. This is done by an improvement of interaction between the search engine and its users. On the other hand, convenience for users is enhanced. This is done by a generation of a so-called report about a search term. The content in these reports is fetched from chosen websites, all stated with its source so that the user has the possibility for further reading. As these sources are fetched in real-time and furthermore displayed to the user only, no copyright infringements are thus possible. In addition, users can rate different aspects of the sources used, for instance, trust, impartiality, readability, and others. Further on, an extension is proposed. Reports could be split into different categories. Imagine, a search for a city. Users might be interested in images, videos, key facts, general information, points of interest, and history. A report generated could consist of these categories with the possibility for the user to hide unwanted sections.

1. Introduction

Further on, challenges appearing are discussed briefly in this thesis. The avoidance of redundancy and the selection of high-quality sources are non-trivial problems. Also, the support for sources containing content in a different language than desired and the effort of real-time generation need further discussion. Furthermore, determination of possible categories for reports is a hard challenge. Anyway, much of this is beyond this thesis.

Moreover, to test the feasibility of the approach proposed, a first version of a search assistant was implemented on a small domain. The domain chosen was "some major cities of German and English speaking countries". For that, a web application based on JSP (Java ServerPages) was developed. Technologies used are Java, Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript, and JSON (JavaScript Object Notation). Finally, the implementation was discussed briefly.

In this thesis, many screenshots are used to provide a better understanding of topics. Some of them contain data and images which might not be open-source. Therefore, these parts are disguised by blurring.

2. World Wide Web and Search Engines

In this chapter, an overview of the role and techniques of the currently most frequently used web search engines is given. These include some statistics about the web, the process of crawling and indexing, and the former key concept of ranking on Google – PageRank. For PageRank, an example and a way used to evade its purpose are given.

2.1. World Wide Web

The information on the web is incredibly huge, widespread, and still growing. The number of websites grew from about 207 million in 2010 to more than 860 million in 2015 (Internet Live Stats, 2017). In addition, the number of search requests rose incredibly. In 2010 100 billion search requests were counted on Google worldwide, this number increased continuously to about 3.300 billion requests in 2016 (Internet Live Stats (2000-2012) and Statistic Brain Research Institute (2013-2015), 2017). It is a hard challenge for search engines to filter this huge amount of information and give users information they might want to see.

The most used search engine on the web is Google¹ (NetMarketShare, 2017). Worldwide more than 77.4% (March 2017) of all desktop searches are sent to Google. There is a huge gap between Google and the second most popular search engine, Baidu², which has a market share of about 8.1%. Bing³ follows

¹<https://www.google.com>

²<https://www.baidu.com>

³<https://www.bing.com>

2. World Wide Web and Search Engines

with a market share of about 7.3%. Other search engines are Yahoo⁴ (5.6%) and Ask⁵ (0.2%).

Statistics with similar results show the usage of search engines for mobile devices as smartphones and tablets (NetMarketShare, 2017). In that area, Google is even more powerful with a market share of about 90.4%. Yahoo (5.3%), Bing (1.5%) and Baidu (1%) are very much below this.

This work mostly focuses on the behaviours of the search engine Google, as it is the most used search engine. Bing and Yahoo behave similar. Other search engines as Wolfram-Alpha⁶ are examined less. The behaviour of Baidu is neglected since it is mainly used in Asian countries but not in Europe. Ask is neglected as well due to its tiny market share.

Search engines highly rely on their web index created by robots, so-called web crawlers. Statistics show that the amount of traffic caused by bots (51.8%) on the internet exceeds the traffic caused by humans (48.2%) in 2016 (Zeifman, 2017). From traffic caused by bots, about 6.6% is due to crawlers from search engines. The others include feed fetchers (bots that ferry content for mobile and web applications), impersonators (bots that assume false identities to bypass security solutions), commercial crawlers, and hacker tools. The size of Google's page index is stated as hundreds of billions of web pages (Google, 2017e). It exceeds 100 million gigabytes (Google, 2017a).

2.2. Crawling and Indexing

The process of crawling starts with a list of web addresses to be crawled and indexed. Major search engines provide webmasters with a way to add their own sites to this list. A crawler visits a website mentioned in the list and collects its information. The first document on the web space to be parsed (if it exists) is a file named "robots.txt". Usually, it contains information about firstly if, secondly, which parts and thirdly, how often a website

⁴<https://www.yahoo.com>

⁵<https://www.ask.com>

⁶<https://www.wolframalpha.com>

2. World Wide Web and Search Engines

should be crawled. So-called "user-friendly" crawlers take care of this information. Harvester, spambots, and other robots might not cooperate with this standard. Major search engines, like Google (Google, 2017j), Bing (Bing, 2008), and Yahoo (Yahoo, 2017) make use of this standard.

Another approach providing search engines information about the website is the use of Sitemaps (Sitemaps.org, 2008). It contains a list of URLs and corresponding metadata about the site. For example, date of the last modification, change frequency, and priority. Some major search engines such as Google, Bing, and Yahoo support Sitemaps. Google recommends using them if the site is large, the site has large archives of content pages that are not linked to each other, the site is new or has few links to it, or the site uses rich media content (Google, 2017g). According to Google (Google, 2017j), a Sitemap does not guarantee that all items in the Sitemap will be crawled. However, in most cases, the site will benefit from having a Sitemap; at least it will never be disadvantaged.

Further on, the crawler visits the different sites and navigates through all links of a page to discover other pages. It also pays attention to new sites, modifications of the site and dead links. Whenever a crawler finds a page it is rendered by the search engine as a browser does it (Google, 2017f). Keywords, titles, heading, the freshness of the site, mobile representation, content, and many more attributes are analysed and sent back to the server. Furthermore, all the information is stored in a so-called web index.

A web index can be compared to an index of a book. It stores information about words and its position(s) within the document. Broadly speaking, search engines go through their index to find relevant documents to the user's search query.

Imagine a website containing 10 links to other pages. By a depth of 5, this results in 510 pages to be crawled (neglecting backlinks and loop links). That is an incredibly huge number to be crawled and indexed. Thus, information gained by crawlers grows extremely fast at the beginning of the crawling and indexing process.

An aspect of a good search engine is its independence of third-party companies that offer payment for benefits without displaying it. Google states that

2. World Wide Web and Search Engines

they do not accept payment for crawling a site more frequently (Google, 2017f).

2.3. PageRank – First Sophisticated Ranking Technology

Today's version of Google's search engine is based on the ideas of Lawrence Page and Sergey Brin in 1998 (Brin and Page, 2012). A major part of this search engine was a precise page ranking. The concept was named PageRank and is explained in this section.

Analogous to (Brin and Page, 2012): Google has an important feature to produce high precision results in form of using the link structure of the Web to calculate a quality ranking for each web page. Counting the citations and backlinks for a page gives an approximation of the importance of it. PageRank extends this idea by not weighing all links to a page equally and normalizing the number of links on a page. The PageRank of a page represents the probability a user visits this page.

- Assume page A has pages T_1, \dots, T_n pointing to it.
- A parameter d is set between 0 and 1 and states a probability that the user clicks on a link on the current page. Usually, d is set to 0.85.
- $1 - d$ states the probability that the user jumps to another page (by not clicking a link) or stops surfing.
- Then, $C(A)$ is defined as the number of outgoing links from page A .
- T is the total number of pages on the web graph.

Then, the PageRank of page A is calculated recursively as

$$\text{PageRank}(A) = \frac{1 - d}{T} + d \sum_{i=1}^n \frac{\text{PageRank}(T_i)}{C(T_i)}$$

The sum of all web page's PageRank is 1. This is due to the division by T . Initial values are set as $T_i = \frac{1}{T}$, after each iteration the PageRank values get

2. World Wide Web and Search Engines

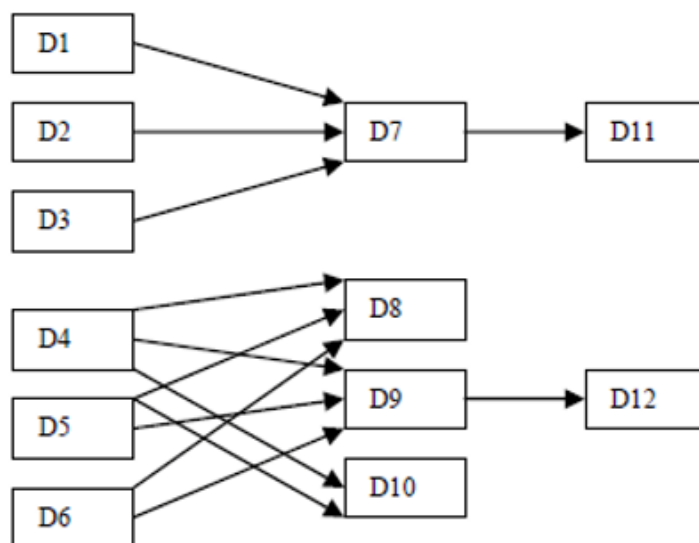


Figure 2.1.: Example graph for pages and links

normalized:

$$c = \frac{1}{\sum PageRank'(T_i)}$$

and $c \cdot PageRank'(T_i) \rightarrow PageRank(T_i)$. If the page has no outbound links, set $d = 1$. Thus, the probability to jump to another page is equally distributed (Gütl, 2014).

2.3.1. Example Calculation for PageRank

The following example is taken from (Gütl, 2014). Given is the graph in Figure 2.1, D_i ($1 \leq i \leq 12$) denote web pages, arrows denote links to other pages. d is set to 0.85. Table 2.1 shows the calculation of all PageRanks in three iterations including normalization.

An early experience from Page and Brin was the convergence of iterations (Gütl, 2014). This is due to a theorem in Markov chain theory that can be formulated in the sense of pages and steps (respectively "clicks on links") as (Golub, 2014):

2. World Wide Web and Search Engines

	It. 1	N.	It. 2	N.	It. 3	N.
D1	0.0125	0.0174	0.0125	0.0125	0.0125	0.0336
D2	0.0125	0.0174	0.0125	0.020	0.0125	0.0336
D3	0.0125	0.0174	0.0125	0.020	0.0125	0.0336
D4	0.0125	0.0174	0.0125	0.020	0.0125	0.0336
D5	0.0125	0.0174	0.0125	0.020	0.0125	0.0336
D6	0.0125	0.0174	0.0125	0.020	0.0125	0.0336
D7	0.2250	0.3140	0.0570	0.092	0.0640	0.1720
D8	0.0951	0.1328	0.0298	0.048	0.0325	0.0874
D9	0.0951	0.1328	0.0298	0.048	0.0325	0.0874
D10	0.0597	0.0833	0.0224	0.036	0.0239	0.0643
D11	0.0833	0.1163	0.2794	0.452	0.0908	0.2439
D12	0.0833	0.1163	0.1253	0.203	0.0534	0.1435

Table 2.1.: Calculation of PageRanks for each page D_i in three iterations

“The probability (before the surfer starts surfing) that he is at a particular page X in k steps tends to a fixed number and this number is independent of the starting point as k grows large.”

For pages with 322 million links, it takes 52 iterations, in general $O(\log n)$, where n is the size of the network (Gütl, 2014).

Hence, a page has a high PageRank if there are many pages pointing to it or there are some high ranked pages pointing to it (Brin and Page, 2012). Further on, they stated that this is quite intuitive as important pages would not point to a low-quality page and well-cited pages from many places are probably worth looking at.

Using the knowledge of how the PageRank algorithm works, it is an easy task to improve the PageRank of a specific page. For example, a webmaster could create several different pages, all pointing to a single page. This results in a higher PageRank for the single page. It is called “link farm”, and (if found) is excluded from the index by Google (Gütl, 2014).

Back in 1998, Brin and Page implemented further factors that improved their ranking technique (Brin and Page, 2012). For instance, Google has location information for all hits and make use of proximity in search. Additionally,

2. World Wide Web and Search Engines

they mentioned that words in a larger or bolder font are weighted higher than other words.

In their work (Brin and Page, 2012) Brin and Page proposed improvements for ranking by personalising PageRank. They suggested considering the user's homepage or bookmarks to rank results.

2.3.2. Googlebombing

In the 2000s, the term "Googlebomb" got widespread attention. Googlebomb is a technique to relate a website to one or more actually unrelated search terms. This is possible because Google associates the text of a link with the page the link points to (Brin and Page, 2012).

With this constructed relation, a search query for the (unrelated) terms results in a high ranking of the corresponding website. For example, in 2004 the search terms "miserable failure" led to the biography of George Bush jun. (Cutts, 2009). In 2009, after some other Googlebombs got global attention, Google developed an algorithm to detect them automatically. They also declared that this algorithm is not running all the time as it takes some computing power to scan the entire web index. Google rejected that Googlebombing is a sign of political bias (Cutts, 2009).

3. Today's Ranking Factors on Google

Since the introduction of PageRank in 1998, many adaptations and extensions, not only for PageRank, have been realised (Google, 2017f). An accurate ranking is a key competence of a search engine, although this is not a trivial thing at all (Maurer, 2012). Most of the latter extensions of the ranking techniques are Google's secret and not meant for the public (Gütl, 2014). In this chapter, a list of the most influencing ranking factors for a search query on Google is given. The information presented is gathered from different sources on the web, including leading companies that offer search engine optimisation (SEO). Further on, a short overview of a machine learning technology for analysing search queries on Google – RankBrain, is given. Last but not least, the influence of personalisation on search results is described.

3.1. List of Ranking Factors on Google

Many companies are dealing with its expertise on SEO, for instance Link-Assistant.Com¹, Backlinko², Searchmetrics³, and Unamo⁴. They offer knowledge about the most important ranking factors on Google, based on research, experiments, and statistics.

¹<https://www.link-assistant.com>

²<https://www.backlinko.com>

³<https://www.searchmetrics.com>

⁴<https://www.unamo.com>

3. Today's Ranking Factors on Google

More than 200 ranking factors are known (Dean, 2016a). Several websites assume approximately this number of ranking factors and offer lists, pointing out the most influencing ranking factors (Maksimava, 2016), (Grabowski, 2014). Some of them are proven, some of them are speculation, and not all these factors are equally weighted (Dean, 2016a).

The following list describes the most important ranking factors and its influence on the ranking based on (Searchmetrics, 2016):

1. Content Relevance

Of course, the content of a website and its URL have a direct influence on its importance. A study (Dean, 2016b) discovered that a site about a single topic has a higher importance than a site touching more topics just briefly.

Another indicator is the title of the site. Is a keyword in the beginning of the title, the site is considered as more important (Dean, 2016a), (Grabowski, 2014). This influencing factor decreased within the last years, only 45% of the top-20-URLs had the keyword in the title in 2016 (Searchmetrics, 2016).

Other signals for relevancy are the usage of meta description tag, H1 and H2 tag (Grabowski, 2014). In addition, images, audio, and videos should have alt text, caption, and description (Grabowski, 2014). Contrary, duplicate content can shrink the ranking of the site. URLs ranked between positions three and six have the highest content relevance as ranking position one (and sometimes two) are mostly considered as "brand-links" (Searchmetrics, 2016). These brand-links are sites that got a higher ranking because they are considered as trustworthy for the query by Google.

A smaller influence is the content length of a website. SEOs discuss this influence but there is no clear sign what the optimal length is, as it varies from niche to niche (Maksimava, 2016).

2. User Signals

Google uses user signals to get information about the quality of its ranking. These data are taken from the user behaviour during the search process, Google Chrome, Google Analytics, and Android among others (Searchmetrics, 2016).

3. Today's Ranking Factors on Google

The click-through rate (CTR) has a massive influence on the ranking. Taken from (Maksimava, 2016), CTR is "a ratio of the number of times a given search listing was clicked on to the number of times it was displayed to searchers". A study found that the CTR has even the highest influence on ranking out of all factors examined (Searchmetrics, 2017). The average CTR of position one to three is about 36% (Searchmetrics, 2016). For results on page two (results 11-20) and above it is estimated to be less than 10% (Sweeny, 2011).

A real-time experiment (Fishkin, 2015) in June 2015 showed the influence of the CTR directly. Users on Twitter were asked to query "best grilled steak" on Google, then click on the first result, click back as soon as the page loads, click on the fourth result and dwell on it. After about 70 minutes and 4-500 interactions, the fourth result moved to the top result. As this is not a scientific approach there is no proof that Google uses CTR.

Hence, the dwell time on the clicked site is also an important user signal (Searchmetrics, 2016). Actually, this depends on the intention of the user. For instance, if users are searching for a quick answer to a specific question, a short dwell time is an indicator for a "good" site. If users search for a broader answer, a long dwell time is an indicator for a site offering many good aspects. Therefore, a good search engine knows the intention of users submitting the query. Google uses RankBrain to analyse entered queries using machine-learning technologies, see Section RankBrain – Analysing Queries using Machine Learning Technologies. Contrary, Gary Illyes from Google declined that CTR is actually used for ranking, as it would be too easy to use it for manipulations (Slegg, 2015).

3. Technical Factors

A rather new influence on ranking is the mobile-friendliness of a website. In November 2016 Google announced that they have begun to index the mobile version first as most of the searches come from mobile devices and are therefore more relevant (Google Webmaster Central Blog, 2016). Furthermore, they revealed that eventually the mobile version of a site's content will be used for ranking.

In April 2010, Google confirmed to use the response time of a website as a ranking factor (Google Webmaster Central Blog, 2010). They integrated this

3. Today's Ranking Factors on Google

feature because faster responses create happier users and lower operating costs. The use of secure communication, as HTTPS, can have a positive influence on the site's rank, about one-third of the top-10 ranked sites are secured using HTTPS (Searchmetrics, 2016).

4. User Experience

Google tries to estimate the user experience of a site. Internal links are an important ranking factor (Searchmetrics, 2016). If there is a clear and structured way to navigate through the whole site then this can boost the ranking. In addition, external links can influence the estimated user experience.

Studies show that the more menus, buttons, and other interactive elements a site contains the better is the ranking (Searchmetrics, 2016). Further on, other factors for good user experience are the use of lists, bullet points and the exclusion of Adobe Flash content.

5. Social signals

Experiments showed that pages with more social interaction rank better (Grice, 2012). Facebook is considered as the most influencing social signal (Searchmetrics, 2016). Still, the discussion whether or not social shares influence the ranking is ongoing (Grice, 2012).

6. Backlinks and Linking Domains

As already discussed in the previous chapter, backlinks were the main concept to derive the importance of a website in the past. Due to its drawbacks (creation of link farm to gain a higher rank), it partly lost its importance. Still, more backlinks will result in a higher ranking, as long as the linking pages are not considered as low quality (Maksimava, 2016). This would have a reverse effect.

Further on, the number of different domains linking to a site has influence, the more different domains the better (Maksimava, 2016). A high-quality link (a link from a high-ranked website) can boost a site; lower quality backlinks may penalize the ranking of the site.

3. Today's Ranking Factors on Google

Domains of sites, which Google trusts, are ranked higher; also, the server location has an influence on ranking for that region or country (Grabowski, 2014). Other influences mentioned are domain registration length (domains registered for more than a year are ranked higher) and domain history (if a spammer has registered a domain before it will be considered less trustworthy).

