From Unreliable Web Search to Information Provisioning based on Curated Data

Florian Stahl, Gottfried Vossen

European Research Center for Information Systems (ERCIS)

University of Münster, Germany

July 2012

Abstract:

The volume of data stored electronically is now superabundant. Accessible data on the Web, whether created by computers, by Web users, or generated within professional organizations, are growing at a tremendous pace. Social networks, search engines, and e-commerce sites generate and store new data in the terabyte range on a daily basis and with the growth of cloud computing, this trend will continue. Yet data is not only generated and permanently stored online, it is also linked to other data, and aggregated in order to form new data. The question of how the most appropriate information required for a specific purpose at a given moment can be sourced has for the past two decades been answered by search: Data underlying the information we need is determined using a search engine. This paper describes parts of the evolution of search engines.

However, a point has been reached where search is no longer good enough in many situations. Based on the historic development of search engines the idea has emerged that selecting and integrating quality-checked (i.e., curated) data can potential improve Web search and information quality. Some effort has been made in that direction, but many open questions remain, which we indicate at the end of the paper.

1 Introduction

Today, we are living in an information society [LM06]. The fact that for modern enterprises information and the quality thereof is of higher importance than ever before [OLC11] is a clear indicator for this. Information technology has penetrated most business areas and has a significant impact on the quality of products and services. Furthermore, not finding information can – despite being annoying to users – be costly for enterprises [FS01]. As Tim O`Reilly, one of the intellectual fathers of the “Web 2.0”, has observed, "*when hardware became commoditized, software was valuable. Now that software is being commoditized, data is valuable*.”[MO07].

While hardware and software were initially (i.e., 30 to 40 years ago) valuable because of their limited availability (scarcity), the case of data is a different one: Even though data available on the Web – created by Web users, computers, or professional organisations – are growing at an enormous speed [7], it is becoming increasingly difficult to satisfy the increasingly complex information needs. This is because if answering one sort of queries becomes possible a plenitude of new queries arises [19]. From this it follows that with information the issues is not scarcity, but superfluity or abundance, i.e., the problem of not knowing the relevant data which are able to satisfy the complex information needs of a given situation.

Since the early days of the Web in the 1990s, people have been striving to maximise their utility of the information available through it, and the tool of choice has become the search engine [MV98]. Moreover, methods from Information Retrieval (IR) have been applied to retrieving information effectively and efficiently from the Web. IR has been established as a field in the early sixties; however, it has only become of broader interest with the emergence of the Web establishing the discipline of Web Information Retrieval (Web IR) [LM06], [Dop09], [Lev10], [BR10], [BH11]. However, unlike in classical IR, where mainly small homogeneous data sources are involved, data sources on the Web are more unstructured and heterogeneous [PBMW99]. Following the emergence of the Web both analysis and retrieval of data and information (data with an interpretation) from the Web have been researched extensively over the last years [BR10, BR11]. Moreover, a shift from document retrieval towards satisfying information needs can be observed. Baeza-Yates et al. describe the problem as: "People do not really want to search, they want to get tasks done" [BR10]. Similarly, in library and information science a focus shift from collections to users and their needs can be observed [SC11].

With the vast amounts of data nowadays present, a point has been reached where search engines alone may no longer be adequate in many situations. The goal of this paper is to argue why this is the case, and to indicate improvements. To this end, a literature review has been conducted, surveying historic as well as current developments and ideas in search as well as Web IR, and outlining advantages, disadvantages, and reasoning behind past and current methods to satisfy information needs. It will be shown that throughout the lifetime of the Web new ways of finding and retrieving information became necessary in order to cope with its exponential growth, as well as with the growing information needs. Both have not stopped yet to grow; together with observation of Buguanza et al. [BD10] that the search engine industry is moving and the current dominant design is being challenged this indicates that a new approach can have significant impact on the way information is retrieved and used in the near future.

The remainder of this paper is organised as follows: Section 2 gives an overview of Web search concepts, particularly highlighting the problems *mining the Deep Web* and *judging relevance*. In Section 3 the limitations of actual search technology are highlighted, before Section 4 concludes this paper by outlining research areas that should be developed in order to provide better information services.

