

# Web searching, search engines and Information Retrieval

**Dirk Lewandowski**

*Department of Information Science, Heinrich-Heine-University Düsseldorf, Universitätsstraße 1, D - 40225 Düsseldorf, Germany.*

*E-mail: [dirk.lewandowski@uni-duesseldorf.de](mailto:dirk.lewandowski@uni-duesseldorf.de)*

To appear in: [Information Services & Use 25\(2005\)3](#)

**This article discusses Web search engines; mainly the challenges in indexing the World Wide Web, the user behaviour, and the ranking factors used by these engines. Ranking factors are divided into query-dependent and query-independent factors, the latter of which have become more and more important within recent years. The possibilities of these factors are limited, mainly of those that are based on the widely used link popularity measures. The article concludes with an overview of factors that should be considered to determine the quality of Web search engines.**

## 1. Introduction

“Ten Years Into the Web, and the Search Problem is Nowhere Near Solved”, was the title of a talk given by David Seuss, founder of the Northern Light search engine, at the Computers in Libraries conference in 2004 [26]. This title sounds rather pessimistic considering the great advances search engines made from early developments such as Webcrawler to the modern engines we all use, such as Google, Yahoo and MSN.

It is interesting to see that the search engine market is shared by just a few companies and dominated by just one, namely Google. But during the last years we saw the other mentioned companies catching up with the fusion of already developed technology (in the case of Yahoo, which bought Inktomi, All the Web and Alta Vista) or with the development of their own search technology (in the case of Microsoft). Apart from the three big players, there are just a few companies mainly with regional impact, such as Ask Jeeves in the US or Seekport in Germany. Many other search sites, first of all the big portals such as AOL, do not provide their own search engines, but use services provided by one of the big three instead. A good and currently updated source that reveals/shows whose search results are provided by which search sites is the Search Engine Relationship Chart [5].

In this article, the problems search engines face in indexing the Web, their solutions to these problems, and the behaviour of their users are discussed. The main goal is to describe the changes from traditional Information Retrieval to Web Information Retrieval, which adds some major problems to

Information Retrieval. Search engines are the most popular implementation of Information Retrieval techniques into systems used by millions of people every day.

## **2. Challenges in indexing the World Wide Web**

An ideal search engine would give a complete and comprehensive representation of the Web. Unfortunately, such a search engine does not exist. There are technical and economical factors that prevent these engines from indexing the whole web every day. On the economic side, it is very expensive to crawl the whole Web. Such a challenge can only be met with the use of server farms consisting of hundreds if not thousands of computers.

On the technical side, the challenge starts with finding all the relevant documents in an environment where no one knows how large it is. Therefore, it is difficult to measure the part of the Web that a certain search engines covers.

### **2.1. Size of the databases, Web coverage**

Search engine sizes are often compared by their self-reported numbers. Google claims to have indexed approx. 8 billion documents and Yahoo claims that its total index size is 19 billion Web documents, which seems to be highly exaggerated. Estimates show that this engine has indexed approx. 5-7 billion documents, while competitor MSN – which does not report numbers – lies between 4 and 5 billion [17].

Some studies tried to measure the exact index sizes of the search engines [21] and their coverage of the indexable Web [13, 14]. They suggest that the data published by the search engines is usually reliable, and some indices are even bigger than the engines claim.

To determine the Web coverage of search engines, one has first to discover how large the Web actually is. This is very problematic, since there is no central directory of all Web pages. The only possibility is to estimate the size based on a representative sample. A recent study [8] found that the indexable Web contains at least 11.5 billion pages, not including the Invisible Web (discussed in section 2.3).

Another important fact is that search engines *should not* index the entire Web. An ideal search engine should know all the pages of the Web, but there are contents such as duplicates or spam pages (see section 2.5) that should not be indexed. So the size of its index alone is not a good indicator for the overall quality of a search engine. But it seems the only factor to compare the competitors easily.