The anchor text of a backlink is used by Google to get information about the content of the linked page (Maksimava, 2016). The title of the linking page is used much weaker to classify the content of the linked page. With this information, it is easier for Google to determine if two pages are related to each other. This could influence the ranking. Earlier in the lifetime of Google, this effect was used for Googlebombing, see Section Googlebombing.

3.2. RankBrain – Analysing Queries using Machine Learning Technologies

In 2015, Google invented a new technology to get more information about the intention of users submitting a query, called RankBrain (Sullivan, 2016). As stated, it is part of Google's overall search algorithm and the third-most important signal.

In 2013, about 15% of queries submitted to Google had never been entered before (Farber, 2013). According to (Sullivan, 2016), most of them are "long-tail" queries. Further on, RankBrain tries to interpret and translate these long-tail queries to find relevant sites. It can discover relations between seemingly unconnected searches by calculating the distance between them. Finally, it is assumed that complex queries are translated to easier queries to find relevant sites.

3.3. Personalisation

Personalisation has a huge influence on the result page of a search engine. Google considers the search history of the currently used browser, inde-

3. Today's Ranking Factors on Google

pendent of whether the user is logged in or not (Föller, 2014). It is also mentioned, that if a user clicked on specific domains in a previous search it is likely that these domains will rank higher in a similar search in the future.

Another personal factor for ranking is the current (network-based) location. A user in Graz gets different results searching for "local music" than a user in Vienna. Assuming users enable their current location.

Even the type of browser influences the ranking results (Föller, 2014). By using Safari and searching for software, it is more likely to get Apple-specific results. Other factors are the current time and selected language. There are probably many more factors, which are Google's secret.

4. Search Engine Result Page (SERP)

A Search Engine Result Page (SERP) is the response of a search engine to a search query submitted by the user. Online search engines usually respond with a Hypertext Markup Language (HTML) file which is presented to the user by a web browser. A list of results is the main component of a SERP. Additionally, a featured snippet (for instance, an extraction of a definition taken from another website) and/or a Knowledge Graph (a knowledge base containing information about a person or an object) are shown for some search terms. Examples of such search terms are famous people and well-known objects. Furthermore, additional advertisements are shown. However, they are neglected here. In this chapter, a close look at the structure of a SERP of Google is given. The different parts are listed and explained. Google was chosen as it has the most sophisticated concept of a search engine and its interaction with users. Also, other search engines behave similarly.

Further on, some implications of the design are pointed out, including some major issues and drawbacks. The main critical aspects are ambiguous search terms (What does the user mean or want to have?), sources of data retrieved (Where do the data in the knowledge base come from?), and interpretation of data retrieved (What does the data in the knowledge base mean?).

For instance, if the user searches for "Galaxy" the search engine does not know exactly what is meant by the user. "Astronomical system" or "a professional soccer team" are only two ways on how the search engine could interpret the search query. To continue this thought, a simple search term is always ambiguous. Consider a search query for "electricity". There are different ways of interpretation as well. Following Wikipedia's site, for

4. Search Engine Result Page (SERP)

"Electricity (disambiguation)"¹, energy, a movie's name, or some relation to music are listed. Anyway, users most likely are interested in electricity in terms of energy. Even with this restriction, users differ in which information they are interested in. Some might want to have a brief overview of electricity, others some particular information about the technology to produce and store it. Furthermore, some might be interested in the distribution of it, or want to have information about its relation to the environment, and so on.

Another example is a search for "London". Google offers some key facts about the city. For instance, the population and area are shown. What they do not do is to state where these data come from, a massive point of criticism. Following the example, Google also does not state how the data should be interpreted. Taking the population, they do not state to which area it applies to. It could be city population, metro population, or a combination of both. For interested users, the data presented in the Knowledge Graph on Google is more or less useless, as the sources and definitions are missing most of the time.

It is the basis of this thesis that search results must be presented in a different and augmented way. The issues addressed are examined and different ways for improvements are suggested. In particular, for ambiguous search terms, the search engine could ask in which categories a user is interested before or after a search query is submitted. After that, the search-area should be restricted to the corresponding categories. For sources, it is obviously necessary to state where the data was taken from. Optimally, the search engine offers different sources. Moreover, the search engine must give at least a definition for the data presented. If data is available in different definitions, the search engine should list all of them. Further on, the search engine should ask for feedback to understand the user. Improvement of interaction between users and the search engine is highly important to find out what a user wants. This is a major aspect demanded in this thesis.

¹[https://en.wikipedia.org/wiki/Electricity_\(disambiguation\)](https://en.wikipedia.org/wiki/Electricity_(disambiguation))

4. Search Engine Result Page (SERP)

4.1. Structure

Google's SERP has several parts, not all of them showing up on every search. Assuming the number of results for a search term is greater than zero, the following parts are shown in descending order:

1. Type of selected search (for instance all, images, videos) and information about retrieval (always shown),
2. Link for (search related) settings (always shown)
3. Button to configure search restrictions (always shown)
4. Button to configure search restrictions (always shown)
5. Featured snippet
6. Knowledge Graph
7. List of related search terms (always shown)
8. User's suggested location (always shown)
9. Links for help, feedback, privacy and usage terms (always shown)

In the following, all these points are examined more closely, including some examples.

1. Type of Selected Search

Searching for "Computer Science" leads to a result page containing all types of search. The user can now restrict the search to a specific predefined topic, as seen in Figure 4.1. The arrangement of topics is referenced to the search term. A search query "Flights to Graz" results in a different arrangement. In that case, the menu item "Flights" is right next to "All", which is always in first place.

Additionally, the user also gets information about retrieval. The number of approximated search results and the query time are shown. It is not stated how these two values are computed.

2. Link for (Search Related) Settings

Users might also configure the behaviour of the search engine. They can switch to Google instant mode. Google instant is an option to give users a kind of autocomplete for their search query.

4. Search Engine Result Page (SERP)

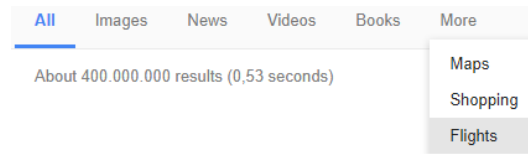


Figure 4.1.: Type of search and retrieval information for search term "Computer Science" on Google

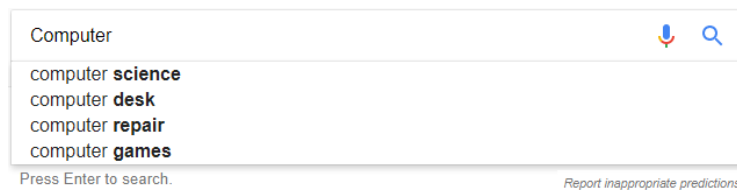


Figure 4.2.: Google Instant for search query "Computer"

Results appear automatically when a user types, see Figure 4.2 for an example. This technology was introduced in September 2008 and was launched widely in September 2010 (Hölzle, 2008).

Other options are to change the language, switch SafeSearch mode (to block inappropriate or explicit material), or perform an advanced search (see Figure 4.3). The advanced search allows techniques to restrict the search, for instance, exclude keywords, narrow region, or domain. See Figure 4.4 for an example. As it can be seen some of these restrictions are also available in the standard search box by using special commands such as quote marks or the minus operator.

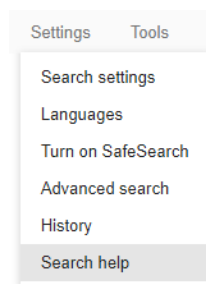


Figure 4.3.: Google's settings for the search

4. Search Engine Result Page (SERP)

Find pages with...		To do this in the search box.
all these words:	<input type="text" value="Computer Science"/>	Type the important words: tri-colour rat terrier
this exact word or phrase:	<input type="text"/>	Put exact words in quotes: "rat terrier"
any of these words:	<input type="text"/>	Type OR between all the words you want: miniature OR standard
none of these words:	<input type="text"/>	Put a minus sign just before words that you don't want: -rodent, -"Jack Russell"
numbers ranging from:	<input type="text"/> to <input type="text"/>	Put two full stops between the numbers and add a unit of measurement: 10..35 kg, £300..£500, 2010..2011

Then narrow your results by...

language:	<input type="text" value="any language"/>	Find pages in the language that you select.
region:	<input type="text" value="any region"/>	Find pages published in a particular region.
last update:	<input type="text" value="anytime"/>	Find pages updated within the time that you specify.
site or domain:	<input type="text"/>	Search one site (like wikipedia.org) or limit your results to a domain like .edu, .org or .gov
terms appearing:	<input type="text" value="anywhere in the page"/>	Search for terms in the whole page, page title or web address, or links to the page you're looking for.
SafeSearch:	<input type="text" value="Show most relevant results"/>	Tell SafeSearch whether to filter sexually explicit content.
file type:	<input type="text" value="any format"/>	Find pages in the format that you prefer.
usage rights:	<input type="text" value="not filtered by licence"/>	Find pages that you are free to use yourself.

Figure 4.4.: Google's settings for the search

4. Search Engine Result Page (SERP)

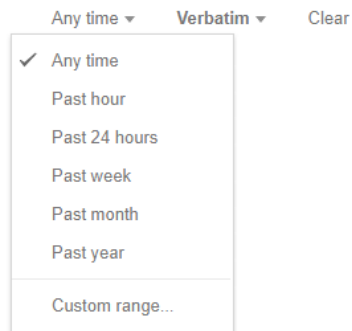


Figure 4.5.: Configure Search Restrictions on Google

3. Configure Search Restrictions

Users can also configure the date range of search results and choose a verbatim handling of the search request, see Figure 4.5. For other types of searches, similar options exist, as the size of images or the currency of news.

4. List of Results

Usually, ten links to results are shown on the page with the possibility to navigate to the next page. Depending on the search query, the user also gets latest tweets, top and fresh stories, images, links to social media profiles, related questions of other people ("People also ask" – in short PAA), placed within the ranking. Obviously, the result page is not a simple list containing a short snippet and a link to the corresponding site anymore. There is much more information presented to users. PAA is a rather new feature. It was introduced in 2015 and officially integrated into the English version of Google in February 2017 (Schwartz, 2017).

Observations show that there is no strict template for the list of results. While one search query results in a view containing PAA, four links to sites, images about the topic top-down (see Figure 4.6), a different search query results in a link, top stories, and some more links top-down (see Figure 4.7).

PAA is currently (July 2017) available in the English version of Google only. It is not permanently shown in the list of results as it highly relates to the

4. Search Engine Result Page (SERP)

People also ask

- What is it all about computer science? ▾
- What is Theory of Computing? ▾
- What is meant by theoretical computer science? ▾
- What is computer science and engineering? ▾

Feedback

Computer science - Wikipedia

https://en.wikipedia.org/wiki/Computer_science ▾

Computer science is the study of the theory, experimentation, and engineering that form the basis for the design and use of computers.

[Computer graphics](#) · [Human-computer interaction](#) · [Computation](#)



Computer Science Courses | edX

<https://www.edx.org/course/subject/computer-science> ▾

Take computer science courses online for free from top universities worldwide. Browse computer science MOOCs in a variety of disciplines and award rates.

Images for Computer Science



→ [More images for Computer Science](#)

[Report images](#)

Home | Department of Computer Science

<https://www.cs.cornell.edu/> ▾

The Computer Science Department at Cornell University and several high school-ranked teams of high school students from across New England compete in ...



Computer science | Computing | Khan Academy

<https://www.khanacademy.org/computing/computer-science> ▾




Learn with Khan's top college professors. Tom Gruber and David Salomon teach algorithms, computer science algorithms, including searching ...

Figure 4.6.: List of results for search query "Computer Science" on Google. Some results were pruned away.

4. Search Engine Result Page (SERP)


Climate change - Wikipedia
https://en.wikipedia.org/wiki/Climate_change ▼
Climate change is a change in the statistical distribution of weather patterns when that change lasts for an extended period of time. Climate change may refer to:
[Global warming](#) · [Climate change in Canada](#) · [Climate change denial](#) · [Weather](#)

Top stories

		
Want to fight climate change? Have fewer children	Anti-Trump climate change alliance may be unconstitutional	Climate change may ground more planes making flying more restrictive
The Guardian · 2 days ago	CNBC · 14 hours ago	Stuff.co.nz · 2 hours ago

[→ More for Climate Change](#)

NASA: Climate Change and Global Warming
<https://climate.nasa.gov/> ▼
The signs of the planet's global climate change and global warming. Current news and data streams about global warming and climate change from NASA.



When Will Climate Change Make the Earth Too Hot For Humans?
nymag.com/daily/intelligencer/2017/07/climate-change-earth-too-hot-for-humans.html
A new report says that the planet will become uninhabitable, and that it could be happening sooner than climate change could be seen to do.

Figure 4.7.: List of results for search query "Climate Change" on Google. Some results were prune away.

4. Search Engine Result Page (SERP)

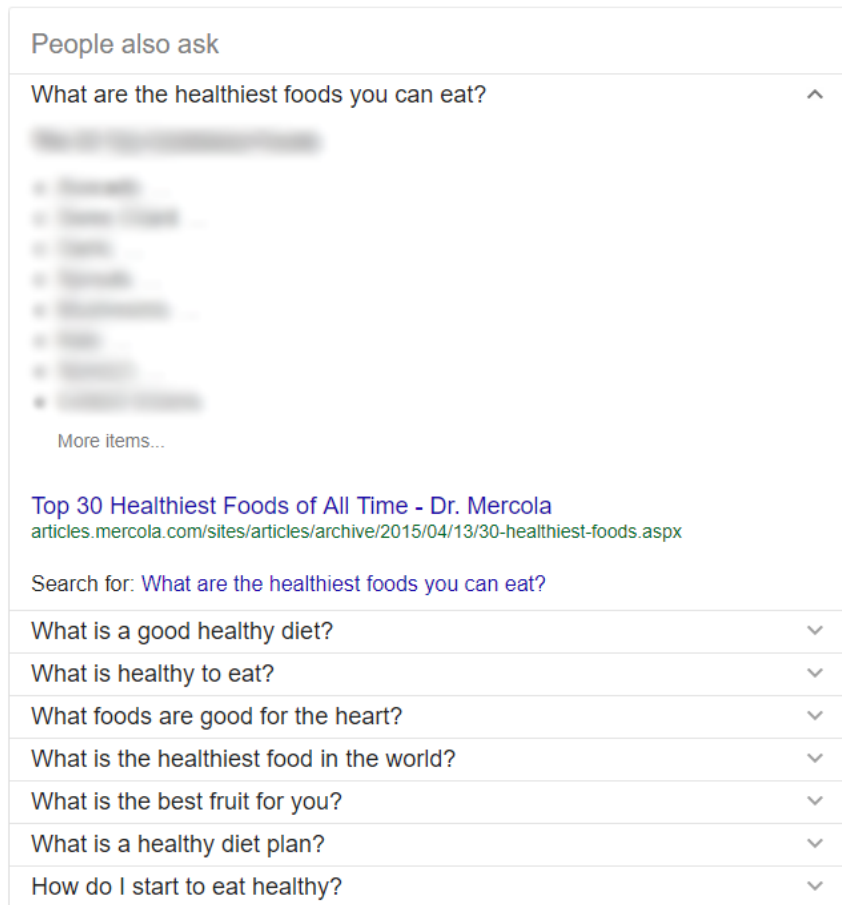


Figure 4.8.: PAA for "Healthy Food" on Google.

submitted search query. Observations show that it is only present if the search query could be interpreted as a question or at least the user might want to know something about the term(s) in the query more precisely. As there is no article about the functionality of PAA by Google itself and not much research done yet, it is impossible to analyse this feature in detail.

The search query "healthy food" results in the PAA shown in Figure 4.8, the first question was already clicked by the user. If the user opens a question, the PAA section is extended by further related questions. This option can be used indefinitely.

4. Search Engine Result Page (SERP)

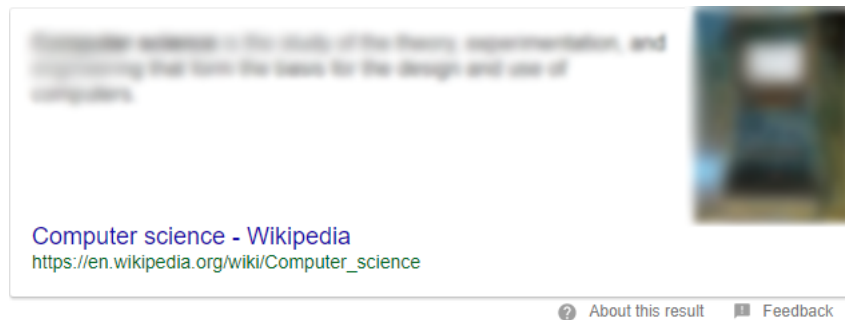


Figure 4.9.: Featured snippet for search query "Computer Science" on Google.

5. Snippets Featured

Google is able to give the user a description of the entered search term in certain cases. For instance, the search query "Computer Science" results not only in a list of ranked sites but also a featured snippet is shown at the top. It contains a description, the corresponding source, and a sample picture, see Figure 4.9. Not only definitions are supported. Users can also get instructions (Google, 2017m) and answers to questions (Google, 2017n).

According to (Google, 2017c), the data shown is extracted programmatically from websites offering answers to the user's search query. It is also stated that webmasters can prevent Google to take snippets from their site using a specific meta tag. Anyway, it is mentioned that it is not possible to mark a specific text as a featured snippet. Also, these data is just a search result with a special layout and not part of the Knowledge Graph which is analysed in further consequence.

6. Knowledge Graph

Imagine a search query for the name of a famous actor. A modern search engine will not only go through its index and list relevant web pages. Much more information is retrieved and presented. For instance, on Google, pictures, a short description, key facts, links to social media profiles and connections to other objects (in the case of an actor: persons and movies) are offered, besides the actual ranking, within a "Knowledge Panel" – hereinafter called Knowledge Graph. This data is collected from different sources on the Internet, for instance, Wikipedia, Freebase, and CIA World

4. Search Engine Result Page (SERP)

Factbook (Singhal, 2012) and is based on real search queries and therefore considered as interesting for the user by Google.

In the case of Arnold Schwarzenegger, the key facts contain information about his date of birth, his height, years of service, TV shows, and children. See Figure 4.10 for the Knowledge Graph of Arnold Schwarzenegger.

A search for "Vladimir Putin" reveals different information in the Knowledge Graph, for instance, his spouse, previous offices, presidential terms, and quotes. See Figure 4.11 for the Knowledge Graph of Vladimir Putin.