2 Web Search Concepts

Right from the start of the Web, technical means to retrieve information became necessary because of its decentralised, heterogeneous nature and exponential growth. First attempts in the pre-Web-based Internet were Gopher, Archie and Wide Area Information Servers (WAIS) [Gil93], [Com95]. Then Tim Berners-Lee proposed a project [9] that later built the basis for the Web, thereby enabling better information interconnection and access. From 1991 on the Web grew rapidly [4] and at the same time the first Web directories (lists of hierarchically sorted hyperlinks) started to appear with the intention of serving as entry points to the Web. These catalogues were mainly concerned with different research areas and can be seen as the application of cataloguing techniques to the Web (examples are [3] and [5]). With the commercialisation of the Web private companies also offered professionally maintained Web directories; one of the most prominent examples is Yahoo!, founded in 1994 as *"Jerry and Davids Guide to the World Wide Web"* [6].

A key characteristic of a Web directory is that it is maintained manually, which means that humans analyse the content of Websites and summarise it in order to categorise it into a hierarchy[GBR09], [BH11]. Theoretically, this should secure quality and lead to an intuitive structure. Using a Web directory, visitors may either browse through the aforementioned hierarchy or utilise a tool that searches thorough titles and descriptions of Websites. Despite these advantages, directories have two obvious drawbacks: maintenance effort and scalability, which are not able to cope with the growing Web [GBR09], [BH11]. This was correctly recognised by Google founders Brin and Page, who came up with the idea of (dynamically) *searching* the Web based on an automatically created and maintained document index [Ba05].

The first Web crawlers – autonomous programs that follow hyperlinks to automatically retrieve new and updated Websites – and Web search engines based on them were built in the early 1990's [MSHP09]. The architecture of search engines can be divided into three parts. Firstly, crawlers "browse" the Web. Secondly, the retrieved Websites are temporarily stored and analysed. In this step keywords are extracted and the search index, which basically links keywords to Web documents, is updated. This process runs infinitely and independent of user queries in the background. The third and last part of a search engine is the user interface or runtime system, which is only invoked if a user poses a query to the search engine; queries are translated into machine-understandable terms and relevant entries from the index are retrieved and ranked. [VH07]

Accordingly, research has been conducted in all three areas. In the area of crawling it is investigated how sources that are hard to access can be tapped, also referred to as *Deep Web Mining*. In regard to analysing it is mainly investigated how non-textual sources can be accessed an evaluated. In terms of presenting results to users it is examined what a user's intentions are and hence in what order results should be presented. Whilst improvements in analysing and indexing of non-textual content are desirable they have smaller influence on the provisioning of high quality information, which – to date – is mainly considered to be textual facts. Therefore, only accessing new sources and user presentation will be discussed in more detail, following a general description of search concepts.

2.1 Approaches to Search

At a high level, three types of search engines can be distinguished: general purpose, special purpose and archive search engines [Lew09]. Whilst the purpose of the first two is to serve immediate information needs, the purpose of the latter is to conserve Web pages permanently in order to make documents available that are no longer present on the Web. A website that falls into this category is the Wayback Archive[1].

General purpose search engines are search engines for the *average* user providing *average* results, which are not necessarily in-depth information and hardly personalised. If such a *horizontal* search (covering the whole breadth of the Web) is performed, there is also the risk of accidentally excluding specific and important information in the course of reducing results to a manageable size. This is overcome by specialised search engines which can focus more precisely on a specific task. Examples are search for up to date information (e.g., news), search for special document types (e.g., images / PDFs) or searches in a certain domain [Lew09], whereas one domain may consist of several sub domains; for instance search on travel may consist of the domains flights, hotels and car rental [Cer10].

[1] See http://www.archive.org/Web/Web.php

Different (types of) search engines offer different search options and return different results. To combine the strength of various search engines meta search engines have evolved. In contrast to conventional search engines they do not have an index of their own but redirect queries to a number of search engines and present their combined results to users. Some of them even eliminate duplicates and re-order the result to offer a persistent ranking [Cer10]. Depending on their use case meta search engines can theoretically fall into each of the three aforementioned categories.

A big advantage of meta search engines is the broad spectrum that can be covered. However, search functionality might be limited, as only search operators that are supported by all underlying search engines can be used [GBR09]. In a study from 2009 Griesbaum et al. [GBR09] mention several examples of meta search engines, of which at the time of writing this paper only two of six were still online. Interestingly, the two sites working were the people search engines pipl and yasni[2].