### **2.2. Up-to-dateness of search engines' databases**

Search engines should not only focus on the sizes of their indices, but also on their up-to-dateness. The contents on the Web change very fast [cf. 23] and therefore, new or updated pages should be indexed as fast as possible. Search engines face problems in keeping up to date with the entire Web, and

because of its enormous size and the different update cycles of individual websites, adequate crawling strategies are needed.

An older study [22] found that the up-to-dateness of current Web pages in the search engines' indices ranges widely. The big search engines MSN, HotBot, Google, AlltheWeb, and AltaVista all had some pages in their databases that were current or one day old. The newest pages in the databases of the smaller engines Gigablast, Teoma, and Wisenut were pages that were quite older, at least 40 days. When looking for the oldest pages, results differed a lot more and ranged from 51 days (MSN and HotBot) to 599 days (AlltheWeb). This shows that a regular update cycle of 30 days, as usually assumed for all the engines, is not used. All tested search engines had older pages in their databases. In a recent study by Lewandowski, Wahlig and Meyer-Bautor [18], the three big search engines Google, MSN and Yahoo are analysed. The question is whether they are able to index current contents on a daily basis. 38 daily updated web sites are observed within a period of six weeks. Findings include that none of the engines is able to keep current with all pages analysed, that Google achieves the best overall results and only MSN is able to update all pages within a time-span of less than 20 days. Both other engines have outliers that are quite older.

### **2.3. Web content**

Web documents differ significantly from documents in traditional information systems (see table 1). On the Web, documents are written in many different languages, whilst other information systems usually cover only one or a few selected languages. Documents are indexed using a controlled vocabulary, which allows it to search for documents written in different languages with just one query. Another difference is the use of many different file types on the Web. Search engines today not only index documents written in HTML, but also PDF, Word, or other Office files. Each file format provides certain difficulties for the search engines. In the overall ranking, all file formats have to be considered. There are some characteristics, which often coincide with certain file formats, such as the length of PDF files, which are often longer than documents written in HTML. The length of documents on the Web varies from just a few words to very long documents. This has to be considered in the rankings.

Another problem is the documents structure. HTML and other typical Web-documents are just vaguely structured. There is no field structure similar to traditional information systems, which makes it a lot more difficult to allow for exact search queries.

### **2.4. The Invisible Web**

The Invisible Web [28] is defined as the part of the Web that search engines do not index. This may be due to technical reasons or barriers made by website owners, e.g. password protection or robots exclusions. The Invisible Web is an interesting part of the Web because of its size and its data, which is often of high quality. Sherman and Price [28, p. 57] say the Invisible Web consists of “text pages,

files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages.”

Surely the most interesting part of the Invisible Web are databases that are available via the Web, many of which can be used free of charge. Search engines can index the search forms of these databases but are not able to get beyond them. The content of the databases itself remains invisible for the search engines. But in many cases, databases offer a large amount of quality information.

Commercial database vendors such as Lexis-Nexis are omitted because they protect their contents, which are only available for paying customers. But other databases can be used for free. For example, the databases of the United States Patent and Trademark Office (like many other patent databases) contain millions of patents and patent applications in full text, but search engines are not able to index these valuable contents.

There are different solutions for this. One is to integrate the most important of these databases manually. Google, for example, does this for patent data, but only when one searches for a patent number. Above the regular hits, Google displays a link to the USPTO database. Another solution is a kind of meta search engine that integrates not only regular Web search engines but also Invisible Web databases (e.g. <turbo10.com>, cf. [9]). Finally, another solution comes from the Webmasters themselves: They convert their databases to regular HTML pages. A well-known example for this is the Amazon website. Originally a large database of books, each database record is converted into HTML and can be found in the search engine indices.

Today, it is unclear to what extent this method is used and if the Invisible Web is still such a big problem as it used to be some years ago. Regarding the size of the Invisible Web, it is surely quite smaller than proposed by Bergman in 2001 [2]. He said the Invisible Web was 400 to 500 times larger than the Surface Web, but his calculations were based on some of the largest Invisible Web databases, which included sites such as the National Climate Data Center (NOAA) and NASA EOSDIS, both of which are databases of satellite images of the earth. For each picture included, its size in kilobytes was added. As a result, Bergman concludes that the NOAA database contains 30 times more data than Lexis-Nexis, which is a mere textual database. But this says nothing about the amount of information. In conclusion, Bergman’s figures seem highly overestimated. Other authors assume that the Invisible Web is 20 to 50 times larger than the Surface Web [27, 32].