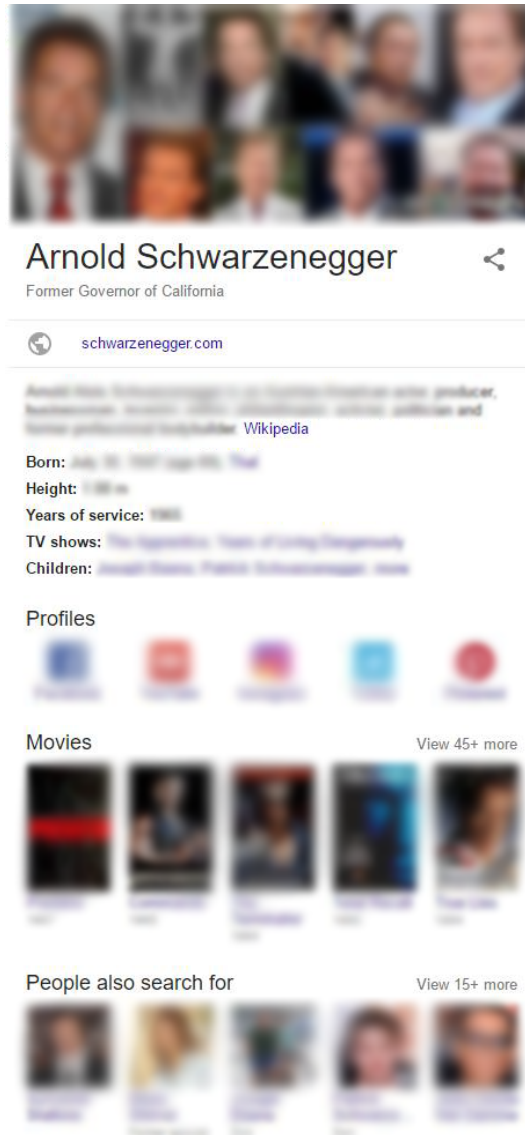
The Knowledge Graph for a big company looks similar. It contains its logo, a short description of the company, and information about the founder, headquarters, CEO, and stock price (if available). See Figure 4.12 for an example of the company "Twitter".

Local or smaller businesses are presented differently. The user gets information about the location, social status of the company, such as reviews, popular times, opening hours, and other reviews on the web. Most of this data is based on Google Reviews and Google+ (Graap, 2015). See Figure 4.13 for an example of the restaurant "Granola".

There are much more examples of objects that are in the Knowledge Graph, for instance, any kind of food. The Knowledge Graph consists of millions of objects (Google, 2015). Even Google+ friends are in the Knowledge Graph (available if the user is logged in). The data is taken from the corresponding Google+ profile. Google also offers the user to give a feedback on the data presented. Every single field can be reported. The interaction with the user does not end with the Knowledge Graph. If a user clicks on selected links within the panel, the process of retrieving and presenting information, without the need to visit an external site, goes on.

For instance, clicking on the entry "Predator" within the "movies"-section of Arnold Schwarzenegger's Knowledge Graph results in a new page with a Knowledge Graph containing information about the movie selected. Additionally, it presents a listing of other movies with Arnold Schwarzenegger at the top of the page. See Figure 4.14 for this example. The user has the opportunity to select the arrangement of presented movies, for instance, most popular first or newest first. After that, the user can continue browsing

4. Search Engine Result Page (SERP)



The image shows a Google Knowledge Graph for Arnold Schwarzenegger. At the top is a collage of photos. Below it, the name "Arnold Schwarzenegger" is displayed with a share icon and the subtitle "Former Governor of California". A link to "schwarzenegger.com" is provided. A brief Wikipedia-style description follows: "Arnold Alois Schwarzenegger is an Austrian-American actor, producer, businessman, investor, bodybuilder, politician, activist, politician and former professional bodybuilder." Below this, key facts are listed: "Born: July 30, 1947 (age 65), Austria", "Height: 1.93 m", "Years of service: 1983", "TV shows: The Appointments of大明王朝, The Expendables, The Terminator", and "Children: Joseph Patrick Schwarzenegger, more".

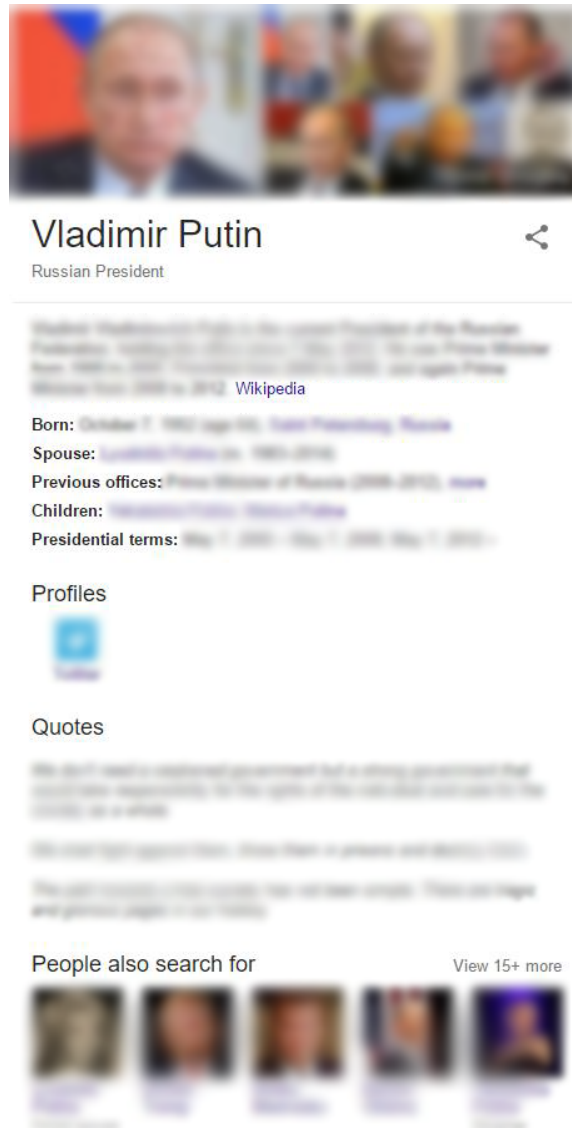
Profiles

Movies View 45+ more

People also search for View 15+ more

Figure 4.10.: Knowledge Graph for "Arnold Schwarzenegger" on Google.

4. Search Engine Result Page (SERP)



The image shows a Google Knowledge Graph for Vladimir Putin. At the top is a photograph of Putin. Below it, his name "Vladimir Putin" is displayed in large text, with "Russian President" underneath. To the right of the name is a share icon. Below this is a brief biographical summary: "Russian politician who is the current President of the Russian Federation. He has been President since 2012. He was Prime Minister from 1999 to 2007, Minister from 2007 to 2008, and again Prime Minister from 2008 to 2012." A "Wikipedia" link is provided. Further down, key facts are listed: "Born: October 7, 1952 (age 69), Leningrad, Russia"; "Spouse: Lyudmila Putina (m. 1996-2014)"; "Previous offices: Prime Minister of Russia (2008-2012), more"; "Children: Maria Vladimirovna Putina, Maria Putina"; "Presidential terms: May 7, 2012 – May 7, 2018, May 7, 2012 –". Below these are sections for "Profiles" (with a YouTube icon), "Quotes" (with several short quotes), and "People also search for" (with five small portrait icons and a "View 15+ more" link).

Figure 4.11.: Knowledge Graph for “Vladimir Putin” on Google.

4. Search Engine Result Page (SERP)

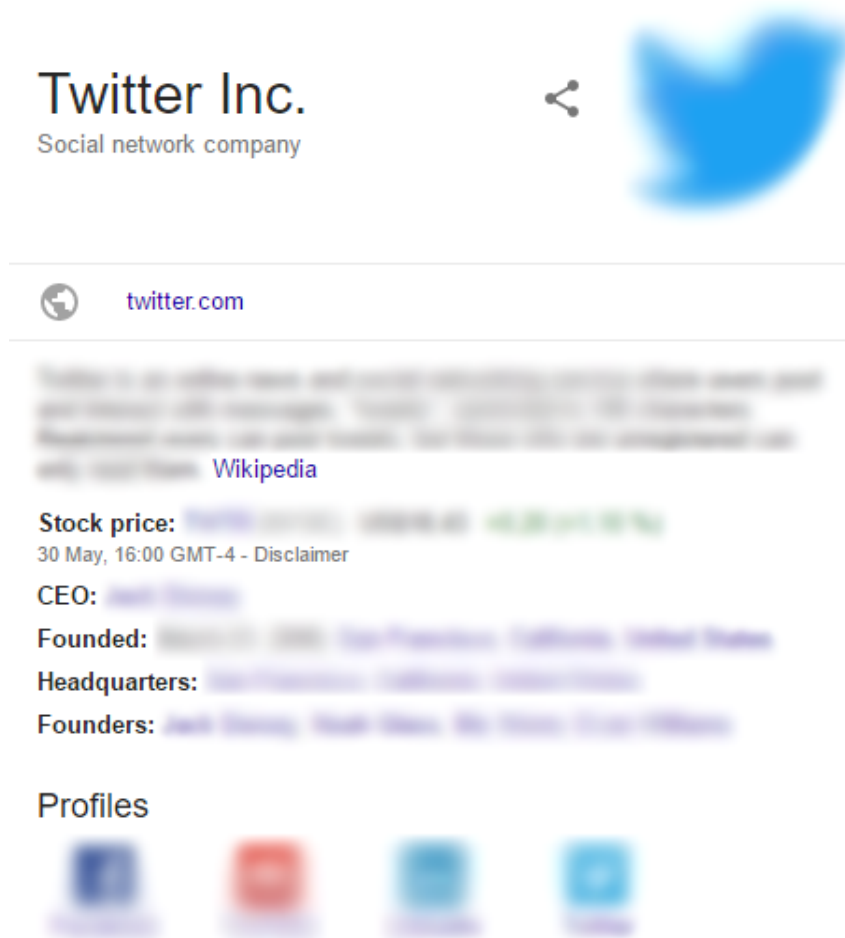


Figure 4.12.: Knowledge Graph for "Twitter".

4. Search Engine Result Page (SERP)

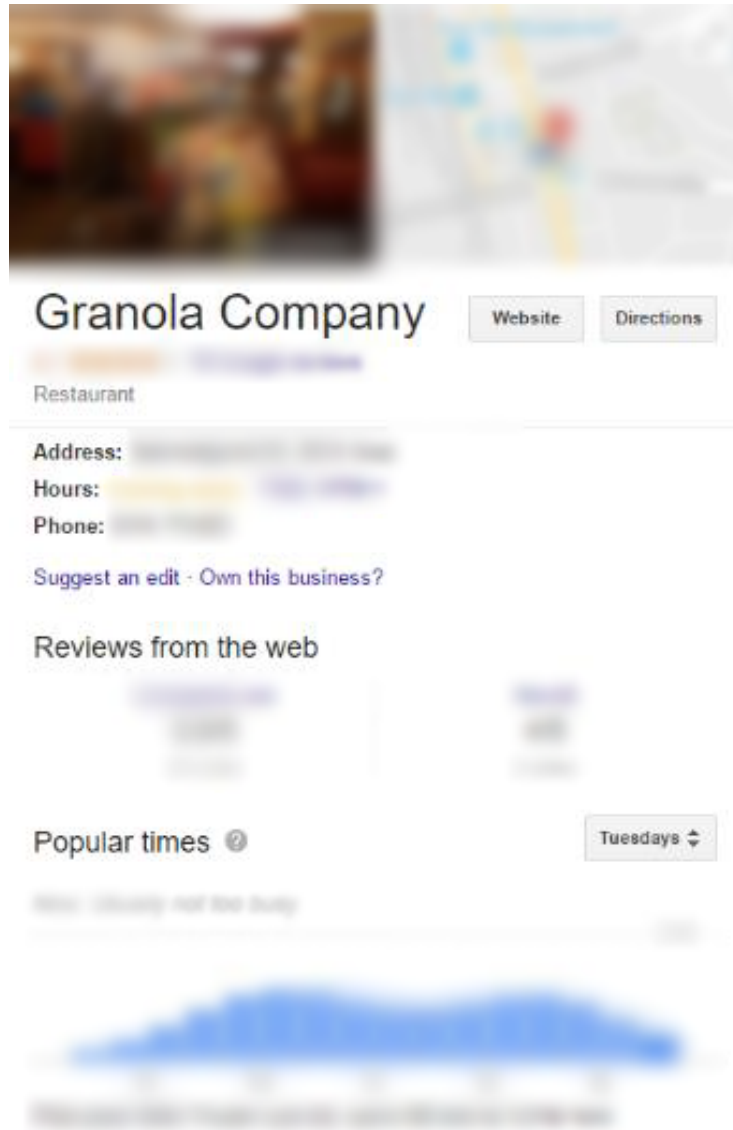


Figure 4.13.: Knowledge Graph for a local restaurant "Granola" in Graz

4. Search Engine Result Page (SERP)

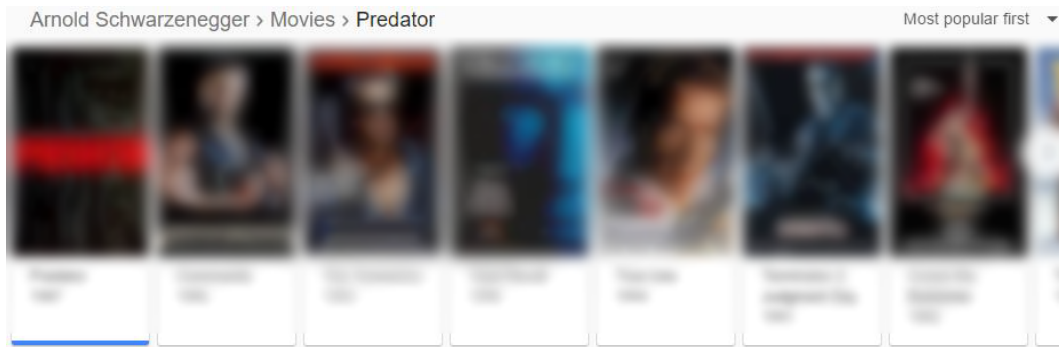


Figure 4.14.: List of movies with Arnold Schwarzenegger.

by selecting an entry within this list or within the Knowledge Graph without even leaving Google.

It is also possible that Google shows two different Knowledge Graphs at once. For instance, the search term "Arnold Schwarzenegger" results in a Knowledge Graph about himself as a person. Additionally, a Knowledge Graph for the Museum of Arnold Schwarzenegger is shown, if searchers are nearby according to their IP location (tested on June 18, 2017).

The technology to store complex structured and unstructured data and retrieve it understandable for humans is called knowledge base. All major search engines support this technology nowadays. Google's implementation is called "Knowledge Graph" and was introduced in 2012 (Singhal, 2012). The Knowledge Graph behaves similarly whether the user searches for a famous person, a big company, a restaurant, food, or even Google+ friends. Google also offers an application programming interface (API) to find entities within their Knowledge Graph (Google, 2015).

The knowledge base technology is a first and major step from a classical information source to a modern knowledge source. Therefore, the integration of Google's Knowledge Graph has many implications which are examined in section Implications, Drawbacks and Suggested Improvements.

7. Related Search Terms

At the bottom of the site, users get a list of related search terms presented. These search terms give users an easy way to specify their interests. To

4. Search Engine Result Page (SERP)

Searches related to Computer Science

computer science **course** computer science **pdf**
computer science **subjects** computer science **notes**
computer science **syllabus** computer science **requirements**
computer science **definition** computer science **careers**

Figure 4.15.: Related searches for "Computer Science" on Google.

● 8010, Graz - From your Internet address - Use precise location - Learn more

Figure 4.16.: Suggested location by Google.

show users a list of related searches was Google's first step to increase the interaction with the user. See Figure 4.15 for an example. Related search terms were introduced by Google in March 2009 (Allon and Wilder, 2009).

8. Suggested Location

To improve the quality of the search result, Google tries to detect the location of the searching user. This is done by checking Internet address, location history (if available), and recent locations a user searched for (Google Search Help, 2017). See Figure 4.16 for an example. Users can refresh their location by clicking "use precise location".

9. Links for Help, Feedback, Privacy, and Usage Terms

The last part of the result page at the bottom gives users possibilities to find help for using Google's search, to send feedback, to check privacy and terms, which a user has to agree within 24 hours using the search engine, see Figure 4.17.

Help Send feedback Privacy Terms

Figure 4.17.: Help, Feedback, Privacy, and Usage Terms on Google.

4.2. Implications, Drawbacks and Suggested Improvements

In this section, some implications and drawbacks, concerning the appearance of search results from the previous section, are pointed out. The focus lies on the issues of ambiguous search terms (What does the user mean or want to have?) and the data of the Knowledge Graph. For instance, mostly it is not stated where the data in the Knowledge Graph come from and how a user should interpret it. Questions arise are "Where does the data in the Knowledge Graph come from?" and "What does the data in the Knowledge Graph mean?". As part of this thesis, the issues addressed are analysed precisely and suggestions for improvements are given.

The Knowledge Graph is a rather new feature but it is widely used and therefore important for research. It is a significant sign of the empowerment of interaction between a search engine and its users. Although, there is still too little interaction in order to understand the user well. This is a massive point of criticism in this work.

A modern search engine gives users information they want to know, even before they are finished typing their search query. This is a hard challenge and requires a lot of learning from previous search queries by other users. As Google is a search engine used worldwide with about 3.300 billion search queries in 2016 (Internet Live Stats (2000-2012) and Statistic Brain Research Institute (2013-2015), 2017), the required data is available. Back in 2008, 2% of all search queries were used to improve Google Instant (former Google Suggest) to give users suggestions for their search query while they are typing (Hölzle, 2008). By submitting search queries, users give back much data to the search engine that can be analysed. This data is used to improve suggestions for users. Therefore, using a search engine is a constant interplay between the search engine and its users.

4.2.1. Ambiguous Search Terms

Once a search query is submitted, the search engine sifts through its index to find relevant content. Thus, there is a huge issue concerning ambiguo-

4. Search Engine Result Page (SERP)

ous search terms. The search engine has to compute (or guess), which interpretation or aspect the user might look for.

Consider the search term "Galaxy". According to Wikipedia's site for "Galaxy (disambiguation)"², it has more than 50 different meanings, beginning from an astronomical system, up to broadcast media, print media, entertainment, sports teams, transportation, and many more usages, all containing or related to "Galaxy". Not considered are different spellings, such as "Galaxie", "Galaxia", and others that are also ambiguous.

Imagine, a user searches for "Galaxy". A fundamental question is how a search engine should treat such a search query. A look at Google is taken, see Figure 4.18 for the result page of "Galaxy", some results were pruned away. For the complete result page, see (Google, 2017d).


In the following, the result page is analysed. About 1.150 billion search results were found. The first entry in the list links to Wikipedia's site for Galaxy (astronomical system). Below, links to the Twitter account of LA Galaxy (an American professional soccer team) and some latest tweets are given. An image gallery with pictures of (astronomical) galaxies is following. The next entry in the list is a link to Samsung Galaxy products. In the following top stories section, some latest news containing Samsung's Galaxy series are presented. Some more links are below, pointing to Samsung's website, the Galaxy community hub (a research platform), the web portal of UT Dallas (The University of Texas at Dallas), and a different link to Galaxy's community hub (pruned away in Figure 4.18). At the bottom of the result page, related searches are listed, giving a hint that the search term might be ambiguous and there are many aspects to be considered.

Looking at the result page reveals that Google mixes many different topics, but they are all related to the original search term "Galaxy". This might be an indicator that the search engine does not know the user's intention. To compensate this lack of information, the search engine presents the probably most related topic (astronomic system), but also some lately popular topics, such as the soccer club or products from Samsung. Submitting the same search query a few days later (or even hours) will most likely lead to a different result page.