Most search engines retrieve text-based information; however, some are concerned with queries regarding multimedia files. Approaches to index these are to analyse related text (meta data or text surrounding, e.g., an image) or to analyse the content itself. [Lew09] The latter – despite being computational more challenging – offers the possibility to enhance results found by annotation analysis and to retrieve multimedia in the absence of annotations. [LSDJ06] Even though, advances are being made research on multimedia information is still active to solve remaining issues in particular regarding semantics and relevance.

The big search engines (such as Google or Bing) commonly offer – besides their general purpose search engine – also special purpose search engines. Taking advantage of having these special results also available they merge them with their general Web search results [14, 16], making them in a sense "universal." Common examples of added specialised contents are news, multimedia (pictures, music, videos) and products, amongst others. Regarding the presentation two forms coexist: Either results are presented separately for each category or results are mixed to present a single result set. In terms of functionality universal search can be compared to meta search engines; a query is redirected to specialised vertical search engines (i.e., search engines focusing on a special domain, or file type) and the results are integrated and displayed. [Qui09], [2, 10]

Other trends in Web search favour more social interaction to enhance search results or data collection. According to Burghardt et al. [BH11] common approaches are social indexing, social question answering, and collaborative filtering and collaborative search. Social indexing describes collaboratively maintained Web directories such as the open directory project and tagging as on video platforms like Youtube or bookmarking services like delicious[3]. The advantage of such systems is that searchers become taggers and therefore a common understanding of terms should exist. In particular for-non texted based content this is a viable solution to make these document types searchable (tag-based search). It also works well in niches such as twitter, however, for search this approach has not yet found broad acceptance. Furthermore, quality cannot be assured in this approach, since taggers are free in their tag choice. Only professional taggers could potentially overcome this. Question and Answer Systems offer users the possibility to pose questions which may be answered by a community or paid experts [AGZ09] [BH11]. If the latter is the case these offers are usually not free of charge. Collaborative filtering mainly describes recommender systems that filter relevant documents based on similar users. Collaborative search can be seen as a prime example of social search for it serves as platform to enable distributed knowledge workers to jointly work on their IR tasks.

[2] See http://www.pipl.com and http://www.yasni.de/, respectively

[3] See http://www.dmoz.org/; http://www.youtube.com/; and http://www.delicious.com respectively

Knowledge-based search systems are delivering answers to queries by *computing* them based on built-in data rather than doing Web search [1]; an example is Wolfram|Alpha[4]. According to their Website Wolfram|Alpha maintains more than 10 trillion (increasing) pieces of data [1], which have been gathered by at least 150 editors prior to the launch of Wolfram|Alpha in 2009 [Gil09]. Along with the collection of data goes an on-going maintenance process with the declared aim of being *"as trustworthy as gold standard sources."* Even though queries can be posed in natural language this is not always working and may annoy users. [Gil09] Also it has to be mentioned that Wolfram|Alpha is superior to conventional search for queries to well-structured data while it is not able to server the whole scope of complex search [Cer10].

2.2 Broadening the Data Basis: Mining and Integrating the Deep Web

As stated before, search engines crawl the Web to retrieve Web pages and, in a second step, analyse their contents in order to index them. This analysis is far from trivial. In order to summarise and categorise retrieved documents, Web Mining – a special form of text mining – can be used. Both Web and text mining try to find patterns in natural language text, which is considerably harder than in structured data. Nevertheless, in Western culture text is the vehicle for information, making the effort worthwhile. Compared to data mining (finding unknown patterns in data), text mining has the clear advantage that the information is not hidden. Beyond that, text on the Web is often enriched with markup that can help to structure the contents.[WFH11]

Despite all crawling, indexing, and retrieval efforts, many Web sources still exist – and may even be growing in number – that currently cannot be accessed in automated ways. This is known since at least 1994 when according to Bergman [Ber01] Ellsworth coined the term "Invisible Web" for non-indexable Websites. Bergman himself rephrased it to *Deep Web* as he considers it to be visible although not accessible. The accessible Web is referred to as *Surface Web*. Similarly Stock [Sto03] divided information on the Web into two categories: information available on the Web (Surface Web) and available through the Web (Deep Web). This Deep Web may be inaccessible because of a) access restrictions, b) because Websites exclude themselves from being indexed, c) because the content is on an "island" in the Web which is not interconnected to the rest of the Web by hyperlinks and can therefore not be found, or d) because of technical restrictions such as database driven designs (i.e. information is a result of user-specific queries to a database) [GBR09]. In particular d) is most commonly understood when using the term Deep Web [CHLP+04], [MKKG+08], [Lew09], [Raj09], [RB09], [XCZY+10].