## **2.5. Spam**

Everyone knows that spam is a problem from his or her own e-mail account. Like with e-mail accounts, spammers try to flood search engine indices with their contents. It is very important for search engines to filter these pages to keep their indices clean and keep a good quality of their results. There are two different kinds of spam pages: First, there is the “classic” spam, which are pages that are created with a commercial purpose and do not offer valuable content to the user. These pages are just

built to sell something, sometimes they claim to have information on a certain subject, but do not offer any information on this, but instead use the keywords searched for in their text and offer something completely different.

The main techniques for classic search engine spamming are the use of misleading keywords, keyword stuffing in the text, extensible building of link structures to pretend an importance of certain pages, and the creation of doorway pages (pages that are optimised for search engines and lead to the actual content page).

But there is also an increasing spam problem with duplicate contents from free sources such as Wikipedia, Open Directory and many product catalogues from online retailers. Webmasters copy these contents (and often add advertisements to them) and use them on their own websites. They earn money from the ads or from commission on sales to people coming from their pages.

There are many attempts to fight spam using various methods [e.g. 6, 36, 9, 7], but as spam is a very large problem for the search engines, the spammers always seem faster and as a search engine user, everyone can tell that the problem is hardly solved.

It would be interesting to know how large the fraction of spam pages is in the indices of the popular search engines, but unfortunately there are no recent studies discussing this point.

### **3. How users search the Web**

The users of Web search engines are very heterogeneous and the engines are used by laypersons, as well as by information professionals or experts in certain fields. Apart from studies discussing the common user behaviour, there are some studies that discuss the behaviour of certain user groups (see [31], pp. 21-25 for an overview). But there are no scientific investigations that discuss how real information professionals in intelligence departments or management consultancies use Web search engines. Instead, most user studies focus on the typical lay user.

The main findings of these studies are that the users are not very sophisticated. Only half of the users know about Boolean operators [20] and only slightly more (59 percent) know about advanced search forms. But knowing them does not mean that they are used: Only 14 percent say that they use them. In a laboratory test in the same study, the use of the advanced search forms was even lower.

In studies based on transaction log analysis, Spink and Jansen [31] found that Boolean operators are only used in one out of ten queries. Half of the Boolean queries are ill-formed [11]; when plus and minus signs are used (which is generally preferred by the users), the fraction of ill-formed queries rises to two thirds.

Users look only seldom at results coming after the first search results page, which means that results which are not among the top 10 are nearly invisible for the general user [31]. There is a tendency that users often only look at the results set that can be seen without scrolling [29].

Within one search session, users look at five documents on average [31, p. 101] and each document is only shortly examined. Sessions are usually terminated when one suitable document is found. A typical search session lasts less than 15 minutes.

## **4. How search engines rank documents**

### **4.1. Generations of search engines**

While early search engines such as Alta Vista mainly relied on techniques from traditional information retrieval, it was soon to be seen that this does not fit for indexing the Web. With these search engines, it was easy for webmasters to manipulate the rankings of the engines by changing the contents of their webpages. A person who wanted his webpage to rank first for a search term just had to repeat the word in the document very often. The search engines soon used techniques to find such manipulations, but with simply content-based approaches they failed and the quality of search engine results soon was very poor.

A second generation of search engines came with the advent of Google. These search engines used link-bases approaches to determine the quality of documents, which are described in section 4.2. These make it a lot more difficult to manipulate the search engines' rankings, but it is far from impossible. A main goal for today's search engines is to keep their indices clean from pages built only to manipulate the rankings (see section 2.5).

Today, it is unclear if the next generation of search engines will be more resistant to spamming attempts, due to their use of personalised ranking techniques or other more user centred approaches.