²[https://en.wikipedia.org/wiki/Galaxy_\(disambiguation\)](https://en.wikipedia.org/wiki/Galaxy_(disambiguation))


4. Search Engine Result Page (SERP)

Galaxy - Wikipedia
<https://en.wikipedia.org/wiki/Galaxy> ▼
Spiral galaxy · Disambiguation · Elliptical galaxy · Compact star

LA Galaxy (@LAGalaxy) · Twitter
<https://twitter.com/LAGalaxy> 

6 hours ago · [Twitter](#) 7 hours ago · [Twitter](#) 8 hours ago · [Twitter](#)


Images for galaxy



→ [More images for galaxy](#) [Report images](#)

Samsung Galaxy Products: Phones, Smartwatches & VR Headsets ...
www.samsung.com > Home > Mobile ▼


Top stories



Samsung's next Galaxy Note will be announced on August 23rd
[The Verge](#) · 12 hours ago

Samsung to unveil Galaxy Note 8 at Unpacked event on Aug. 23
[CNET](#) · 12 hours ago

Samsung will reveal the Galaxy Note 8 on August 23rd
[Engadget](#) · 12 hours ago



Searches related to galaxy

- galaxy **samsung** galaxy **movie**
- galaxy **definition** galaxy **tumblr**
- galaxy **phone** galaxy **hd**
- galaxy **space** galaxy **game**

Figure 4.18.: Result page for "Galaxy" on Google. Some results were prune away.

4. Search Engine Result Page (SERP)

There are many more such examples, as there are many ambiguous words in English. Some of them indicate the same behaviour as "Galaxy", for instance "Domino" (Google, 2017b), "Sugar" (Google, 2017k), and "Time" (Google, 2017l). Actually, it is hard to find a search term that is not ambiguous.

Further on, the same applies for search terms whose interpretation has already been restricted. Consider the search term "electricity" in terms of energy. There are different ways to view at this topic. Users might be interested in a short overview of different aspects, some particular information about the technology to produce and store it, some technologies to distribute it, or its relation to the environment. Anyway, there are many more aspects. In this thesis, possible ways to address that issue are pointed out. These require more interaction with users to understand their intention.

For instance, after users submit a search query, they could be asked to specify their area of interest. The search engine could ask, "In which of the following area(s) are you interested in?" and give some suggested related topics. At least, users must be informed about the issue that they entered an ambiguous search term and that there are many aspects to consider. That is done partly on Google's SERP by showing related searches at the bottom of the site.

Another solution is to give users the possibility to specify the area even before they enter a search query. After this specification, all results are restricted to the selected area(s) only. Austria-Forum³ uses this approach, among others.

A third possible way to handle that problem is to give users, after the search query is submitted, results mixed up from different areas (similar to Google's current SERP), but labelled with the associated topic. By clicking on a label, the search can be restricted to that area.

To sum up, if users enter an ambiguous search term they must be informed about the issue that there are many ways to interpret the search query. There should be more interaction between the search engine and its users to get more information about the user's intention. With the intention available it is easier to derive which kind of information a user wants.

³<https://www.austria-forum.org>

4. Search Engine Result Page (SERP)

4.2.2. Knowledge Graph

As mentioned previously, Knowledge Graph's information is taken from free data platforms (such as Wikipedia and Freebase) and is based on real search queries (Singhal, 2012). This has some implications for all parties involved, the search engine, its users, and the websites providing the information.

1. Number of Clicks on Information-Providing Websites

Firstly, it affects the number of clicks on a website, providing information. For instance, Wikipedia lost more than 550 million search visits within 6 months in 2015 (Hinkis, 2015), most likely related to knowledge base search engines.

Thus, these information-providing sites lose viewers, therefore attention, and in further consequence money due to a lack of sold products, shown advertisements, or other relevant benefits. In addition, Google itself might lose important feedback from users, for instance, the CTR (see Section List of Ranking Factors on Google).

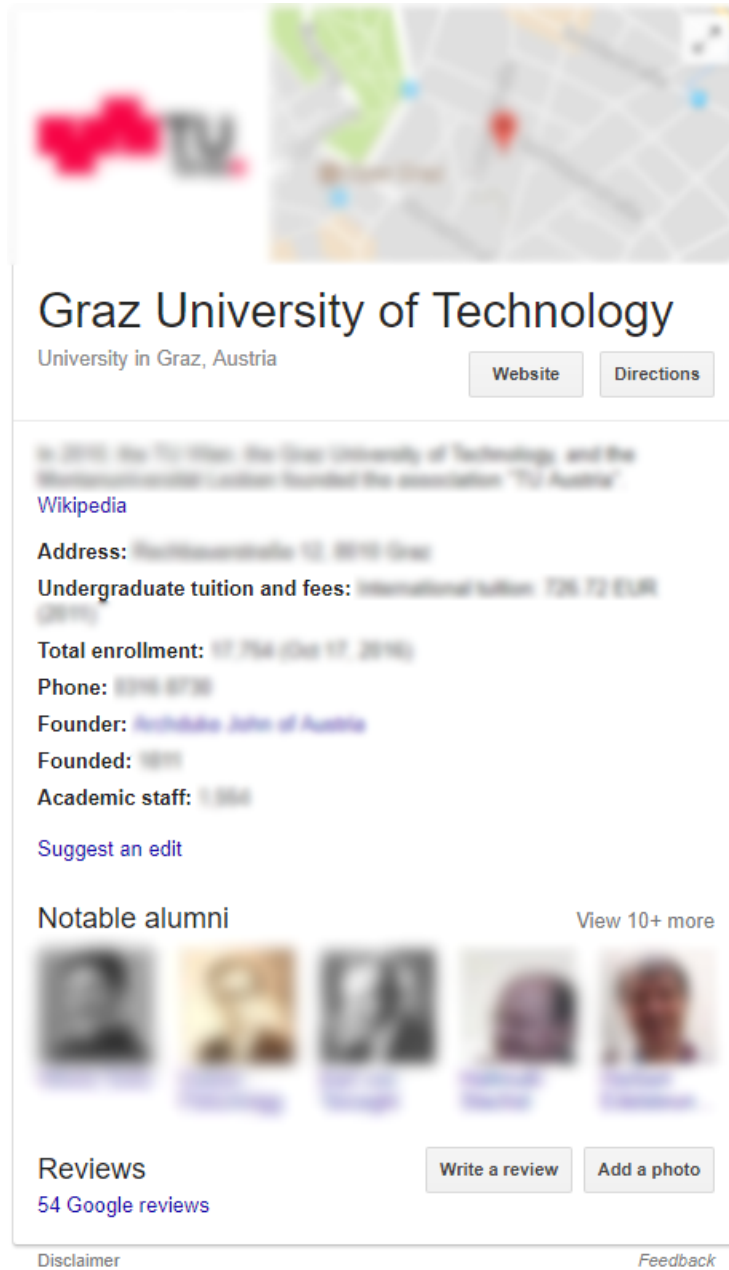
It is beyond debate that, on the one hand, websites have a huge benefit from search engines such as Google and Bing. On the other hand, Google would not exist without web pages supplying content that users want to see. Therefore, it should always be a trade-off between taking content from websites to display it in the SERP and linking to the original website to generate clicks for them.

2. Source of Data Retrieved

Secondly, it is not always clear, where the data retrieved was taken from. Imagining the search query "Graz University of Technology". It leads to the Knowledge Graph shown in Figure 4.19.

For the image (logo of TU Graz), the map (geographical position of TU Graz), the title, some key facts, and notable alumni no source is stated. For users, it is not clear where this information comes from. Only for the short description text, a link to Wikipedia is shown. Similar examples are the Knowledge Graph for Arnold Schwarzenegger (see Figure 4.10),

4. Search Engine Result Page (SERP)



The image shows a Google Knowledge Graph for the Graz University of Technology. At the top, there is a banner with the university's logo (a red 'TU' with a stylized 'G' and 'U') and a map of the university's location in Graz, Austria. Below the banner, the title "Graz University of Technology" is displayed, followed by the subtitle "University in Graz, Austria". To the right of the subtitle are two buttons: "Website" and "Directions".

Below this, there is a snippet of text: "In 2010, the TU Wien, the Graz University of Technology, and the Montanuniversität Leoben founded the association 'TU Austria'". A link to the Wikipedia page is provided. The following information is listed:

- Address:** Inffeldgasse 17, 8010 Graz
- Undergraduate tuition and fees:** International tuition: 726.72 EUR (2019)
- Total enrollment:** 17,764 (as of 17, 2016)
- Phone:** 0316 8730
- Founder:** Friedrich-Johann von Auersperg
- Founded:** 1815
- Academic staff:** 1,364

A link to "Suggest an edit" is also present. Below this is a section for "Notable alumni" with five profile pictures and a link to "View 10+ more".

At the bottom, there is a "Reviews" section with a link to "54 Google reviews" and two buttons: "Write a review" and "Add a photo".

At the very bottom, there are links for "Disclaimer" and "Feedback".

Figure 4.19.: Knowledge Graph for "Graz University of Technology" on Google.

4. Search Engine Result Page (SERP)

Vladimir Putin (see Figure 4.11), Twitter (see Figure 4.12), and Granola (see Figure 4.13).

To stress this fact: Users get almost no information whether the data in the Knowledge Graph is reliable or not. Even if the sources are considered as reliable by Google does not mean users see them as trustworthy. For instance, in most information-providing sources in the area of geography, facts are usually taken from one single source and are not cross-checked against other providers (Mehmood, Kualthuramaiyer, and Maurer, 2016). The Knowledge Graph provides a link to a disclaimer (see Figure 4.19 at the bottom). Currently, there is only a disclaimer for data retrieved from several stock exchanges, investment fund, indexes, bonds, and currencies (tested on July 16, 2017) but not for the data in the Knowledge Graph.

To address this issue, a search engine should always state the source of the information displayed. If data is available on different platforms, a search engine should provide different sources. Thus, users can decide on their own if or which source they trust or not.

3. Interpretation of Data Retrieved

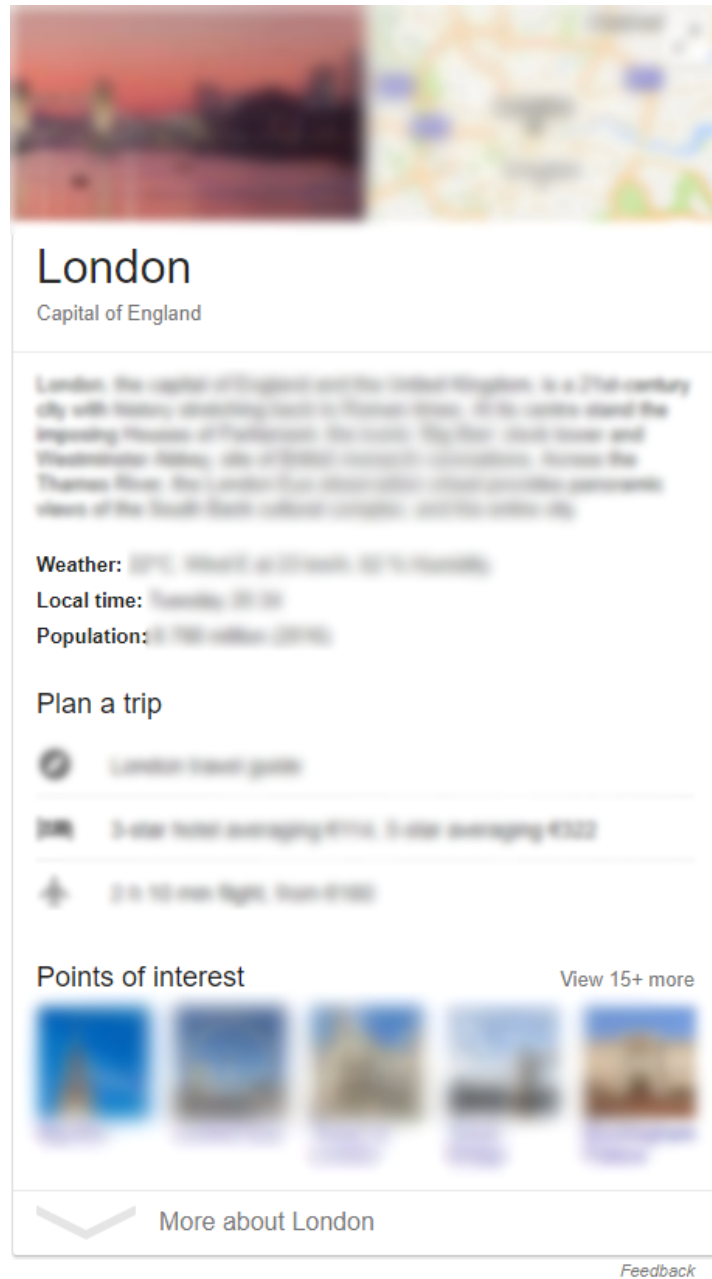
Thirdly, it is not clear how this data should be interpreted. Imagine a search query "London". On Google, users get some information about the city, an image, a map, a short description, some facts, some points of interest, and others. See Figure 4.20 for the Knowledge Graph of "London".

For users, this seems straightforward. The way the summary of the city is presented might be understandable. However, consider the fact that London had 8.788 million inhabitants in 2016 (see Figure 4.20).

Firstly, there is no source stated for this information (see Source of data retrieved). Secondly, there is no information given on which area this number applies to. It could be the city population, the metro area population, or a combination of both. In comparison, Wolfram-Alpha distinguishes between city population and metro area population, see Figure 4.21 (retrieved on July 21, 2017).

Another example is a search for "List of Austrian Nobel Prize winners" on Google. The response is a list of persons who received the Nobel Prize, displayed at the top of the page, see Figure 4.22 for a snippet. The full result

4. Search Engine Result Page (SERP)



The image shows a Google Knowledge Graph for the city of London. At the top, there are two images: a sunset over the London skyline and a map of London. Below the images, the word "London" is displayed in a large font, with "Capital of England" underneath. A short paragraph of text describes London as the capital of England and the United Kingdom, a 21st-century city with history, mentioning landmarks like Big Ben and the Houses of Parliament. Below this, there are three lines of information: "Weather: 21°C, Most at 23°C, 19°C tonight", "Local time: Tuesday, 10:10", and "Population: 9,100,000 (2016)". A section titled "Plan a trip" includes three items: "London travel guide", "3-star hotel averaging \$174, 5-star averaging \$322", and "2 to 10 one-way flights from \$180". A "Points of interest" section shows five small image thumbnails and a "View 15+ more" link. At the bottom, there is a "More about London" link with a downward arrow icon and a "Feedback" link.

Figure 4.20.: Knowledge Graph for "London" on Google.

4. Search Engine Result Page (SERP)

city population	8.674 million people (country rank: 1 st) (2015 estimate)
metro area population	12.58 million people (London metro area) (2007 estimate)

Figure 4.21.: Stated population for "London" on Wolfram-Alpha.

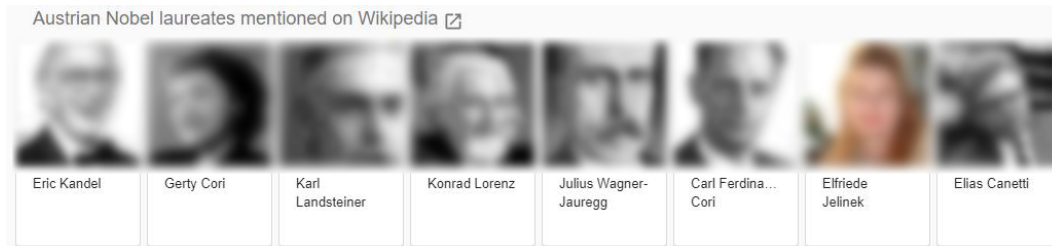


Figure 4.22.: Snippet of the list of Austrian Nobel laureates on Google.

page is provided in (Google, 2017h). The source for that list is stated as Wikipedia, a link is provided, see Figure 4.22.

Thus, users get a simple list of persons who received the Nobel Prize. What they might not know is that this list is rather arbitrary. Firstly, it is not explicitly stated what "Austrian Nobel laureates" mean. Either this could be people with Austrian citizenship or nationality, or a receiving of the prize during their work in Austria, or winners born in Austria if the location is or was part of Austria in its history. Indeed, examining the list more closely by following the links to Wikipedia reveals their relationship to Austria, which is rather different for each person.

For instance, while Erwin Schrödinger was Austrian by nationality, Max Perutz was Austrian-born but British by nationality. Elias Canetti's relationship to Austria is rather tenuous. He was born in the Principality of Bulgaria, received the British citizenship in 1952. From 1912 – 1916 and 1924 – 1938 he lived and studied in Vienna. Despite this loose connection to Austria, he is mentioned in that list.

Thus, sometimes the information presented is inaccurate and users might be confused. One could even go so far as to say that the information displayed is wrong.

4. Search Engine Result Page (SERP)

year	recipient	field	country of birth
2013	Martin Karplus	chemistry	Austria
2005	International Atomic Energy Agency (IAEA)	peace	
2004	Elfriede Jelinek	literature	Austria
2000	Eric Kandel	medicine	Austria
1998	Walter Kohn	chemistry	Austria
1974	F. A. Hayek	economics	Austria
1973	Karl von Frisch	medicine	Austria
1973	Konrad Lorenz	medicine	Austria
1962	Max F. Perutz	chemistry	Austria
1945	Wolfgang Pauli	physics	Austria
1938	Richard Kuhn	chemistry	Austria
1936	Otto Loewi	medicine	Germany
1936	Victor Francis Hess	physics	Austria
1933	Erwin Schrödinger	physics	Austria
1930	Karl Landsteiner	medicine	Austria

Figure 4.23.: List of Austrian Nobel Prize winners on Wolfram-Alpha.

Alternatively, Figure 4.23 shows the same search query on Wolfram-Alpha (retrieved on July 22, 2017). It can be seen easily that this result differs from Google's result. Nevertheless, even on Wolfram-Alpha a concrete definition for "Austrian Nobel Prize winners" is missing. Again, it is not stated what "country of birth" means as, keeping in mind that the border of Austria has changed over time.

Contrary, on Global-Geography⁴, a definition is given for Nobel Prize winners (Global-Geography, 2017):

"We are displaying Nobel Prize Winners using their country of

⁴<https://www.global-geography.org>

4. Search Engine Result Page (SERP)

birth, if the location they were born is still today within that country.”

Anyway, there are more examples. For instance, in the Knowledge Graph of the German movie “Suck Me Shakespeer 2” the box office is stated in US-Dollars, although this movie was released in German-speaking countries only. There is no information on how the revenues were calculated. Further on, it is not stated which exchange rate was taken to convert the earnings into Dollars. See Figure 4.24 for the Knowledge Graph of “Suck Me Shakespeer 2”. Similar inconveniences occur for other movies, for instance, “Men in Black” (Google, 2017i).

In this work, there are different possible solutions suggested. The main claim is: The search engine has to give the user a clear and precise information about the data displayed. This could be either a concrete definition or at least a link to the website containing both, data and definition.

If data is available in different definitions, users should be informed about this fact. The search engine should try to display all of them (or at least provide a link to the source) or state the lack of availability. Global-Geography is doing that by showing different sources and all available definitions for them, see Figure 4.25.