In a not primarily scientific marketing whitepaper Bergman [Ber01] estimated the size of the Deep Web to be roughly 400 to 550 times larger than the indexable Web, with approximately 200,000 Deep Web sites. In 2004 Chang [CHLP+04] found the number of Deep Web databases to be around 450,000 (with a total of 1.258 million query interfaces) by extrapolating from a sample of 1,000,000 IP addresses to the whole IP space (excluding reserved areas). Researchers of Google stated that there were about 10 million high quality Deep Web forms [MKKG+08] or about one billion pages of structured data [CHM11]. Even though these figures may not be validly compared with one another they indicate the existence and growth of the Deep Web.

Madhavan et al. [MKKG+08] propose three principles to enhance Deep Web access. First, judging the quality of sources; second, crawl only subsets to ensure the load on crawlers does not become too high; third, develop heuristics to discover similarities between data sources as it is not likely that

[4] http://www.wolframalpha.com/

domain specific methods scale to the Web. As further research topics they mentioned discovery of dependencies between form inputs and the ability to deal with JavaScript forms. Regarding judging the quality Xian et al. [XCZY+10] developed an utility maximisation model that can be used to assess the value of Deep Web sources and help to decide which sources to choose from a given list.

In 2011 Carfarella et al. [CHM11] reemphasise Google's goal of making the Deep Web accessible to search engine users and propose two approaches to Deep Web data collection. First, they mention vertical search engines described earlier. Nevertheless, they consider these search engines impractical because a subject area may be hard to define and a significant proportion of human interaction is need to integrate the various sources. Next, they argue that surfacing Deep Web content by means of posing queries and indexing the resulting pages Google was able to index content of "several million" Deep Web databases in a completely automated manner, which they consider superior to any manual approach. Despite this success, however, they confess that there are still major challenges to meet. Some of those are semantic services to improve posing queries to Deep Web databases and gathering data from other sources such as the social Web. Ontological approaches that address semantic problems have been conducted, for instance, by [Jer10] and [LLD11].This shows – leaving quality concerns aside – that retrieval and indexing of documents is no longer a severe problem, and that automatically exploiting the Deep Web is currently being addressed by various researchers (although far from being solved) [BR10]. However, there are more issues to address in order to fully satisfy current information needs. In 2009 Dopichaj [Dop09] stated that at that time it was technically impossible to meaningfully answer queries such as "return all pages that contain product evaluations of fridges by European users;" to a similar conclusion came Ceri [Cer10]. Masermann and Vossen [MV00] have made a proposal in that direction which, however, required considerable programming effort. Whilst Dopichaj advocates the Semantic Web (elaborated on in the next Section), Ceri proposes the so called Search Computing (SeCo) paradigm, which tries to integrate (i. e. combine) Deep Web source, the need for which was also identified by Chang et al [CHLP+04].

Ceri [Cer10] envisions a framework in which a query to a SeCo search engine is processed by a query optimiser which chooses suitable underlying search services for different parts of the query. Results of these search services are joined and displayed to users who are given the chance to modify their queries dynamically, referred to as liquid query processing [BBCF]. The entire framework builds on the creation of two new user groups or communities: data providers offering data services and developers who build search services based on these data services. Data sources in this scenario are collections of data regarding similar domains and may consist of Web Services or scraped Web pages. Search services would integrate data services in a transitive manner.

How a data service (or data mart) can be built using Deep Web mining on a given source is described by Baumgartner et al. [BCGH10]. Their approach called *Lixto* is to create wrappers for Deep Web data sources (and others) by manually training computers to fill out forms and to retrieve data using a graphical user interface. These wrappers can then build the basis for Web services to be included in data marts based on Service Oriented Architecture (SOA) principles. Campi et al. [CCGM+10] describe how service marts can be build and registered with the framework. Building on this, Bozzon et al. suggest an overall architecture for search computing [BBCC+10].