### **4.2. Ranking factors**

Search engines use two different kinds of ranking factors: query-dependent factors and query-independent factors (for an extensive discussion see [16]. Query-dependent are all ranking factors that are specific to a given query, while query-independent factors are attached to the documents, regardless of a given query.

Table 2 shows the query-dependent factors used by search engines. On the one hand, these are measures such as word documents frequency, the position of the query terms within the document or the inverted document frequency, which are all measures that are used in traditional Information Retrieval. On the other hand, there are measures such as an emphasis on anchor text, the language of the document in relation to the language of the query or the measuring of the "geographical distance between the user and the document". These are added to the classic IR measures, but they also focus on finding the most relevant documents to a given query mainly by comparing queries and documents. The second group of measures used by search engines are query-independent factors that are used to determine the quality of a given document. Such measures are necessary because there is a wide range

from low quality to high-quality documents on the Web. Search engines should provide the user with the highest possible quality and should omit low-quality documents.

Query-independent factors (see table 3) are used to determine the quality of documents regardless of a certain query. The most popular of these factors is PageRank [24], which is a measure of link popularity used by the search engine Google. While early approaches to link popularity just counted the number of in-links to a page, PageRank and other link based ranking measures take into account the link popularity of the linking pages or try to measure link popularity within a set of pages relevant to a given query. Some search engines also count the number of clicks a document gets from the results pages and thereby count a measure of click popularity.

Another query-independent factor considered by some search engines is the directory level of a given document, whereby documents on a higher (or a certain) level in the hierarchy are preferred. The document length can also be a ranking factor, because it is known that users prefer short documents in general, but not too short documents that consist of just a few words. Also, the size of the website hosting the document can be used as a ranking factor. Here it is assumed that it is more likely that a document on a larger website is authoritative than another on a small website.

A very important factor is the up-to-dateness of a document: For some queries, newer documents are more important than older ones. Even though this assumption provides some problems in general, the age of a document should be considered as a ranking factor. In some cases, older documents should be preferred. An overview on how one could use date-information for ranking is given in [1].

Finally, even the filetype can be used as a ranking factor as well. Usually, search engines prefer regular HTML documents over PDF or Word files because the user can see these files in his browser without opening another program or plug-in.

### **4.3. Problems with link-based ranking algorithms**

Link-based ranking algorithms are dominant in today's search engines and it is often forgotten that these approaches face some difficulties and provide some kind of bias in the results. Here, the most important bias factors of these algorithms are described (for an extensive discussion see [16]).

Therefore, not the dominant algorithms themselves are criticised, but some of their basic assumptions. Firstly, they are based on a certain quality model. Quality is equated with authority (a notion used by Kleinberg in his seminal paper [12]) or (link) popularity. Other quality factors are disregarded and the algorithms are solely based on an improved quality model as used in citation indexing. The reason for this lies mainly in the link structure of the Web, which can be exploited relatively easy, the regress on well-established bibliometric methods and the plausibility of the basic assumption.

In link-based ranking algorithms, every link is counted as a vote for the linked page. But there are several reasons for linking to a certain page, so links cannot be seen as analogous to citing literature [29]. Some links are just put for navigational purposes, some are indeed pointing to content, but they

are used as a deterring example. Link-based ranking algorithms cannot differentiate between these and links pointing to good content.

Other links are placed out of favour or for promotional purposes. There is no strict border between “good” link exchange and manipulation, and therefore it is difficult for search engines to find links that should not be counted.

Each link within a document is usually counted the same, regardless of its position within the document [4, p. 219]. But the position of the link is important for the user. It is more likely that he will click on the link that is prominently placed.

There is also some bias in link counting by the search engines. The most important anomalies are site selflinks, replicated links, interlinked databases and mirror sites [33, p. 26]. Site selflink (links that point from a page of a certain website to another page of the same website) are not differentiated from external links, while replicated links and interlinked databases are not created by humans but automatically by machines and should therefore not be equal in weight as “real” links. Links that are replicated on mirror sites are counted more often than links from sites that are not mirrored.