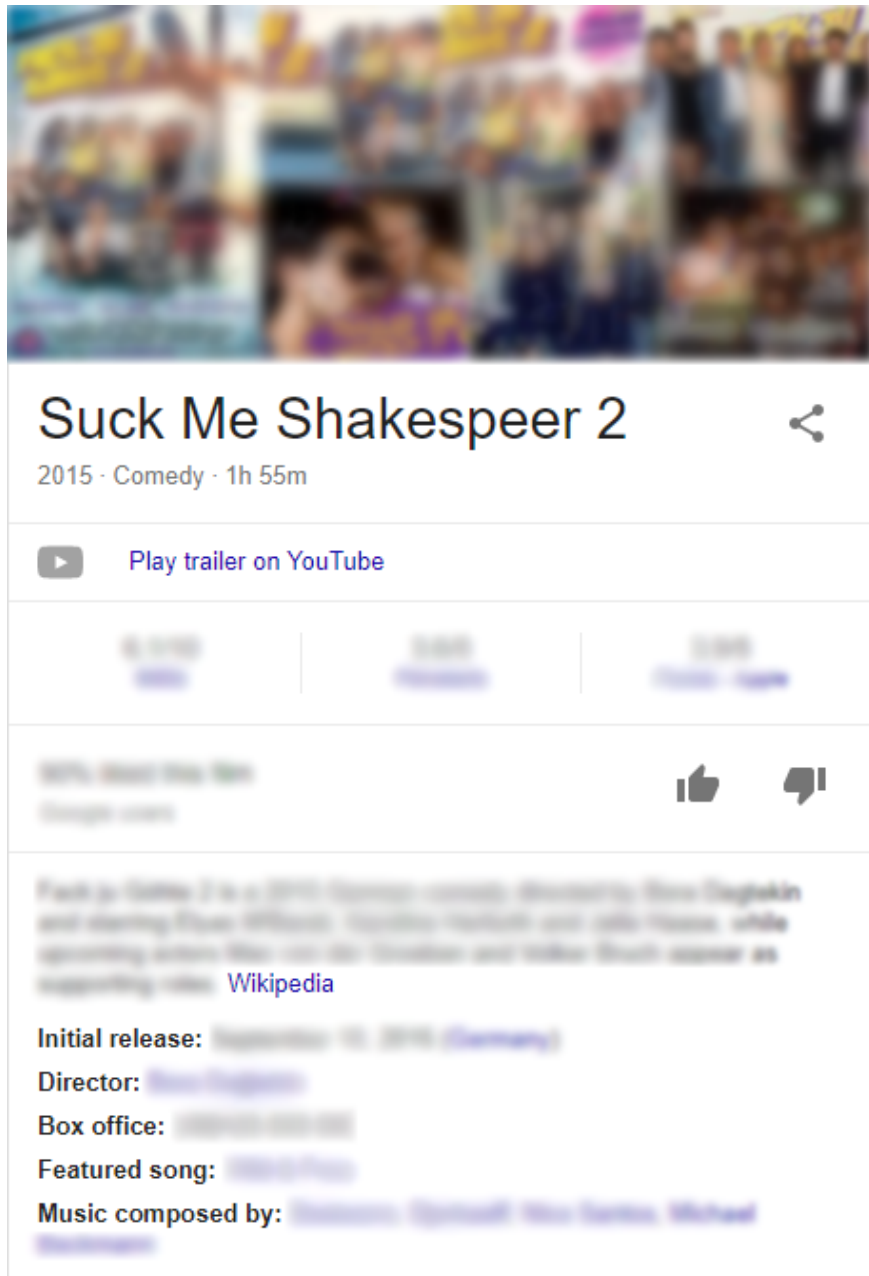
In short, currently, it is not common to give users a definition of results presented - a big critical point examined in this thesis. While in some cases this is negligible, there exist many examples where a definition is indispensable in order to make sense of the information presented.

4. Correctness of Data Retrieved


Fourthly, it is not clear whether the data retrieved (assuming a definition is provided) is correct or not. This is a rather trivial argument as any information anywhere could possibly be wrong. Anyway, users might trust information within Google’s Knowledge Graph more than other sources with the argument “If it is on Google, it must be true.”

Indeed, there are statistics revealing that search engines tend to have the highest ratings for trust (Bush, 2016). As stated, about 63% trust search engines and about 58% trust traditional media for news and information.


4. Search Engine Result Page (SERP)






The image shows a Google Knowledge Graph for the movie "Suck Me Shakespeer 2". At the top is a blurred movie poster. Below it, the title "Suck Me Shakespeer 2" is displayed in large black font, with a share icon to its right. Underneath the title, it says "2015 · Comedy · 1h 55m". A "Play trailer on YouTube" button is visible. Below that are three columns of statistics, which are mostly blurred. There are also icons for "Like" and "Dislike". A short synopsis follows, with a "Wikipedia" link. At the bottom, there are several key-value pairs: "Initial release:" followed by a date and location, "Director:" followed by a name, "Box office:" followed by a value, "Featured song:" followed by a name, and "Music composed by:" followed by a list of names.

Suck Me Shakespeer 2 

2015 · Comedy · 1h 55m

 [Play trailer on YouTube](#)



Suck Me Shakespeer 2 is a 2015 German comedy starring [Oliver Diekmann](#) and starring [Oliver Diekmann](#), [Oliver Diekmann](#) and [Oliver Diekmann](#), while appearing [Oliver Diekmann](#) and [Oliver Diekmann](#) as supporting roles. [Wikipedia](#)

Initial release: [September 11, 2015 \(Germany\)](#)

Director: [Oliver Diekmann](#)

Box office: [\\$1,111,111](#)

Featured song: [Oliver Diekmann](#)

Music composed by: [Oliver Diekmann, Oliver Diekmann, Michael Diekmann](#)

Figure 4.24.: Knowledge Graph for "Suck Me Shakespeer 2" on Google.

4. Search Engine Result Page (SERP)

Area	<i>total:</i> 243,610 sq km <i>land:</i> 241,930 sq km <i>water:</i> 1,680 sq km <i>note:</i> includes Rockall and Shetland Islands <i>We are unable to verify those figures from Factbook, since the 6 main sources we used for checking give different figures as follows:</i> <i>Factbook</i> •: 243610 <i>DBpedia</i> •: 243610 <i>Geoname</i> •: 244820 <i>Infoplease</i> •: 244820 <i>Britannica</i> •: 243073 <i>Wolfram</i> •: 243610 <i>Attempted Explanation:</i> Please help us to try to explain the discrepancies by sending us helpful information to office@global-geography.org •
-------------	---

Figure 4.25.: Area stated for "United Kingdom" on Global-Geography.org.

The third-most trusted source is online media (53%) followed by owned media (46%) and social media (44%).

4.2.3. People Also Ask (PAA)

PAA is - like the Knowledge Graph - a huge step to more interaction between a search engine and its users. As users get more related questions to the search term continuously by opening questions, they can explore a topic in a deep way without even leaving Google's site. As there is no official explanation for the technology used, it is hard to determine how relevant the offered questions are.

5. The Future: Search Assistance

Keeping the issues and drawbacks of Google's current SERP in mind, in this chapter, a new approach for a search engine and the main contribution of this thesis is proposed¹. Its purpose is to get rid of the issues mentioned above. In addition, the convenience of users should be enhanced by selecting and extracting whole parts from other websites they might want to see automatically. To find out which sources are "useful" for a specific topic, an improvement of interaction between the search engine and its users is required.

A new approach for a search engine is that so-called reports containing information about the search query submitted are generated. The content of these reports is fetched from different sources on the Internet. All sources are stated in the report to give users the possibility for further reading. Additionally, users can rate different aspects of sources. These ratings represent the quality of a source and are obtained by users. The information on the ratings is stored in a database on an external server to avoid manipulation. There are many aspects to state if a source is "good" or not. For instance, the server should have a high reliability (high quality of content, availability, and low response time), content should be trustworthy, impartial, relevant, readable, well-cited, and up-to-date. Another point to mention is how to determine which users are allowed to give quality-feedback (special/registered users or all users). These issues are discussed more precisely in section Discussion.

As an extension, reports contain different categories or views at a topic. This is included in order to get rid of the problem of ambiguous search queries mentioned in the previous chapter. For instance, a report about the search

¹Many of the ideas presented have emerged from discussions with the supervisor of this thesis, Hermann Maurer.

5. The Future: Search Assistance

query "London" contains categories like "Images", "General Information", "Points of Interest", and "History". A huge challenge is to fill the categories accordingly and not display content twice within a report. This requires having a set of metadata associated with a source. These metadata describe the content. For instance, that a text is about a city's history. Thus, this needs an automatic way to obtain the meaning of a text and requires language analysis tools to be used in the future but not yet in this thesis.

Another extension is to select a master source which provides most of the content of a report. The selected master source has to fulfil the requirements of a "good" source as effectively as possible, a hard challenge to obtain. Additionally, some minor sources fill-up missing information. Again, this requires clever language analysis tools.

In the following, the main idea is explained in greater detail. Afterwards, the idea of a search assistant is applied to a small domain. The domain chosen is "some major cities of German and English speaking countries". After a motivation for the domain chosen, an example of a conventional search on that domain is illustrated. Subsequently, the new approach is shown in the domain chosen. Finally, the idea of a search assistant is discussed. The main advantages of this approach, as a simplification for users when searching or well-citing of information, are pointed out. Also, drawbacks and challenges, as the questions how to get rid of redundancies or how to obtain sources of high quality, are analysed. In the next chapter, an implementation of a very early version of a search assistant (named *Sistant* – a short version of Search assISTANT) is introduced and discussed.

5.1. Main Idea of a Search Assistant

In this section, a new approach for a search engine is proposed. By submitting a search query to Google, users get a SERP in a structure explained in the previous chapter. The first step for an improvement of interaction between a search engine and its users is to inform users if they entered an ambiguous search term. Users are asked to specify their area of interest. Then, the search engine suggests some related topics. Alternatively, users specify their area of interest even before entering a search query.

5. The Future: Search Assistance

The main parts of a SERP are the list of results (links to other websites), a short snippet (for instance, a definition for a search term), and the Knowledge Graph (containing some key information about the search term). Users have the possibility to "consume" the presented information of the featured snippet or the Knowledge Graph, or they can visit the linked sites to get a broader overview of the topic. Although Google tries to rank "useful" sites at the very top, it is an annoying task for users to visit external sites. The external sites might have an unusual or unintuitive design for the user. Furthermore, they might contain tons of advertisements covering actual useful information, or they are not available at the moment, probably resulting in a long but not successful loading of the page.

As the technology gets more sophisticated continuously, the task of visiting external sites should be automated by the search engine. In the process of crawling and indexing, these tasks are already performed by a search engine. The information of millions of websites gets sifted to find content. The next step for improvement is an extraction of some information automatically and present it to the user so that they do not have to visit the external sites on their own. Optimally, the search engine has the ability to extract content from "relevant" websites automatically. In the beginning, lists created manually have to be used. These lists contain topics and corresponding URLs of servers providing "useful" information. Then, the search engine has to call these URLs and fetch, parse, and present the content of them to the user. A similar approach was already done for a server providing geographical data. The generation of non-trivial stories for countries by extracting automatically "interesting" facts, combined with manual cleaning-up, was introduced, tested, and deduced as actually working (Mehmood and Maurer, 2017). Obviously, that is not possible for every search query. A starting point is to offer additional information about search terms, as it is already done with the featured snippet and the Knowledge Graph. For instance, it is a rather easy task for a search engine to give users information about a city, a famous person, or an object of general interest. The search engine could extract whole parts from different external sites, concatenate them to one document, a so-called report, and present it to users. Of course, each extracted snippet should be linked to the original content to firstly, state the source and secondly, give users the availability for further reading.

A question immediately arising is how the content should be extracted.

5. The Future: Search Assistance

For instance, Wikipedia offers an API² to retrieve data from its articles. Another approach is to use HTML-based communication. The search engine establishes parallel connections to different servers by calling deposited URLs. In further consequence, these URLs are obtained automatically. Then, the content of these websites is fetched, parsed, and finally presented to the user. A major drawback of this approach is that parsing of a server-response is very individual as HTML content has to be analysed. Another challenge is a possible format-change on the providing servers and the question is how the search engine should deal with it. One more drawback of this approach is that it is very time-consuming to call URLs and fetch data just in time. An alternative approach is to store data which is known to be open-source and therefore retrieve it much faster from a local storage. These data have to be refreshed in certain time intervals to be constantly up-to-date. An alternative is to ask major content providers for permission to use their data. Another issue concerns the language of a source. Wikipedia is available in many languages but local providers may not support multiple languages. Anyway, the aim is to translate information to the desired language automatically by using tools like Google Translate³ and others. All these aspects are discussed more precisely in section Discussion.

A further extension is to split up the content of these reports into meaningful categories. This is especially useful for ambiguous search terms. For instance, users searching for a city might be interested in pictures, video clips, key facts, general information, points of interest, or history. By default, the search engine does not know exactly what users want to see in particular. An automatically generated report consists of all mentioned categories. Each category has some sources for the content deposited. Of course, a user has the possibility to switch off categories of no interest. The list of sources can be deposited on the server or in further versions obtained by the search engine automatically as it ranks different sites accordingly already. As mentioned before, to fill categories with associated content is a hard challenge. For that, it needs to have metadata that describes corresponding content. Therefore, it needs software to automatically determine the meaning of text. This is only possible with sophisticated language analysis tools in the future but not yet in this thesis.

²https://www.mediawiki.org/wiki/API:Main_page

³<https://translate.google.com>

5. The Future: Search Assistance

Another immediate influencing factor for ranking is the feedback from users. A user might rank different aspects of a source either by choosing "positive"/"negative", a 5-star-rating system⁴, or by using a Likert scale⁵. By collecting votes from many users the best content would come out on top. This can be done on an extra page which is available for chosen users only. Chosen users can be registered or trusted users by the search engine. As mentioned above, it is hard to obtain the quality of a source. There are many aspects to consider, as the reliability of the server, up-to-dateness and impartiality of the content, and many others. Nevertheless, the rating for different sources should be visible for all users. Furthermore, users also have the possibility to give feedback about the report generated directly to the search engine, for instance, a request to add a potentially good source. This approach is discussed in greater detail in section Discussion.

Furthermore, a huge challenge is to ensure that no content is displayed more than once. Information consolidation, this means taking information from various sources with as little redundancy and as little information loss is a major challenge. In general, no satisfying approach has been discovered yet (Maurer and Mehmood, 2015).

A possible modification to be examined is a selection of a master source which provides most of the information presented in the report. Besides that, the content of some minor sources is collected to fill-up missing information on a topic/category. Again, determine the meaning of text and filtering for redundant content are hard task, to be discussed in greater detail in section Discussion.

5.2. Search Assistant Applied on a Small Domain

In this section, the idea of a search assistant is applied to a small domain. Its main focus is to give users a broad knowledge about a topic by automatically generating a report so that they do not have to visit several websites on

⁴[https://en.wikipedia.org/wiki/Star_\(classification\)](https://en.wikipedia.org/wiki/Star_(classification))

⁵https://en.wikipedia.org/wiki/Likert_scale

5. The Future: Search Assistance

their own. The domain chosen is "some major cities of German and English speaking countries".

After a motivation for the domain chosen is given, the behaviour of a conventional search engine on that domain is examined. Finally, the idea of a search assistant is explained on the domain chosen. In the next chapter, an implementation of a very early version of a search assistant is illustrated and analysed.

5.2.1. Motivation for the Domain Chosen

The domain "some major cities of German and English speaking countries" was chosen as it is a small and independent domain. The content of the domain comprises images, key facts (for instance, area and population), some general information, points of interests, and history. The data for that is rather stable as it is mostly independent of today's political influences. Furthermore, it is widely provided and available (respectively accessible) on the Internet.

5.2.2. Example of a Conventional Search on Domain Chosen

Conventional search engines give users a SERP as described in chapter Search Engine Result Page (SERP). As mentioned there, the site is split into different parts, not all of them showing up on every SERP. In the following, the SERP for the search term "Vienna" on Google is examined. Figure 5.1 illustrates the upper SERP, Figure 5.2 illustrates the lower SERP. A splitting was done due to lack of space here. Anyway, it can be seen that the SERP is split into different parts.

The first part (marked as "1" in Figure 5.1) is an entry which corresponds to Wikipedia's article of "Vienna". A link with the title of the site is shown, always in bigger font. Then, the exact URL (sometimes the nesting of the linked site instead) is shown. Additionally, some options are available, as a link to the cached version of the site or a link to retrieve similar sites. Then, a short snippet of the linked site is following, where the search term is

5. The Future: Search Assistance

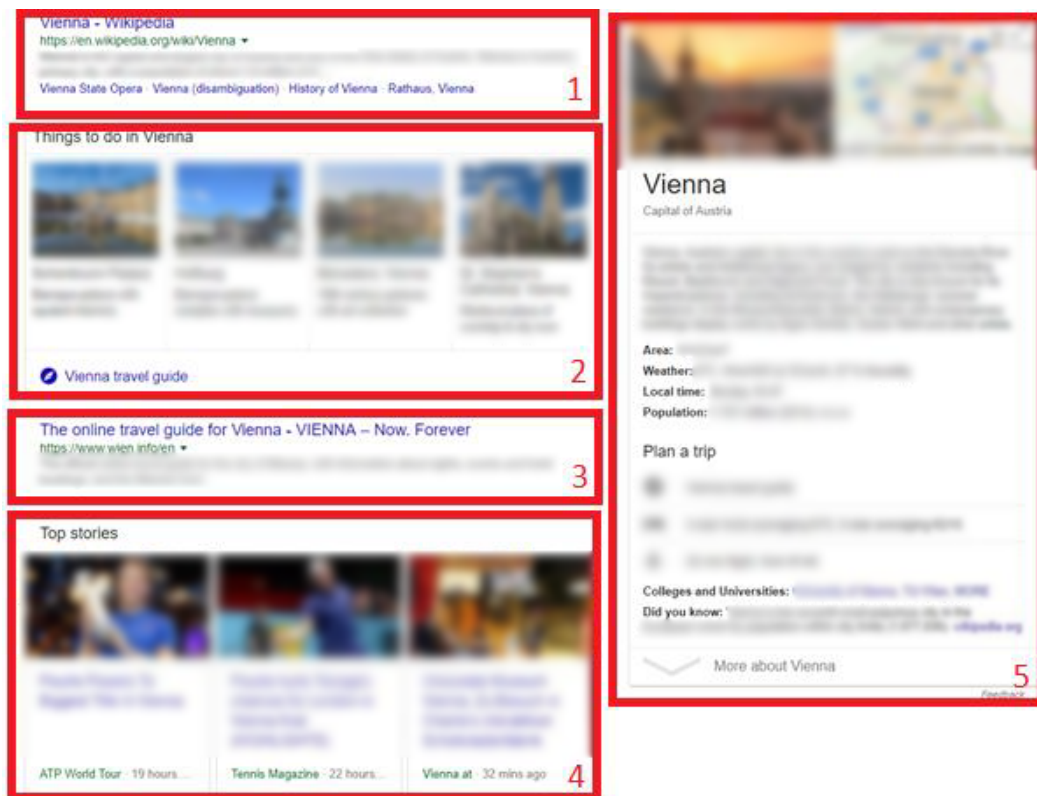


Figure 5.1.: Upper SERP for "Vienna" on Google

5. The Future: Search Assistance

highlighted. At last, there are links to possibly useful sites in the subdomain of the linked site. These links are not always shown but mostly when the linked site is an article on Wikipedia.