As described above there are two approaches to deep Web access. Rajaraman refers to them as Deep Web crawl (querying deep Web sources and indexing the results) and Federate Search (using APIs to access sources at query time) [Raj09]. His suggestion – implemented in the Kosmix Explore Engine – is using a hybrid approach combining the comprehensiveness of Web search with specificity of federate search.

2.3 Judging Relevance: Retrieving and ranking results from an index

Before results can be retrieved, a query has to be pre-processed to determine a user’s intention, for human language is not (yet) machine understandable. Natural language entails several challenges, in particular synonyms and polysemy [LM06]. This is addressed in various ways, one being semantic search *[Dop09]*. The semantic Web, the idea of which is to enhance text on the Web by semantic information to make it not only machine readable, but machine understandable, was initially proposed by Berners Lee in 2001 [BHL01]. In a personal view, Berners-Lee [23] states that “*the semantic Web is not just about putting data on the web. It is a about making links …*” Links that enable humans and machines to understand relations of data and to discover new data by exploration. As technical means he suggested the resource description framework (RDF)[24] which is a standard intended to facilitate data interchange. It uses unique resource identifiers (URIs) to name the connection between two things as well as the thing themselves.

Until 2009, however, the semantic Web and its enabling technologies were not really being used, most likely because of the maintenance efforts it requires [Dop09]. In June 2011, to increase semantic Web usage Bing, Google, and Yahoo! announced their joint efforts to foster the semantic Web by focusing on one standard. However, apparently this shall only be used for displaying additional information on result pages rather than using it for the actual search [12]. Most recently news spread that Google is planning to introduce semantic Web technologies in their actual search [11]. Nevertheless, it remains to be seen whether the semantic Web will eventually be better supported by Websites and thus offer more possibilities for search based on natural language.

Once a user’s intentions are clear the results from the index can be presented. Early search engines were only able to determine whether a search string was contained in a document or not. Albeit reducing the search space, this left users with potentially thousands of results through which they had to manually browse in order to find the sites most relevant to them. This made it necessary to arrange or order results according to a “guessed” relevance to users [Dop09]. To this end, [BP98] suggested the *PageRank* as a surprisingly accurate and robust ordering criterion. According to Google, their search nowadays incorporates more than 200 factors to judge this relevance [13, 19, 21, 22], which are not published by Google since they are regarded as trade secrets. However, in 2009 Smarty published a community generated list of 120-odd factors (grouped in 14 categories) that are likely to influence Google search. Independently Griesbaum et al. [GBR09] named four categories of factors on which ranking of search results are based on: on-page, on-site, linking and user behavioural factors. Most of the Smarty categories can be fit in to the broader categories defined by Griesbaum et al.; only the groups *penalties* (mainly flags for unwanted behaviour or spamming) and *more factors* (mainly registration with various Google services) were not covered. Since the broader categories cover the most important factors they will be described here in the order in which they were implemented by various search engines.

On-page factors – such as proximity, function, and format of words within a given text [GBR09] – were initial attempts to solve the ranking problem. Regarding function and formatting HTML mark-up was used to determine the importance of words. To judge similarity and proximity vector space models that translate documents into vectors on which mathematical calculations are based can be used [Dop09]. Lycos gained significant market share in the mid nineteen-ninety’s by including the word proximity, i.e., the proximity of (multiple) search terms within a document, in their ranking algorithm [MSHP09]. This simple mechanism, however, contains the risk of spam. In the context of search engines spam refers to Web pages of usually low quality which receive a high ranking through exploiting knowledge of how search engines work. For instance it is simple to generate Web pages containing the same words many times in important functions such as headings [Dop09]. On site factors are more technical factors such as update frequencies, geo location, and architectural factors [21] and have probably been incorporated from an early point on.

A breakthrough in search technology was achieved in 1998 when Google entered the market with a ranking algorithm incorporating linking factors [LM06], [MSHP09], [GBR09], which led to the already mentioned PageRank. This technique does not stem from information retrieval but from an analysis of citation or recommendation behaviour, where it is used to judge the importance of a person or source and to distinguish relevant sources ("authorities") from non-relevant ones [BP98], [VH07], [Dop09]. Today, a number of ranking algorithms building on linking factors exist, of which PageRank [PBMW99] and Hypertext Induced Topic Search (HITS) based on [Kle99] are most well-known [FLMR+06], [LM06], [GBR09]. The underlying idea is that the Web can be seen as a graph (pages being nodes, hyperlinks being edges) in which hyperlinks to other pages can be seen as citations. Therefore, it is reasonable to interpret a link as recommendation. [BP98], [Kle99], [LM06], [GBR09]. Given this assumption, the PageRank algorithm assigns a value (PageRank) to each Web page based on the number of links pointing to it and the PageRank of these. This algorithm was later further improved and additional criteria were taken into account to result in better rankings [VH07].