In addition, some pages are preferred when placing a link. These are pages that are already visible in the search engines. Because of their visibility, they have a higher chance to get additional links, which is called preferential attachment.

Link-based ranking algorithms are very good for navigational queries, in which a user searches for a certain homepage, but not – as in informational queries – for an amount of documents. In a study it was shown that link-based algorithms only perform better in navigational, but not in informational queries [25] (for a distinction of certain kinds of queries see [3]). It should be investigated which algorithms suits best for navigational, informational and transactional queries.

In conclusion, link-based algorithms helped to improve the quality of search engine rankings to a high degree, but there is also an immanent kind of bias that is not considered enough yet.

## **5. Measuring the quality of Web search engines**

In this article, it was shown that there are many factors that together determine the quality of a Web search engine. But usually, the quality of information retrieval systems in general and search engines in particular is measured only with retrieval tests. These take into account standard measures like recall and precision but omit other factors that are not relevant in traditional information retrieval. To consider the specific characteristics of Web information retrieval, apart from the standard measures, tests should also take into account the index quality, the search features (which vary strongly, cf. [15], the retrieval system and the user behaviour.

The index quality of a certain search engine is a combination of the size of the database, its up-to-dateness, the indexing depth, and hopefully low indexing bias, e.g. bias in the coverage of documents from different countries [cf. 35]. It should also be kept in mind that search engines offer additional

databases, e.g. for pictures, audio files, and special news databases [19]. These special collections are valuable additions to the Web document.

Advanced search features are often regarded as not so important because only a relatively low fraction of users avail them [20, p. 168]. But for the professional use of the search engines, they are indispensable and should therefore be taken into account when discussing the quality of search engines.

Surely, the retrieval system as the core of each engine should be tested in studies discussing quality. In addition to traditional retrieval measures, extended measures specifically for search engines [e.g. 34] should be developed and used.

And last, the user behaviour should be the centre of attention of search engine quality studies.

Although there have been some studies on this topic [e.g. 31, 20], research should be extended, because although we know a lot about the general user, we do not know much about certain user groups, such as how information professionals or members of a certain occupational group use – or would like to use – search engines.

## References

1. Acharya, A.; Cutts, M.; Dean, J.; Haahr, P.; Henzinger, M.; Hoelzle, U.; Lawrence, S.; Pflieger, K.; Sercinoglu, O.; Tong, S. (2005): Information retrieval based on historical data. Patent Application US 2005/0071741 A1 (published: 31.3.2005)
2. Bergman, M. K. (2001): The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* 7(1). <http://www.press.umich.edu/jep/07-01/bergman.html> [22.8.2005]
3. Broder, A. (2002): A taxonomy of web search. *SIGIR Forum* 36(2). <http://www.acm.org/sigir/forum/F2002/broder.pdf> [22.8.2005]
4. Chakrabarti, S. (2003): *Mining the Web: Discovering Knowledge from Hypertext Data*. Amsterdam (u.a.): Morgan Kaufmann
5. Clay, B. (2004): Search Engine Relationship Chart. <http://www.bruceclay.com/searchenginechart.pdf> [22.8.2005]
6. Fetterley, D.; Manasse, M.; Najork, M.: Spam, Damn Spam, and Statistics. *Seventh International Workshop on the Web and Databases (WebDB 2004)*, June 17-18, 2004, Paris, France, pp. 1-6
7. Gee, K.R.: Using Latent Semantic Indexing to Filter Spam. *Proceedings of SAC 2003*, Florida, USA. pp. 460-464
8. Gulli, A.; Signorini, A. (2005): The Indexable Web is More than 11.5 billion pages. *Proceedings of the Special interest tracks and posters of the 14th international conference on World Wide Web*, May 10-14, 2005, Chiba, Japan. pp. 902-903