The second part (marked as "2" in Figure 5.1) represents a list of things to do in Vienna including a link to a travel guide for the city. Part three (marked as "3" in Figure 5.1) is very similar to part one. Part four (marked as "4" in Figure 5.1) represents a list of top stories concerning Vienna. It looks similar to part two. The fifth part (marked as "5" in Figure 5.1) is the Knowledge Graph of Vienna containing an image, a map and some key facts.

The lower SERP looks similar, see Figure 5.2. Part six and eight (marked as "6" respectively "8" in Figure 5.2) are analogous to part one (and three). The seventh part (marked as "7" in Figure 5.2) is similar to part two (and four). The ninth part (marked as "9" in Figure 5.2) is a list of related searches.

All these different parts are more or less classifications of (different) topics and interests, for instance general information, things to do, top stories, and images. This approach to split up a SERP into different parts or topics is used for almost every search term which is part of the Knowledge Graph, see Figure 4.6 and Figure 4.7 for two more examples. Depending on the search term, different templates for the SERP are used. For example, a search for a city results in general information, things to do, top stories and images. Contrary, a search for a famous person results in general information and top stories mostly, see Figure 5.3.

5.2.3. Search Assistance on Domain Chosen

The proposed idea of a search assistant on that domain is to take the approach of splitting the content into different parts and enhance it even more. Hence, users get a report for a city which contains various categories or perspectives. The content of that reports is collected from various sources on the Internet, for instance, Google Knowledge Graph and Wikipedia but also local providers as websites on tourist information and municipal governments. Information under Creative Commons license⁶ can be stored

⁶https://en.wikipedia.org/wiki/Creative_Commons_license

5. The Future: Search Assistance

Things to do in Vienna, Austria - Austria.info
<https://www.austria.info> > Home > Where to Go > Cities > Vienna ▼
Vienna's history dates back to the 3rd year Christian century when the Romans established the military camp *Vindobona*. Today's cityscape is characterized by ...


wien.at - Information and services of the City of Vienna available on ...
<https://www.wien.gv.at/english/> ▼
wien.at - the online information service of the city of Vienna

Visiting Vienna - wien.at
<https://www.wien.gv.at> > English > Culture & History ▼
The Vienna Tourist Board provides you with the most current and comprehensive information about sights, hotels and leisure activities in Vienna to go.

Vienna 2017: Best of Vienna, Austria Tourism - TripAdvisor
<https://www.tripadvisor.com> > Europe > Austria > Vienna Region ▼
Where to stay, restaurants, nightlife, reviews of Vienna hotels, attractions, and transportation making your trip to Vienna easier.

6

Images for vienna



→ More images for vienna Report images **7**

About Vienna: Your Vienna travel guide | Tourism Vienna, Wien ...
<https://www.aboutvienna.org/> ▼
About Vienna - Get to know Vienna with its great city and culture with: Sightseeing, History, Culture, Accommodation, Study and Work, Best Addresses, Travel ...

8

Searches related to vienna

vienna points of interest	vienna things to do
vienna italy	vienna austria weather
vienna language	vienna flights
vienna austria map	wien austria

9

Figure 5.2.: Lower SERP for "Vienna" on Google

5. The Future: Search Assistance

The image shows a Google search result for "Ryan Reynolds". On the left, there are several links for general information: IMDb, Twitter (@VancityReynolds), Wikipedia, and Instagram (@vancityreynolds). On the right, there is a detailed information card for Ryan Reynolds, including his birth date (October 23, 1976), height (1.83 m), spouse (Blake Lively), and a list of movies like "Deadpool" and "The Hitman's Bodyguard". Below the information card, there are social media profiles for Twitter, Instagram, Facebook, and LinkedIn. At the bottom, there are "People also search for" suggestions.

Top stories

- Ryan Reynolds Wished His Brother A Happy Birthday The Only Way He Knows How**
HuffPost
1 day ago
- Ryan Reynolds shares adorable throwback photo for his brother's birthday – see it**
Today.com
22 hours ago
- Ryan Reynolds Wishes Brother Happy Birthday with Throwback Photo**
PEOPLE.com
1 day ago

[More for Ryan Reynolds](#)

Ryan Reynolds - Home | Facebook
<https://www.facebook.com/VancityReynolds/>

Figure 5.3.: Upper SERP for “Ryan Reynolds” on Google containing general information and top stories.

5. The Future: Search Assistance

locally to improve performance as it is fetched from local storage. Although, these data have to be checked for changes on a regular basis.

Optimally, the different sources for a category are obtained automatically from a ranking which is computed by the search engine. As mentioned previously, the process of fetching and parsing has to be supported for all these sources. This is a hard challenge, to be discussed in the next section.

Each snippet extracted is stated with its source to give users the possibility for further reading. In addition, a rating for each source is available to give users a kind of measurement of the quality. As the rating is obtained from users, they may have the possibility to evaluate them. Further on, users can exclude unwanted sources from the report.

As some content of the report is fetched live from different sources on the Internet, it might contain copyright material. Therefore, the report generated is for the individual user only. If they store it, they must not disseminate it under any circumstances. The content is stored temporarily during the generation of a SERP for a user only.

By this approach, users get a broad summary on a topic containing different categories/perspectives without leaving the website of the search engine. The displayed ratings for the sources give users a kind of a value on the quality of the extracted content. Interaction is improved as users can rate sources. The search engine gets an immediate feedback from its users. If they want to explore a topic in depth, they can follow the stated links to the original sources. By offering checkboxes to switch on/off categories in the report, users get even more flexibility on the content of them. Figure 5.4 illustrates a possible structure of a SERP by the proposed search assistant rudimentary.

5.3. Discussion

In this section, the advantages and drawbacks of the search assistant proposed are discussed. This work tries to point out that benefits by far outweigh disadvantages. As this proposal is based on theoretical thoughts,

5. The Future: Search Assistance

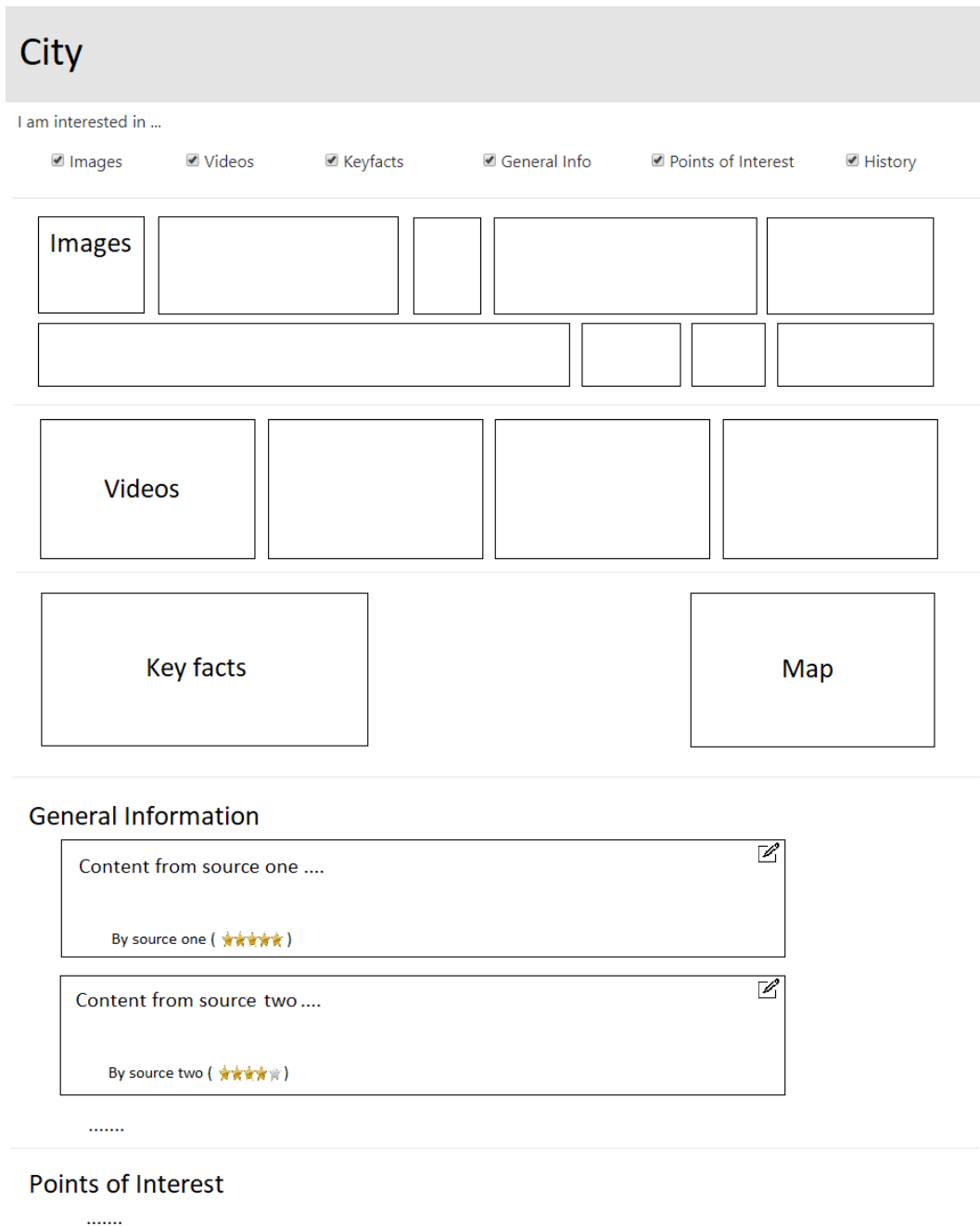


Figure 5.4.: Possible structure of a SERP by a search assistant

5. The Future: Search Assistance

some additional drawbacks might appear during the process of implementation.

5.3.1. Improvements in Search

In the following, the main advantages of the proposed search assistant are listed. Most of them were already touched briefly in the previous sections. A key advantage is the involvement of users. By improving the interaction between the search engine and its users, it is easier to obtain the thoughts of the individual user.

1. Convenience of Searching

The first improvement to mention is the convenience for users when they do a search. As the search assistant sifts through sources automatically, users are not obliged to navigate through a list of search results on their own. For the user, calling external sites is an annoying task. Although a conventional search engine tries to rank "useful" at the top, many troubles might occur. For instance, the external site has an inconvenient design for the user or it is filled up with advertisements which hide actual content. Other possibilities which might occur are a time-out when calling the site. This is bothersome for the user. A search assistant fetches content using parallel threads. If a source/server is not available users would not notice that in particular.

Further on, users get a prepared report of a topic containing different aspects or categories. Imagine a search query for a city. It is easy to derive that users are most interested in one or more of the following aspects: images, videos, key facts (population, area, etc.), general information, points of interest, or history. By using a search assistant, a report containing all these facets is generated, all automatically. The user can unselect categories of no interest to get a report even more custom. Here, an improvement of interaction is the key-concept used to understand the user.

By including language tools to translate content to the desired languages automatically, much more information is available and reachable for users. They may not even recognize that parts of the presented report are written

5. The Future: Search Assistance

in another language originally. Only by following the stated link to the source they may notice it.

2. Sources for Data Retrieved

Using the proposed approach of a search assistant, every single snippet within a report is stated with its source. Therefore, for users, it is transparent where the data comes and if they trust the source. As a consequence, users do also have the possibility for further reading. If an extracted snippet in the report misses a definition users might find it following the link to the original content.

3. User Interaction and Feedback System

There are many positive effects regarding an enhancement of interaction. This work tries to point out that an improvement of interaction between the search engine and its users is required to get the intention of a user.

A first example was given with ambiguous search terms. If users enter an ambiguous search term they must be informed about that issue. By asking users to specify their area of interest the search engine gets information on the field of interest at first hand. This applies not only for ambiguous search terms but for all search queries that have different aspects. Thus, almost every search query. By specifying the interest, the search engine can react and restrict results accordingly. Alternatively, users specify their area of interest even before they enter a search query. In that case, the search engine can restrict the results even before.

Another aspect of an improvement of interaction is that the search engine gets information about the quality of the content (and source) immediately by users. They have the possibility to rate different aspects of the sources for the content on an extra page. This is not a trivial thing and very subjectively. It has to be done carefully. Therefore, it is an option to restrict the process of rating to chosen users only. These chosen users can be either trusted users by the search engine or registered users to open contribution for everyone (if they are willing to register). The more users contribute to the rating system, the more the rating of a source is informative. Although, if all users have the opportunity to rate, it would be easier to manipulate it. The information of the rating is stored externally in a database to avoid direct manipulation.

5. The Future: Search Assistance

Also, it is not a trivial thing which aspects are open for rating. This work follows the guidelines elaborated in (Handschack et al., 2006). They propose ten criteria to check if a source is reliable:

- *Core Requirement*: Does the site have an imprint? Who is the owner? Who is running the server?
- *Publisher and Authors*: Who is the publisher? A well-known organisation, a company, a private person, or someone else? Is the author an expert in that field?
- *References*: Who links to that site? Where do links on the site point to? Are these references trustworthy?
- *Content and Scientificity*: Is the content (text, images, etc.) well-cited? How detailed is the information provided described? Is the content provable by using cross-checks with other sites?
- *Up-To-Dateness*: When was the content modified at last? How old are referenced sources?
- *Functionality and User-Friendliness*: Is the site intuitive and well-structured? Is additional media (images, videos, etc.) provided?
- *Formal Writing*: Is the content free from grammatical errors, misspelling, and wrong punctuation?
- *Impartial or Subjective Content*: Is the content impartial or is it an opinion of a single author? Is there an intentional lack of information?
- *Pop-ups, Cookies, Plugins*: Are there any annoying pop-ups? What are the usage terms for cookies? Does the site force an installation of software?
- *Advertisements*: Are there advertisements shown? Which kind of advertisements are shown?

Other aspects are the response-time of the server and its general availability (up-time). The whole process of rating sources has to be done much more precisely. In this work, only a brief touch on this topic is given.

Further on, a possibility to submit feedback directly to the search engine is given. Users may have found feasible sources for a topic on their own. By offering a request form to add sources the interaction gets improved even more. Of course, a user can also report sources that may not be reliable for a topic.

5. The Future: Search Assistance

4. No Copyright Infringements

By fetching all content live and automatically from information-providing websites, no copyright infringement is done. The content is stored during the generation of a SERP in the main memory of the server only. No files are stored on the disk. As mentioned before, an extension is to store data known as open-source locally on the server to improve performance. Anyway, users must not disseminate the report generated under any circumstances. It is for the individual user only. Of course, users must be informed about that issue.

5.3.2. Challenges and Problems

The idea of a search assistant poses both challenges and drawbacks. In the following, a list of possibly hard tasks is given. Certainly, the list given is not complete. Further problems might occur during the process of realisation.

1. Avoidance of Redundancy

As mentioned before, the avoidance of redundancy in a report is a huge challenge. The collection of information from various sources with as little redundancy and as little loss is called information consolidation (Maurer and Mehmood, 2015). There are several approaches for information consolidation. For images, feature-based similarity algorithms could be used, as explained in (Schreck, 2016). However, it has to be evaluated if that approach could be used in a real-time application. Text similarity is even more difficult to compute. Requirements for an algorithm are not only a comparison of text but also for meaning. In this work, that issue is just mentioned without further investigations. In general, no satisfying algorithm for information integration has been discovered yet (Maurer and Mehmood, 2015), but progress is continuing.

Alternatively, the selection of a master source which provides most of the content of the report could be tackled. To fill-up missing information, some minor sources could be used. In that case, a comparison of images and text has to be done as well. Again, this is a hard task and not covered in this work.

5. The Future: Search Assistance

2. Selection of High-Quality Sources

Another major challenge is to choose high-quality sources for a report. In the beginning, lists containing topics and corresponding URLs may be deposited. These lists are created mostly manually. The search engine calls the URLs and fetches, parses, and finally presents the content to users. Regardless of the topic, Google (Knowledge Graph) and Wikipedia are good providers of information.

It is a huge challenge to create lists manually and certainly not possible for every topic. Thus, this process should be automated. During a search, the search engine sifts through its index to find relevant results already. Therefore, access to relevant URLs for every search query is given.

3. Support of (Local) Providers

Not only the selection of high-quality sources is a challenge. Also, communication with APIs and parsing of HTML responses are very individual for a source/website. For each source, a parsing algorithm which filters information has to be implemented. That results in a huge overload for programmers. Moreover, format-changing is another critical aspect. While APIs (for instance, Wikipedia) may not change frequently, it is probable that HTML structure is updated continuously by information-providing sites. If that happens, the parsing algorithm has to be adopted which is a great deal of work.

Also, the support for local providers is worth to mention. Local providers are information-providing sites besides big-players like Google and Wikipedia which offer information on a very specific topic. Besides that it is exhausting to implement many parsing algorithms, the more different sources are available for a report the better. By supporting local providers another drawback arises. These websites often supply content in a different language than the desired one. The aim is to use software to translate content in a foreign language, for instance, Google Translate. This takes additional computation power and therefore time. Finally, these services and also information-providing websites may have a restriction on automatic requests or at least a maximum number per IP address. This thesis tries to examine the possibility of a search assistant on a large scale. Much more work is to be done for realisation. Without finding and cooperating with powerful

5. The Future: Search Assistance

partners like Google, it is unlikely that a search assistant as described can be offered on a large scale.

4. Report Generation in Real-Time

Calling an URL and fetching, parsing, and finally presenting content to users is very time-consuming compared to retrieval of information from an index. Using a laptop with Intel(R) Core(TM) i5-3317 dual-core (each 1.70 GHz), 8 GB RAM, and 4G (20 MBit/s) Internet connection, fetching, parsing, and presenting of 15 URLs takes up to 10 seconds. By increasing computer power and Internet connection speed, these tasks will take at least a few seconds nonetheless. Besides that, storing open-source material locally to potentially improve performance has its disadvantages. Such data has to be kept up-to-date. That means it has to be checked for modification in certain time intervals, resulting in additional effort. As most of the data is fetched at run-time the improvement in performance is probably negligible. Anyway, in this work, the focus is on other aspects of a search assistant. Therefore, no precise results can be provided according to the practicability of a real-time generation on a large scale.

5. No Consistent Structure of a Report

Taking the example of a search query for a city. For a human, it is quite intuitive that a report generated has categories like images, videos, key facts, general information, points of interest, and history. Anyway, a search for a book results in different aspects users are interested. However, they might want categories like images, videos, key facts as well. Further categories of interest might be the synopsis, details about development and publication, legacy, and others. As mentioned before, almost every search query has different aspects a user might be interested in. As the task of finding categories manually is very exhausting and not doable for every topic, this process has to be done automatically. It needs to attach content with some metadata describing it. Therefore, software tools to determine the meaning of text are needed which is a hard problem. Anyway, this aspect was not the focus of this thesis. It is left to further research.

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

For demonstration and testing purposes, a web application was implemented. In this section, it is explained and discussed. It is a pre-version of a search assistant for the domain chosen ("some major cities of German and English speaking countries"), implemented using JSP (JavaServer Pages). Technologies used were Java, HTML, CSS (Cascading Style Sheets), JavaScript and JSON (JavaScript Object Notation). Special attention was paid to HTML validity and responsive design. Latter was very much easier with Bootstrap¹. The search assistant was named *Sistant*, a short version for the term "Search assISTANT". For a detailed look at the implementation, design and the motivation behind it, see Appendix Implementation Details of Sistant.