Whilst PageRank only includes in-links pointing to a Web page, HITS also considers out-links pointing from a Web page. Accordingly, two values are assigned to nodes: a hub and an authority value. A hub is a page citing many others, while an authority is a page that is cited by many others. They are circularly linked in the way that the hub value is calculated by summing the authority values of the out-linked Web pages and the authority value by summing the hub value of the in-linked Web pages. Even though developed at roughly the same time as PageRank, HITS was not intended to be commercialised and was only used from 2001 by a commercial search engine [LM06]. A major difference between PageRank and HITS is the document basis. For HITS it takes potentially very long to execute it, so that it is only applied to a subset of potentially relevant documents [Dop09]. This offers the possibility of a more precise result in terms of answering the query at the cost of potentially leaving important documents out of the scope. SALSA (Stochastic Approach for Link-Structure Analysis) by Lempel et al. [LM00] includes ideas of both of the previous [FLMR+06].

One of the problems with link analysis-based ranking algorithms is their selectivity towards established sites. It is easier for these sites to gain new in-links which increases their ranking value. This effect is even amplified by the fact that highly ranked pages are more likely to gain new in-links [GBR09]. Even though these linking mechanisms are harder to exploit than the simpler on-site-factors it is still possible to create spam when linking algorithms are used, for instance by creating Websites linking to each other with the sole purpose of increasing the ranking. [GBR09], [LM06]. Websites doing this are referred to as *link farms* [Dop09].

Human behavioural factors were relatively recently integrated into search engines. The basic assumption is that if two people pose an identical query their information need is not necessarily the same [Dop09]. At Microsoft Research it was first shown that implicit measures of human interest – such as the number of results that was visited or the time spent on a particular result – can reliably be related to ratings of user satisfaction [FKMD+05]; building on this and other works it was later also shown that considering implicit measures in rankings can improve the ranking [ABD06]. The same conclusion, namely that incorporating human behavioural factors can increase the perceived relevance by taking personal characteristic of users into account, was reached by Riemer et al. [RB09]. Their work also gives an overview of personalisation techniques and shows which search engines utilise them. Another approach that also fits in the category of human behavioural factors is to include social factors. For instance, Google offers the possibility to recommend search results to Google+ friends. A search engine solely focusing on social search is Eurekster[5]. Whilst it is a reasonable assumption that something relevant to a friend of a user may as well be relevant to the user themselves, some would rather rely on objective standards than on their social graphs [17].

Another approach suggested in [AGZ09] tries to lever collaboratively generated content (CGC) for search. CGC refers to content that is created by a user community, e.g., by aforementioned question and answer tools. As a challenge the authors highlight estimating content quality and usefulness to use both for ranking of CGC. Also they discuss the unification of rankings for CGC and Web documents for what they advocate using two ranking functions because of the nature of the data underlying.

3 Limitations

The previous section has given an overview of important historical developments and recent approaches to Web search. In conclusion it can be said that search engines evolve for mainly one reason – to produce better search results in order to better address increasingly complex information needs. This section will discuss some points of criticism and outline key features an improved search system should offer.

Search engines in general and Google in particular have been accused for not publishing their algorithms and reasoning behind rankings, also referred to as black box nature. However, it is understandable that search engines do not want to publish their methods as their business model is built upon them [22]. At the same time this black box nature fuels suspicion of manipulation [8]. This lack of trust has been discovered and is now actively been fought by search engines as is evident by this Google blog post [15].

Personalisation and socialisation, albeit offering chances for better results [Dop09], are commonly observed with scepticism. Firstly, because using these techniques arguably presents a different Web to different people [18]; and secondly, because they raise privacy concerns as they make heavy use of profiling users [FKMD+05], [GBR09], [Dop09]. Wiechert [Wie09] concludes a privacy study conducted mainly on Google that at least European privacy laws are violated by personalised search. He demands that companies improve their privacy standards, that internationally agreed-upon common privacy standards are enacted and that existing privacy standards are updated to be applicable the Internet era. We believe privacy concerns and lack of trust can be addressed by transparency; i.e., a search system should be designed in a way that users understand how it works and in what way their personal data is used.