9. Gyögyi, Z.; Garcia-Molina, H.; Pedersen, J.: Combating Spam with TrustRank. Proceedings of the 30<sup>th</sup> VLDB Conference, Toronto, Canada, 2004, pp. 576-587
10. Hamilton, N. (2003): The Mechanics of a Deep Net Metasearch Engine. <http://turbo10.com/papers/deepnet.pdf> [22.8.2005]
11. Jansen, B. J.; Spink, A.; Saracevic, T. (2000): Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management* 36(2), pp. 207-227
12. Kleinberg, J. (1999): Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM* 46(5), pp. 604-632
13. Lawrence, S., Giles, C. L. (1998): Searching the World Wide Web. *Science* 280, pp. 98-100
14. Lawrence, S., Giles, C. L. (1999): Accessibility of information on the web. *Nature* 400(8), pp. 107-109
15. Lewandowski, D. (2004): Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen. *Information: Wissenschaft und Praxis* 55(2), pp. 97-102
16. Lewandowski, D. (2005): Web Information Retrieval. Frankfurt am Main, DGI, 2005
17. Lewandowski, D. (2005): Yahoo - Zweifel an den Angaben zur Indexgröße, Suche in mehreren Sprachen. *Password* 20(9) [to appear]
18. Lewandowski, D.; Wahlig, H.; Meyer-Bautor, G.: The Freshness of Web Search Engines' Databases. [to appear]
19. Machill, M.; Lewandowski, D.; Karzauninkat, S. (2005): Journalistische Aktualität im Internet. Ein Experiment mit den News-Suchfunktionen von Suchmaschinen. In: Machill, M.; Schneider, N. (Hrsg.): *Suchmaschinen: Herausforderung für die Medienpolitik*. Berlin: Vistas 2005, pp. 105-164
20. Machill, M.; Neuberger, C.; Schweiger, W.; Wirth, W. (2003): Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: Machill, M.; Welp, C. (Hrsg.): *Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen*. Gütersloh: Verlag Bertelsmann Stiftung, pp. 13-490
21. Notess, G. (2003): Search Engine Statistics: Database Total Size Estimates. <http://www.searchengineshowdown.com/stats/sizeest.shtml> [7.7.2005]
22. Notess, G. (2003): Search Engine Statistics: Freshness Showdown. <http://www.searchengineshowdown.com/stats/freshness.shtml> [7.7.2005]
23. Ntoulas, A.; Cho, J.; Olston, C. (2004): What's New on the Web? The Evolution of the Web from a Search Engine Perspective. Proceedings of the Thirteenth WWW Conference, New York, USA. [http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas\\_new.pdf](http://oak.cs.ucla.edu/~ntoulas/pubs/ntoulas_new.pdf) [22.8.2005]
24. Page, L., Brin, S., Motwani, R., Winograd, T. (1998): The PageRank citation ranking: Bringing order to the Web. <http://dbpubs.stanford.edu:8090/pub/1999-66> [22.8.2005]

25. Savoy, J.; Rasolofo, Y. (2001): Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. <http://trec.nist.gov/pubs/trec9/papers/unine9.pdf> [22.8.2005]
26. Seuss, D. (2004): Ten Years Into the Web, and the Search Problem is Nowhere Near Solved. Computers In Libraries Conference, March 10-12, 2004. <http://www.infotoday.com/cil2004/presentations/seuss.pps> [22.8.2005]
27. Sherman, C. (2001): Search for the Invisible Web. Guardian Unlimited 6.9.2001. <http://www.guardian.co.uk/online/story/0,3605,547140,00.html> [22.8.2005]
28. Sherman, C.; Price, G. (2001): The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today
29. Singhal, Amit (2004): Challenges in Running a Commercial Search Engine. <http://www.research.ibm.com/haifa/Workshops/searchandcollaboration2004/papers/haifa.pdf> [22.8.2005]
30. Smith, A. G. (2004): Web links as analogues of citations. Information Research 9(4). <http://informationr.net/ir/9-4/paper188.html> [22.8.2005]
31. Spink, A.; Jansen, B. J. (2004): Web Search: Public Searching of the Web. Dordrecht: Kluwer Academic Publishers
32. Stock, W. G. (2003): Weltregionen des Internet: Digitale Informationen im WWW und via WWW. Password Nr. 18(2), pp. 26-28
33. Thelwall, M. (2004): Link Analysis: An Information Science Approach. Amsterdam [u.a.]: Elsevier Academic Press
34. Vaughan, L. (2004): New measurements for search engine evaluation proposed and tested. In: Information Processing and Management 40(4), pp. 677-691
35. Vaughan, L.; Thelwall, M. (2004): Search Engine Coverage Bias: Evidence and Possible Causes. Information Processing & Management, 40(4), pp. 693-707
36. Wu, B.; Davison, B.D.: Identifying Link Farm Spam Pages. Proceedings of WWW 2005, May 10-14, Chiba, Japan, pp. 820-829