For each supported city a file is deposited on the server containing information about sources. For each category, a list of server URLs is stored. Within the generation of a report, these lists of servers are sifted and the content of the URL is fetched, parsed, and finally presented to the user. Users can configure these lists of servers on an extra page, see Section Server Configuration in Sistant.

Currently, 24 cities in three countries are supported. Main sources for reports are Google, Wikipedia, Infoplease², and Britannica³. For the full list of supported sources per category, see Section Supported Sources per

¹<https://getbootstrap.com>

²<https://www.infoplease.com>

³<https://www.britannica.com>

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

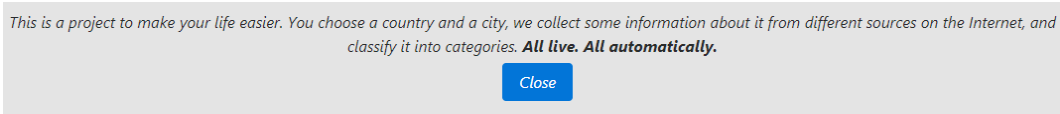


Figure 6.1.: Hint in Sistant when a user opens the website for the first time

Category. As these sources offer a broad number of cities, an extension to more cities can be done easily. An extension to more sources per city takes more effort as local providers (for instance, tourist information sites) have to be supported. This requires a selection of appropriate sources and distinct implementations of parsers for each source.

6.1. SERP in Sistant

By calling the website of Sistant the very first time, a hint at the top of the page is shown, explaining what this project is actually about, see Figure 6.1. Furthermore, a SERP for "London" is shown. All interests and sources are selected by default. In the following, the particular parts of this SERP are examined.

1. Header

At the top of the page, users can switch between three different countries (namely *Austria*, *United Kingdom*, and *United States*) and its (bigger) cities, see Figure 6.2. For the full list of supported cities per country, see Appendix Supported Countries and Cities. As mentioned before, the number of supported cities can be increased easily, as Google, Wikipedia, Infoplease, and Britannica support a wide range of cities. Furthermore, buttons for generating a report (after a change of the country/city) and for configuring the sources are available.

2. Interest Selection

Further on, users can configure the categories the corresponding report should contain by using checkboxes. Currently, images, key facts (area,

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

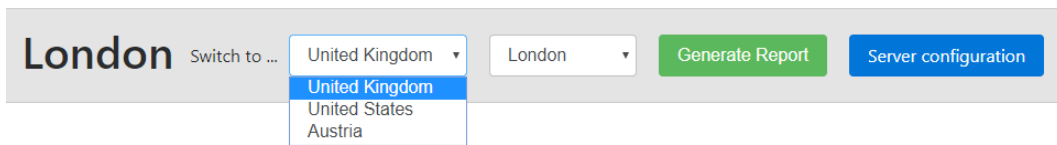


Figure 6.2.: Header in Sistant

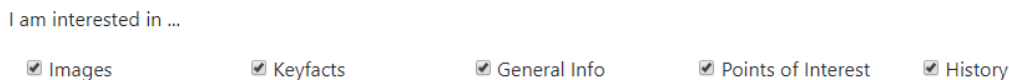


Figure 6.3.: Selectable interests in Sistant

population, and a map), general information, points of interest, and history are available, see Figure 6.3.

3. Images

The images are taken from different image-search portals. It is worth to mention that no images are stored on the local server. Only links are extracted and set as image sources in the corresponding HTML tags. In that version of Sistant, the most popular websites offering an image search are fetched. These are Google Images⁴, Pixabay⁵, and Flickr⁶. Images for a “coats of arms” or a “seal of the city” are shown always in first place. They are taken mostly from Wikipedia. Anyway, these links are hardcoded in the corresponding configuration file of the city.

For the layout of the images, a pure CSS implementation from GitHub⁷ was taken, it was created by the user xieranmaya (Ran, 2016). Its special feature is that independent from the number of images, a grid is shown which fulfils the available width and keeps the ratio of images. Also on resizing, the layout is not recalculated as it is a CSS-only implementation. See Figure 6.4 for London’s image category.

⁴<https://images.google.com>

⁵<https://www.pixabay.com>

⁶<https://www.flickr.com>

⁷<https://www.github.com>

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant



Figure 6.4.: Image category for London in Sistant

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

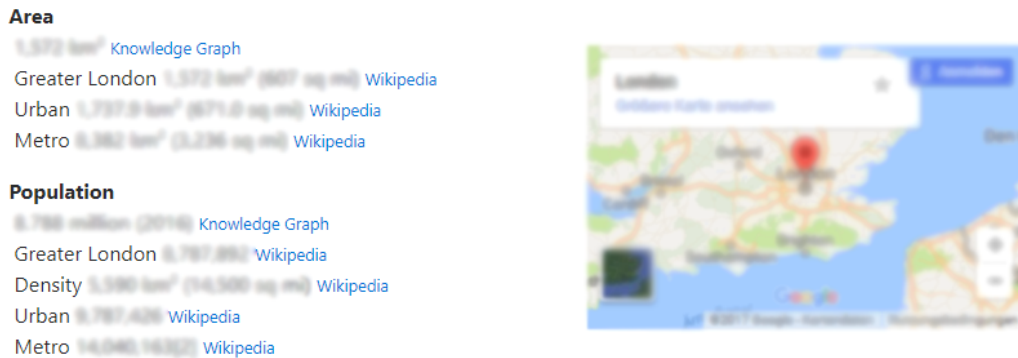


Figure 6.5.: Key facts category for London in Sistant

4. Key Facts

Currently, key facts of a city comprise its area, population, and a map with all its facets. Two different sources are used for this data, Wikipedia and Google. In the most cases, Wikipedia offers a distinction between urban area/population and metro area/population. Google offers a single number only. Anyway, Sistant parses the content of the sites and extracts all available area/population data. The map shown is an embedded version of Google Maps, its link is hard-coded in a configuration file of a city. See Figure 6.5 for London's key facts category.

5. General Information

The general information of a city is split into different paragraphs. Each of them is taken from a different source. The length of a single paragraph is limited to 2500 characters. If a source supplies more than 2500 plus 40 (tolerance) characters of content, the exceeding content is cut off. This mechanism was introduced to avoid too long paragraphs and therefore reports. At the end of every paragraph, the source is stated as a link to the page where the information was retrieved. Currently, four providers are supported, namely Wikipedia, Google, Infoplease, and Britannica. All these sources supply content for all supported cities. In this version, no functionality to rate the sources is implemented yet. Figure 6.6 shows two paragraphs of London's general information in Sistant.

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

General Information

London, the capital of England and the United Kingdom, is a 21st-century city with history stretching back to Roman times. At its center stand the imposing Houses of Parliament, the iconic Big Ben clock tower and Westminster Abbey, sites of British monarch coronations, across the Thames River. The London Eye observation wheel provides panoramic views of the South Bank cultural complex, and the entire city.

By Google.

London (/ˈlʌndən/ *listen*) is the capital and most populous city of England and the United Kingdom.^[32] Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.^[16] London's ancient core, the City of London, largely retains its 1.12-square-mile (2.9 km²) medieval boundaries. Since at least the 18th century, 'London' has also referred to the metropolis around the core, historically split between Middlesex, Essex, Surrey, Kent and Northamptonshire,^[17]^[22] which today largely makes up Greater London,^[14]^[15] (note 1) a region governed by the Mayor of London and the London Assembly.^[14] (note 2)^[17]

London is a leading global city^[18]^[19] in the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation.^[20]^[21]^[22] It is the world's largest financial centre^[23]^[24]^[25] and has the 9th- or 10th-largest metropolitan area GDP in the world.^[note 3]^[26] London is a world cultural capital^[27]^[28]^[29] It is the world's most-visited city as measured by international arrivals^[30] and has the world's largest city airport system measured by passenger traffic.^[31] It is the world's leading investment destination,^[32]^[33]^[34]^[35] hosting more international retailers^[36]^[37] and ultra-high-net-worth individuals^[38]^[39] than any other city. London's universities form the largest concentration of higher education institutes in Europe.^[40] In 2012, London became the first city to have hosted the modern Summer Olympic Games three times.^[41]

London has a diverse range of people and cultures, and more than 300 languages are spoken in the region.^[44] Its estimated mid-2016 municipal population (corresponding to Greater London) was 8,787,882,^[2] the largest of any city in the European Union^[45] and accounting for 13.8% of the UK population.^[46] London's urban area is the second most populous in the EU, after Paris, with 8,787,438 inhabitants at the 2011 census.^[47] The city's metropolitan area is the most populous in the EU with 14,898,162 inhabitants in 2016.^[note 4]^[48] While the Greater London Authority states the population of the city-region (covering a large part of the south east) as 21.7 million.^[49]^[50] London was the world's most populous city from around 1821 to 1925.^[51]

London contains four World Heritage Sites. ...

By Wikipedia.

Figure 6.6.: General Information category for London in Sistant

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

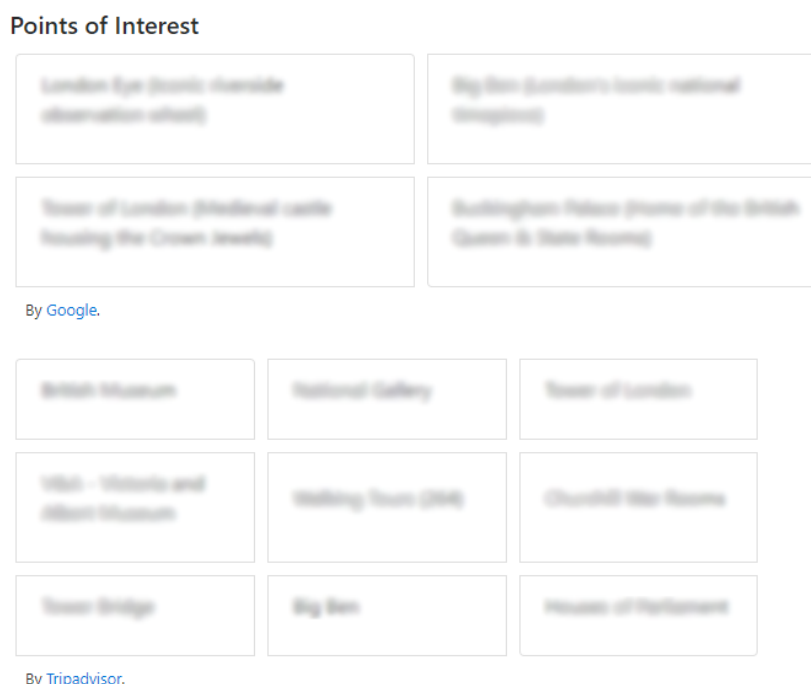


Figure 6.7.: Points of interest grids for London in Sistant

6. Points of Interest

The points of interest are displayed depending on the source. For the sources Wikipedia and Infoplease the same approach as in category "General Information" is used. The same mechanism for the length of paragraphs is applied. For the sources Google and Tripadvisor⁸ a grid is shown containing the highest-ranked points of interest for the city chosen, see Figure 6.7 for London's points of interest grids.

7. History

The history of a city looks pretty much like the category of "General Information". Again, the same mechanism for the length of paragraphs is applied. Currently, three providers are supported, namely Wikipedia, Google, and Britannica.

⁸<https://www.tripadvisor.com>

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

8. Footer

The footer signals users that the tailored report might contain copyright information. Furthermore, a link to the imprint of the site is shown. The imprint contains information about the project Sistant itself and contact details.

6.2. Server Configuration in Sistant

For each city, all categories with its available sources are listed. The user can switch on/off selected servers by unchecking them. For images, the user has the possibility to configure how many images should be shown at most from a specific server. This feature gives users the possibility to kind of weigh the image servers. Figure 6.8 shows the server configuration for London.

6.3. Summary

The purpose of an implementation of a first version of a search assistant was to show that the concept proposed actually works. With Sistant, users get a broad overview of a topic (in that case, a city) without leaving the website of the search engine. As all data is fetched live and automatically, users can be sure to have the latest data available. Sources are always stated for further reading and users can switch off categories of no interest. Also, no copyright infringements are to be expected as no data is stored on the server of Sistant. Further on, the waiting time for a single user to get a report is a few seconds (Sistant running on a laptop providing Intel(R) Core(TM) i5-3317 dual-core each 1.70 GHz, 8 GB RAM, and 20 MBit/s Internet connection) and therefore tolerable.

Anyway, the issues addressed in section Challenges and Problems still persist. Sistant does not pay attention to a possible redundancy of content at all. No text is checked for similarity to other content. Moreover, all sources were picked by hand. Sources providing information with a wide coverage

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

I am interested in the following sources ...

Images

Google Images ([Link](#))

Max. number of pictures:

Flickr ([Link](#))

Max. number of pictures:

Pixabay ([Link](#))

Max. number of pictures:

Keyfacts

Google Knowledge Graph ([Link](#))

Wikipedia ([Link](#))

General Info

Google ([Link](#))

Wikipedia ([Link](#))

Infoplease ([Link](#))

Britannica ([Link](#))

Points of Interest

Google ([Link](#))

Wikipedia ([Link](#))

Infoplease ([Link](#))

Tripadvisor ([Link](#))

History

Google ([Link](#))

Wikipedia ([Link](#))

Infoplease ([Link](#))

Britannica ([Link](#))

Figure 6.8.: Server configuration for London in Sistant

6. Discussion of a First Version of a Search Assistant as Implemented – Sistant

of cities and content available in English were chosen. Thus, it would be easy to extend the range of cities supported. Furthermore, no rating for these sources was implemented. Although, sources can be unchecked if unwanted. As there is only one structure for a report needed it was created by hand and not automatically.

7. Summary and Future Work

Search engines have a powerful and indispensable standing nowadays. Without a functionality to get at least a brief overview of what is available on the web, most of the content would never be visible to the vast majority of users. As seen, Google, the most-used search engine, has a very sophisticated way to rank websites according to a search query.

Although ranking was improved over the last years, presentation of results has been made little progress. At least, users get some additional data presented for chosen search terms. These data are presented in addition to an actual ranking of websites within a knowledge panel.

In this work, some major drawbacks of Google's current structure of a SERP were pointed out. These include ambiguous search terms, the lack of sources stated for data presented, and the lack of definitions for data presented. This thesis proposed suggestions to get rid of these problems. For ambiguous search terms, more interaction between the search engine and its users is required to "understand" users. Further on, all data presented has to be stated with its source and also a definition for these data has to be given.

The main contribution of this thesis was a new approach for a search engine - a search assistant. On the one hand, a search assistant has the purpose to solve the issues mentioned before. On the other hand, it improves the convenience of searching for users massively by taking a lot of annoying work off the hands of them.

The idea of a search assistant was applied on a small domain ("some major cities of German and English speaking countries") and a discussion on the advantages and disadvantages was given. In this work, it was pointed out that the advantages by far outweigh the challenges. Main benefits are an enhancement of convenience for users, a rating system for all sources,

7. Summary and Future Work

an improvement of user interaction, and that all along with no copyright infringements. Besides that, some problems and drawbacks arose. The avoidance of redundant content is a non-trivial problem. It is an issue of general interest. As mentioned, no really satisfying approach has been discovered yet. Also, the selection of high-quality sources has to be automated as far as possible. Further on, approaches for including local providers in different languages have to be investigated. As far as possible, this process has to be automated as well. Moreover, the feasibility of report generation in real-time has to be examined in greater detail, taking in account many users in parallel. Finally, an automation for obtaining various categories for a generated report has to be found.

Appendix

Appendix A.

Implementation Details of Sistant

The software is web based and was implemented with Java Server Pages (JSP)¹, HTML, CSS, JavaScript, and JSON. The information on categories, supported cities, and its sources is stored in JSON-files which are loaded on application's start up.

As the different categories are independent from each other, the fetching process is split up into parallel working-threads to gain performance. Each category is packed into a Java package containing a Java class for each source (derived from a shared parent class) and a controller class which handles these sources. This design was chosen to facilitate extensions. For instance, a new source can be added simply by creating a new Java class implementing a specified fetch-method and an invocation in the controller.

For the design, some external libraries were used. These are described in the next section. Hence, the whole design is responsive and valid HTML5.

A.1. Libraries Used

The following libraries were used to improve beauty of the website and readability of the source code. Without these libraries the project would not be that well advanced in design and functionality.

¹https://en.wikipedia.org/wiki/JavaServer_Pages

Appendix A. Implementation Details of Sistant

- Bootstrap v.4.0.0 (<https://v4-alpha.getbootstrap.com>) for front-end design supporting design and responsiveness
- JQuery v.3.2.1 (<https://developers.google.com/speed/libraries/#jquery>) along with Bootstrap and for simplification of client-side scripting
- Gson v.2.2.4 (<https://github.com/google/gson>) for converting JSON files to Java Objects
- Jsoup v.1.10.3 (<https://jsoup.org/download>) for fetching and parsing HTML easily

A.2. JSON Files

In this section the handling of the configuration files and its structures are illustrated. One file contains information on the supported countries and cities. Another one contains all available interests. The third one contains information on all supported sources for a city.