More of a technical issue are improvements to deep Web access (be it manually training or automated) and multi-domain queries on multiple heterogeneous data sources such as *name all universities that offer an information systems degree in cities larger than 200k inhabitants and within 50km to an airport in Europe*, which are not yet ready to be extended to a larger scale.

When it comes to data quality the spam problem touched upon earlier must not be underestimated because it is an on-going competition between spammers and search engines, i.e., whenever search engines can detect one kind of spam another evolves [Dop09], [BR10].

As seen throughout the existence of the Web high quality data was always connected to human interaction in the beginning with Web directories and nowadays with Wolfram|Alpha. It is evident in

[5] See http://www.eurekster.com/

itself that only humans can decide what is really relevant to them and therefore, detect spam in all cases. A sensible step towards this human-computer hybrid is the *search computing* paradigm. Though this mainly technology driven idea seems to be able to advance Web search, the initiators themselves state that it is all but clear whether there is demand for SeCo as most users are (most of the time) happy with the simple search offered by major search engines [BR11]. Likewise, it was recognised by Ceri [Cer10] that generally special purpose search engines outperform general purpose search engines; nevertheless, SeCo targets a relatively broad user base with heterogeneous interests.

Therefore, it should be carefully evaluated who is actually in need of advanced search capabilities and an application should be built according to these needs (fit for purpose) rather than building an application which lacks approval. When doing so it should also be considered that there is undoubtedly value in mining the Deep Web for the average user but the real value can be generated for businesses [25].

Additionally there is no clear concept for an organisational or for incentives for the various parties (data providers, developer, and domain experts) interacting in the complex organisation of SeCo. Furthermore, a loosely connected organisational structure that is imposed on the different participants rather than evolved is intuitively harder to manage than having a single organisational body.

From this, some conclusion regarding an ideal search system can be drawn. First and above all, it should solve a user’s problems better than existing solutions. This is also where the concept of transparency comes in. The technical means to achieve high quality results are secondary. Google gained dominance not because of the introduction of PageRank but because users were happier with the results (rankings) Google offered after its introduction. Thus far the average user is contented with results offered by general purpose search engines. Ergo new search strategies have to focus on, and provide solutions for, users that are currently not able to satisfy their complex information needs. Only if this need is identified first, a suitable solution in terms of technical means and organisational form can be found. Starting based on a specific subset of domains will furthermore simplify the implementation because fewer peculiarities have to be considered.

Secondly, historically high quality results for complex queries are achieved by systems relying on humans, who select and organise the data underlying also referred to as curation [20]. This suggests high quality sources that are able to offer the required functionality at the right level of detail should be maintained in one or more centrally curated data repositories. Rather than containing the actual data themselves, the repositories should contain metadata about all subordinate data sources. The difference to SeCo would be that the repository is centrally maintained rather than having sources registering themselves with the repository. To elaborate more on this the next section will suggest how data curation can enhance IR.

4 Future Work

In order to build a new innovative search system, next steps should be to clearly envision and to conceptualise a use case and a target market for high quality data. It has to be decided whether consumer or business are addressed and how high quality data can be delivered. In this context the idea of digital curation as institutionalised by the Digital Curation Centre [RBRB+05] has to be critically examined in order to judge whether curation can advance search. Also collaborative data curation (as done by Wikipedia[6]) should be considered. In two regards ideas can potentially be

[6] http://www.wikipedia.org/

gathered from research on Wikipedia: a) regarding textual data quality and b) regarding incentives to provide data.

Once mechanisms for generating high quality curated data are established, means of delivering these data have to be found. A sensible approach could be to build a service-based architecture. As shown in Section 2, technical issues of such a search systems are being researched extensively and implementations are realised. Our idea is to combine these existing IR techniques (intelligent search, semantic Web, linked data) with approaches from data warehousing and recommender systems as well as with digital curation to deliver the results to any kind of device, including mobile devices. By combining these approaches, which on their own have – as shown in the previous discussion section – only partly solved the problem, we believe a superior search system can be built to satisfy more challenging information needs tailored towards personal and situational usage [DSV12].