**Table 1: Differences between Web Information Retrieval and traditional Information Retrieval**

Differentiator	Web IR	Traditional IR
<b><i>Documents</i></b>		
Languages	Documents in many different languages. Usually search engines use full text indexing; no additional subject analysis.	Databases usually cover only one language or indexing of documents written in different languages with the same vocabulary.
File types	Several file types, some hard to index because of a lack of textual information.	Usually all indexed documents have the same format (e.g. PDF) or only bibliographic information is provided.
Document length	Wide range from very short to very long. Longer documents are often divided into parts.	Document length varies, but not to such a high degree as with the Web documents. Each indexed text is represented with one documentary unit.
Document structure	HTML documents are semi-structures.	Structured documents allow complex field searching.
Spam	Search engines have to decide which documents are suitable for indexing.	Suitable document types are defined in the process of database design.
Hyperlinks	Documents are connected heavily. Hyperlink structure can be used to determine quality.	Documents are usually not connected. Sometimes citation data is used to determine quality.
<b><i>Web characteristics</i></b>		
Amount of data, size of databases	The actual size of the Web is unknown. Complete indexing of the whole Web is impossible.	Exact amount of data can be determined when using formal criteria.
Coverage	Unknown, only estimates are possible.	Complete coverage according to the defined sources.
Duplicates	Many documents exist in many	Duplicates are singled out

	copies or versions.	whilst documents are put into the database. No versioning problems because there is usually a final version for each document.
<b><i>User behaviour</i></b>		
User interests	Very heterogeneous interest.	Clearly defined user group with known information seeking behaviour.
Type of queries	Users have little knowledge how to search; very short queries (2-3 words).	Users know the retrieval language; longer, exact queries.
<b><i>IR system</i></b>		
User interface	Easy to use interfaces suitable for laypersons.	Normally complex interfaces; practice needed to conduct searches.
Ranking	Due to the large amount of hits relevance ranking is the norm.	Relevance ranking is often not needed because the users know how to constrain the amount of hits.
Search functions	Limited possibilities.	Complex query languages allow narrowing searches.

**Table 2: Query-dependent ranking factors**

Word document frequency	Counting the relative frequency of a query term in a document.
Search term distance	
Search term order	
Position of the query terms	Documents in which the search terms appear in prominent places such as in the title or in headings are preferred.
Metatags	Search terms appear in meta information such as keywords or description.
Position of the search terms within the document	If the terms appear at the beginning of the documents, this document is seen as more important than others.
Emphasis on terms within the document	Terms that are emphasised (e.g. with HTML tags like <b> or <i>) are regarded as more important than terms in regular expression.
Inverted document frequency (IDF)	Counting the relative frequency of a term in all documents; rarely occurring terms are preferred.
Anchor text	Query terms appearing in anchor text are counted higher.
Language	Documents written in the same language as the used user interface are preferred.
Geo targeting	Pages that are “closer” to the location of the user are preferred.

**Table 3: Query-independent ranking factors**

Directory hierarchy	Documents on a higher hierarchy level are preferred.
Number of incoming links	The higher the number of incoming links, the more important the document.
Link popularity	Quality/authority of a document is measured according to its linking within the Web graph.
Click popularity	Documents visited by many users are preferred.
Up-to-dateness	Current documents are preferred to older documents.
Document length	Documents within a sudden length range are preferred.
File format	Documents written in standard HTML are preferred to documents in other formats such as PDF or DOC.
Size of the Website	Documents from larger Web sites (or within a sudden size range) are preferred.