A.2.1. File Handling

For processing the JSON files an external library was used, Gson. Gson supports both directions, conversion of Java Objects into JSON and the other way around. The library is released under the Apache 2.0 License². For Sistant the way of converting a JSON string into Java Objects was used only. The following example illustrates a conversion for London's JSON representation into a Java Object, SourcesOfCity describes the corresponding object.

```
URL url = new
    URL("http://localhost:8080/data/United%20Kingdom/London.json");
BufferedReader in = new BufferedReader(new
    InputStreamReader(url.openStream()));
Gson gson = new GsonBuilder().create();
this.city = gson.fromJson(in, SourcesOfCity.class);
```

²<https://github.com/google/gson/blob/master/LICENSE>

A.2.2. Regions.json

The JSON file for supported regions contains an array of countries and for each country an array of cities. The "selected" attribute specifies default values.

```

1  [{
2    "name": "United Kingdom",
3    "cities": [
4      {"name": "London", "selected": "true"},
5      {"name": "Birmingham", "selected": "false"},
6      {"name": "Edinburgh", "selected": "false"},
7      {"name": "Glasgow", "selected": "false"},
8      {"name": "Manchester", "selected": "false"},
9      {"name": "Liverpool", "selected": "false"},
10     {"name": "Leeds", "selected": "false"},
11     {"name": "Belfast", "selected": "false"}
12   ],
13   "selected": "true"
14 },
15 {
16   "name": "United States",
17   "cities": [
18     {"name": "New York City", "selected": "true"},
19     {"name": "Washington, D.C", "selected": "false"},
20     {"name": "Chicago", "selected": "false"},
21     {"name": "Los Angeles", "selected": "false"},
22     {"name": "Houston", "selected": "false"},
23     {"name": "Phoenix", "selected": "false"},
24     {"name": "Philadelphia", "selected": "false"},
25     {"name": "San Antonio", "selected": "false"},
26     {"name": "San Diego", "selected": "false"},
27     {"name": "Dallas", "selected": "false"},
28     {"name": "San Jose", "selected": "false"}
29   ],
30   "selected": "false"
31 },
32 {
33   "name": "Austria",

```

Appendix A. Implementation Details of Sistant

```
34 "cities": [  
35   {"name": "Graz", "selected": "true"},  
36   {"name": "Vienna", "selected": "false"},  
37   {"name": "Salzburg", "selected": "false"},  
38   {"name": "Innsbruck", "selected": "false"},  
39   {"name": "Klagenfurt", "selected": "false"}  
40 ],  
41 "selected": "false"  
42 ]]
```

A.2.3. Interests.json

Each supported interests is an element containing information about the display name and if the interest is selected by default.

```
1 {  
2   "images":  
3   {  
4     "displayName": "Images",  
5     "selected": "true"  
6   },  
7   "keyfacts":  
8   {  
9     "displayName": "Keyfacts",  
10    "selected": "true"  
11  },  
12  "generalinfo":  
13  {  
14    "displayName": "General Info",  
15    "selected": "true"  
16  },  
17  "poi":  
18  {  
19    "displayName": "Points of Interest",  
20    "selected": "true"  
21  },  
22  "history":
```

Appendix A. Implementation Details of Sistant

```
23 {
24   "displayName": "History",
25   "selected": "true"
26 }
27 }
```

A.2.4. City Files

The structure of the JSON file of a city is illustrated schematic only. The real URLs were substituted by the string "url".

```
1 {
2   {
3     "name": "London",
4     "country": "United Kingdom",
5     "gmapsurl": "url",
6     "coatOfArmsUrl": "url",
7     "categories": [
8       {
9         "name": "images",
10        "displayName": "Images",
11        "servers": [
12          {
13            "name": "Google Images",
14            "url": "url",
15            "numberOfPictures": 5,
16            "active": "true"
17          },
18          {
19            "name": "Pixabay",
20            "url": "url",
21            "numberOfPictures": 5,
22            "active": "true"
23          },
24          {
25            "name": "Flickr",
26            "url": "url",
```

Appendix A. Implementation Details of Sistant

```
27     "numberOfPictures": 5,  
28     "active": "true"  
29   }]  
30 },  
31 {  
32   "name": "keyfacts",  
33   "displayName": "Keyfacts",  
34   "servers": [  
35     {  
36       "name": "Google Knowledge Graph",  
37       "url": "url",  
38       "active": "true"  
39     },  
40     {  
41       "name": "Wikipedia",  
42       "url": "url",  
43       "active": "true"  
44     }]  
45   },  
46   {  
47     "name": "generalinfo",  
48     "displayName": "General Info",  
49     "servers": [  
50       {  
51         "name": "Google",  
52         "url": "url",  
53         "active": "true"  
54       },  
55       {  
56         "name": "Wikipedia"  
57         "url": "url",  
58         "active": "true"  
59       },  
60       {  
61         "name": "Infoplease",  
62         "url": "url",  
63         "active": "true"  
64       },  
65     ]  
66   }
```

Appendix A. Implementation Details of Sistant

```
66     "name": "Britannica",
67     "url": "url",
68     "active": "true"
69   }]
70 },
71 {
72   "name": "poi",
73   "displayName": "Points of Interest",
74   "servers": [
75     {
76       "name": "Google",
77       "url": "url",
78       "active": "true"
79     },
80     {
81       "name": "Wikipedia",
82       "url": "url",
83       "active": "true"
84     },
85     {
86       "name": "Infoplease",
87       "url": "url",
88       "active": "true"
89     },
90     {
91       "name": "Tripadvisor",
92       "url": "url",
93       "active": "true"
94     }
95   ],
96 {
97   "name": "history",
98   "displayName": "History",
99   "servers": [
100    {
101      "name": "Google",
102      "url": "url",
103      "active": "true"
104    },
```

Appendix A. Implementation Details of Sistant

```
105     {
106         "name": "Wikipedia",
107         "url": "url",
108         "active": "true"
109     },
110     {
111         "name": "Infoplease",
112         "url": "url",
113         "active": "true"
114     },
115     {
116         "name": "Britannica",
117         "url": "url",
118         "active": "true"
119     }
120 ]
121 }
122 }
```

A.3. Fetching and Parsing of HTML

For fetching and parsing HTML content from the different sources, the Java library "Jsoup" was used. It is an open source project distributed under MIT License. It offers an easy API for fetching, extracting and manipulating (HTML) data, using DOM based, CSS and JQuery-like methods. Taken from Jsoup's website³, the following example fetches Wikipedia's homepage and stores all "In the news" headlines in a list.

```
Document doc = Jsoup.connect("http://en.wikipedia.org/").get();
log(doc.title());
Elements newsHeadlines = doc.select("#mp-itn b a");
for (Element headline : newsHeadlines) {
    log("%s\n\t%s", headline.attr("title"), headline.absUrl("href"));
}
```

³<https://jsoup.org>

A.4. Supported Countries and Cities

Currently, the following countries and cities are supported:

- Austria
 - Graz
 - Vienna
 - Salzburg
 - Innsbruck
 - Klagenfurt
- United Kingdom
 - London
 - Birmingham
 - Edinburgh
 - Glasgow
 - Manchester
 - Liverpool
 - Leeds
 - Belfast
- United States
 - New York City
 - Washington, D.C.
 - Chicago
 - Los Angeles
 - Houston
 - Phoenix
 - Philadelphia
 - San Antonio
 - San Diego
 - Dallas
 - San Jose

A.5. Supported Sources per Category

The focus on implementing parsers for specific sources was a high level of reuse. Most of the chosen sources offer information for a huge number of bigger cities. Especially, Google and Wikipedia have information about almost every city or even village. Currently, the following sources per category are supported:

- Images
 - Google (<https://images.google.com>)
 - Pixabay (<https://www.pixabay.com>)
 - Flickr (<https://www.flickr.com>)
 - Static image by link
- Key Facts
 - Google (<https://www.google.com>)
 - Wikipedia (<https://en.wikipedia.org>)
- General Information
 - Google (<https://www.google.com>)
 - Wikipedia (<https://en.wikipedia.org>)
 - Infoplease (<https://www.infoplease.com>)
 - Britannica (<https://www.britannica.com>)
- Points of Interest
 - Google (<https://www.google.com>)
 - Wikipedia (<https://en.wikipedia.org>)
 - Infoplease (<https://www.infoplease.com>)
 - Tripadvisor (<https://www.tripadvisor.com>)
- History
 - Google (<https://www.google.com>)
 - Wikipedia (<https://en.wikipedia.org>)
 - Infoplease (<https://www.infoplease.com>)
 - Britannica (<https://www.britannica.com>)

Bibliography

- Allon, Ori and Ken Wilder (2009). *Two new improvements to Google results pages*. URL: <https://googleblog.blogspot.co.at/2009/03/two-new-improvements-to-google-results.html> (visited on 07/14/2017) (cit. on p. 34).
- Bing (2008). *Robots Exclusion Protocol: joining together to provide better documentation*. URL: <https://blogs.bing.com/webmaster/2008/06/03/robots-exclusion-protocol-joining-together-to-provide-better-documentation/> (visited on 05/29/2017) (cit. on p. 7).
- Brin, Sergey and Lawrence Page (2012). "Reprint of: The anatomy of a large-scale hypertextual web search engine". In: *Computer networks* 56.18, pp. 3825–3833 (cit. on pp. 8, 10, 11).
- Bush, Michael (2016). *2016 Edelman Trust Barometer Finds Global Trust Inequality is Growing*. URL: <http://www.edelman.com/news/2016-edelman-trust-barometer-release/> (visited on 07/18/2017) (cit. on p. 45).
- Cutts, Matt (2009). *Detecting new Googlebombs*. URL: <https://publicpolicy.googleblog.com/2009/01/detecting-new-googlebombs.html> (visited on 06/05/2017) (cit. on p. 11).
- Dean, Brian (2016a). *Google's 200 Ranking Factors: The Complete List*. URL: <https://www.link-assistant.com/news/ranking-factors.html> (visited on 06/08/2017) (cit. on p. 13).
- Dean, Brian (2016b). *We Analyzed 1 Million Google Search Results. Here's What We Learned About SEO*. URL: <http://backlinko.com/search-engine-ranking> (visited on 06/08/2017) (cit. on p. 13).
- Farber, Dan (2013). *Google Search scratches its brain 500 million times a day*. URL: <https://www.cnet.com/news/google-search-scratches-its-brain-500-million-times-a-day/> (visited on 06/14/2017) (cit. on p. 16).
- Fishkin, Rand (2015). *Twitter post from June 21, 2015*. URL: <https://twitter.com/randfish/status/612730826615230464> (visited on 06/11/2017) (cit. on p. 14).

Bibliography

- Föllner, Patrick (2014). *Unterschiedliche Ergebnisse bei gleicher Suchanfrage? Wie Google Suchergebnisse personalisiert*. URL: <https://www.seonative.de/unterschiedliche-ergebnisse-bei-gleicher-suchanfrage-wie-google-suchergebnisse-personalisiert/> (visited on 06/14/2017) (cit. on p. 17).
- Global-Geography (2017). *Virtual Lab for Exploring Countries*. URL: <https://global-geography.org/af/Geography/Lab> (visited on 08/11/2017) (cit. on p. 44).
- Golub, Benjamin (2014). *Why does PageRank converge?* URL: <https://www.quora.com/Why-does-PageRank-converge> (visited on 06/05/2017) (cit. on p. 9).
- Google (2015). *Google Knowledge Graph Search API*. URL: <https://developers.google.com/knowledge-graph/> (visited on 05/31/2017) (cit. on pp. 28, 33).
- Google (2017a). *Crawling und Indexierung – Alles über die Suche*. URL: <http://www.google.com/insidesearch/howsearchworks/crawling-indexing.html> (visited on 04/26/2017) (cit. on p. 6).
- Google (2017b). *Domino - Google Search*. URL: <https://www.google.com/#q=Domino> (visited on 07/21/2017) (cit. on p. 38).
- Google (2017c). *Featured snippets in search*. URL: <https://support.google.com/webmasters/answer/6229325?hl=en> (visited on 06/05/2017) (cit. on p. 27).
- Google (2017d). *Galaxy - Google Search*. URL: <https://www.google.com/#q=Galaxy> (visited on 07/21/2017) (cit. on p. 36).
- Google (2017e). *How Google Search Works*. URL: <https://www.google.com/search/howsearchworks/algorithms/> (visited on 05/29/2017) (cit. on p. 6).
- Google (2017f). *How Google Search Works — Crawling & Indexing*. URL: <https://www.google.com/search/howsearchworks/crawling-indexing/> (visited on 05/29/2017) (cit. on pp. 7, 8, 12).
- Google (2017g). *Learn about sitemaps*. URL: <https://support.google.com/webmasters/answer/156184?hl=en> (visited on 06/01/2017) (cit. on p. 7).
- Google (2017h). *List of Austrian Nobel Prize winners - Google Search*. URL: <https://www.google.com/#q=List+of+Austrian+Nobel+Prize+winners> (visited on 07/22/2017) (cit. on p. 43).
- Google (2017i). *Men in Black - Google Search*. URL: <https://www.google.com/#q=Men+in+Black> (visited on 12/26/2017) (cit. on p. 45).

Bibliography

- Google (2017j). *Robots.txt-Spezifikationen*. URL: https://developers.google.com/webmasters/control-crawl-index/docs/robots_txt (visited on 05/29/2017) (cit. on p. 7).
- Google (2017k). *Sugar - Google Search*. URL: <https://www.google.com/#q=Sugar> (visited on 07/21/2017) (cit. on p. 38).
- Google (2017l). *Time - Google Search*. URL: <https://www.google.com/#q=Time> (visited on 07/21/2017) (cit. on p. 38).
- Google (2017m). *What is the best way to lose weight - Google Search*. URL: <https://www.google.com/search?q=What+is+the+best+way+to+lose+weight> (visited on 07/23/2017) (cit. on p. 27).
- Google (2017n). *What size does a soccer ball have - Google Search*. URL: <https://www.google.at/search?q=What+size+does+a+soccer+ball+have> (visited on 07/23/2017) (cit. on p. 27).
- Google Search Help (2017). *Update your location on Google*. URL: <https://support.google.com/websearch/answer/179386> (visited on 07/13/2017) (cit. on p. 34).
- Google Webmaster Central Blog (2010). *Using site speed in web search ranking*. URL: <https://webmasters.googleblog.com/2010/04/using-site-speed-in-web-search-ranking.html> (visited on 06/11/2017) (cit. on p. 14).
- Google Webmaster Central Blog (2016). *Mobile-first Indexing*. URL: <https://webmasters.googleblog.com/2016/11/mobile-first-indexing.html> (visited on 06/11/2017) (cit. on p. 14).
- Graap, Andreas (2015). *Wikipedia und Co.: So entsteht der Google Knowledge Graph*. URL: <https://blog.hubspot.de/marketing/wikipedia-und-co-so-entsteht-der-google-knowledge-graph> (visited on 05/30/2017) (cit. on p. 28).
- Grabowski, Paweł (2014). *30 Most Important Google Ranking Factors A Beginner Should Know*. URL: <https://unamo.com/blog/seo/30-important-google-ranking-factors-beginner-know> (visited on 06/08/2017) (cit. on pp. 13, 16).
- Grice, Tim (2012). *Tweets are an SEO ranking factor and this study proves it*. URL: <https://www.branded3.com/blog/tweets-vs-rankings/> (visited on 06/11/2017) (cit. on p. 15).
- Gütl, Christian (2014). "Slides from Lecture: Information Search and Retrieval VU, Part 8". In: (cit. on pp. 9, 10, 12).

Bibliography

- Handsack, Sandra et al. (2006). *Handbuch zur Recherche*. URL: <http://files.infokompetenz.de/Recherchehandbuch.pdf> (visited on 01/01/2018) (cit. on p. 62).
- Hinkis, Roy (2015). *Google steals over 550 million clicks from Wikipedia in 6 months*. SimilarWeb LTD. URL: <https://www.similarweb.com/blog/google-steals-over-550-million-clicks-from-wikipedia-in-6-months> (visited on 05/30/2017) (cit. on p. 39).
- Hölzle, Urs (2008). *Update to Google Suggest*. URL: <https://googleblog.blogspot.co.at/2008/09/update-to-google-suggest.html> (visited on 07/16/2017) (cit. on pp. 21, 35).
- Internet Live Stats (2017). *Anzahl der Webseiten weltweit in den Jahren 1992 bis 2015*. In *Statista - Das Statistik-Portal*. URL: <https://de.statista.com/statistik/daten/studie/290274/umfrage/anzahl-der-webseiten-weltweit/> (visited on 04/13/2017) (cit. on p. 5).
- Internet Live Stats (2000-2012) and Statistic Brain Research Institute (2013-2015) (2017). *Anzahl der Suchanfragen bei Google weltweit in den Jahren 2000 bis 2016 (in Milliarden)*. In *Statista - Das Statistik-Portal*. URL: <https://de.statista.com/statistik/daten/studie/71769/umfrage/anzahl-der-google-suchanfragen-pro-jahr/> (visited on 05/29/2017) (cit. on pp. 5, 35).
- Maksimava, Masha (2016). *Google 9 main ranking factors: the ultimate cheat sheet*. URL: <https://www.link-assistant.com/news/ranking-factors.html> (visited on 06/14/2017) (cit. on pp. 13-16).
- Maurer, Hermann (2012). "A linear ordering of a multi-parameter universe is usually nonsense". In: *Theoretical Computer Science* 429, pp. 222-226 (cit. on p. 12).
- Maurer, Hermann and Rizwan Mehmood (2015). "Merging image databases as an example for information integration". In: *Central European Journal of Operations Research* 23.2, pp. 441-458 (cit. on pp. 52, 63).
- Mehmood, Rizwan, Narayanan Kualthuramaiyer, and Hermann Maurer (2016). "A New Look at Geography of the World". In: *Special issue on New developments on the Web* 12.1 (cit. on p. 41).
- Mehmood, Rizwan and Hermann Maurer (2017). "Semi-Automatic Story Generation for a Geographic Server". In: *Kartografija i geoinformacije* 16.27, pp. 12-25 (cit. on p. 50).
- NetMarketShare (2017). *Marktanteile der Suchmaschinen weltweit nach mobiler und stationärer Nutzung im März 2017*. In *Statista - Das Statistik-Portal*.

Bibliography

- URL: <https://de.statista.com/statistik/daten/studie/222849/umfrage/marktanteile-der-suchmaschinen-weltweit/> (visited on 04/13/2017) (cit. on pp. 2, 5, 6).
- Ran, Xie (2016). *Pure CSS implementation of Google Photos / 500px image layout*. URL: <https://github.com/xieranmaya/blog/issues/6> (visited on 12/01/2017) (cit. on p. 68).
- Schreck, Tobias (2016). *Slides from Lecture: Digital Libraries VU, Part 9* (cit. on p. 63).
- Schwartz, Barry (2017). *Google's 'people also ask' search feature officially gains dynamic loading*. URL: <http://searchengineland.com/googles-people-also-ask-search-feature-officially-gains-dynamic-loading-269299> (visited on 07/16/2017) (cit. on p. 23).
- Searchmetrics (2016). *Rebooting Ranking-Faktoren: Google.de*. URL: <http://www.searchmetrics.com/de/knowledge-base/ranking-faktoren> (visited on 06/08/2017) (cit. on pp. 13–15).
- Searchmetrics (2017). *General Ranking Factors: Rank Correlations Top20*. URL: <http://www.searchmetrics.com/wp-content/uploads/Searchmetrics-Ranking-Factors-Infographic-EN-Print.pdf> (visited on 06/09/2017) (cit. on p. 14).
- Singhal, Patrick (2012). *Introducing the Knowledge Graph: things, not strings*. URL: <https://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html> (visited on 06/18/2017) (cit. on pp. 28, 33, 39).
- Sitemaps.org (2008). *What are Sitemaps?* URL: <https://www.sitemaps.org> (visited on 05/29/2017) (cit. on p. 7).
- Slegg, Jennifer (2015). *How Google Uses Clicks in Search Results, According to Google*. URL: <http://www.thesempost.com/how-google-uses-clicks-in-search-results-according-to-google/> (visited on 06/11/2017) (cit. on p. 14).
- Sullivan, Danny (2016). *FAQ: All about the Google RankBrain algorithm*. URL: <http://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440> (visited on 06/14/2017) (cit. on p. 16).
- Sweeny, Marianne (2011). *How many Google searchers go to page two of their search results?* URL: <https://www.quora.com/How-many-Google-searchers-go-to-page-two-of-their-search-results> (visited on 07/14/2017) (cit. on p. 14).

Bibliography

- Wall, Aaron (2017). *Search Engine History.com*. URL: <http://www.searchenginehistory.com> (visited on 01/02/2018) (cit. on p. 1).
- Yahoo (2017). *Submit your website to Yahoo Search*. URL: <https://help.yahoo.com/kb/search/SLN2217.html?impressions=true> (visited on 05/29/2017) (cit. on p. 7).
- Zeifman, Igal (2017). *Bot Traffic Report 2016. Imperva Incapsula*. URL: <https://www.incapsula.com/blog/bot-traffic-report-2016.html> (visited on 04/13/2017) (cit. on p. 6).