That said, it is not our aim to replace existing general purpose search. As state above, classical search is still sufficient for many cases. Our aim is rather to provide a solution for search scenarios were simple query matching does not suffice. One such scenario is topic-centric data collections. As an example one can think of a person who has been diagnosed with a certain disease (e.g., blood cancer). This person may have some information (in form of digital literature about the condition) available already but wishes to collect more recent data and to add documents to their collection of papers regarding this disease [DSV12].

Apart from technical challenges there are two research areas that were only touched upon: organisational form and business models of innovative search systems. Only small sections – if any – of the cited papers were dedicated to business models (e.g. [BBCC+10]) and only few papers exist that solely focus on this subject area (e.g. [BCDP11]) although it has been recognised that the Deep Web offers significant value [Cer10]. Regarding business potential the advertising turnovers are quoted and dissatisfaction with existing solutions is mentioned but to our knowledge there has been no investigation yet regarding the size and kind of a market for high quality data; initial work in this direction is reported in [MSLV12]. In terms of an actual organisational form of such a data provider concepts are very sparse, too, touched up for instance in [BCDP11].

Bibliography

[ABD06] Agichtein, E.; Brill, E.; Dumais, S**.: Improving web search ranking by incorporating user behavior information**. In Proc. 29th Annual International ACM SIGIR Conference, Seattle, WA, USA. Association for Computing Machinery, New York, N.Y, 2006; pp. 19–26.

[AGZ09] Agichtein, E., Gabrilovich, E.. Zha, H: **The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated Content**. In: IEEE Data Engineering Bulletin, 2009, 32; p. 12.

[BR10] Baeza-Yates, R.; Raghavan Prabhakar: Chapter 2: **Next Generation Web Search**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 11-23.

[BR11] Baeza-Yates, R.; Ribeiro-Neto, B.: **Modern information retrieval. The concepts and technology behind search**. Addison Wesley, New York, 2011.

[Ba05] Battelle, J.: **The Search – How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture**; Portfolio (Penguin Group), New York, 2005.

[BCGH10] Baumgartner, R.; Campi, A.; Gottlob Georg; Herzog, M.: Web Data Extraction for Service Creation. In (Ceri, S.; Brambilla, M. Eds.): **Search Computing. Challenges and Directions.**

Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 94–113.

[Ber01] Bergman, M. K.: White Paper: **The Deep Web: Surfacing Hidden Value**. In Journal of Electronic Publishing, 2001, 7.

[BHL01] Berners-Lee, T.; Hendler, J.; Lassila, O.: **The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities**. In Scientific American, 2001.

[BBCC+10] Bozzon, A.; Brambilla, M.; Ceri, S.; Corcoglioniti, F.; Gatti, N.; Milano, P.: Chapter 14: **Building Search Computing Applications**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 268-290.

[BBCF] Bozzon, A.; Brambilla, M.; Ceri, S.; Fraternali P.: Chapter 13: **Liquid query: Multi-domain exploratory search on the web**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010.

[BP98] Brin, S.; Page, L.: **The anatomy of a large-scale hypertextual Web search engine**. In Computer Networks and ISDN Systems, 1998, 30; pp. 107-117.

[BCDP11] Buganza, T.; Corubolo, M.; Della Valle, E.; Pellizzoni, E.: **Analysis of Business Models for Search Computing**. In (Ceri, S.; Brambilla, M. Eds.): Search computing. Trends and Developments. Springer, Berlin ;, New York, 2011; pp. 256–271.

[BD10] Buganza, T.; Della Valle, E.: Chapter 4: **The Search Engine Industry**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 45-71.

[BH11] Burghardt, M.; Heckner, M.; Wolff Christian: **Social Search**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen 2. Neue Entwicklungen in der Web-Suche. AKA, Akad. Verl.-Ges., Heidelberg, 2011; pp. 3–28.

[BR10] Baeza-Yates, R.; Raghavan Prabhakar: Chapter 2: **Next Generation Web Search**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 11-23.

[CHM11] Cafarella, M. J.; Halevy, A.; Madhavan, J.: **Structured data on the web**. In Communications of the ACM, 2011, 54; p. 72.

[CCGM+10] Campi, A.; Ceri, S.; Gottlob, G.; Maesani, A.; Ronchi, S.: Chapter 9: **Service Marts**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 163–187.

[Cer10] Ceri, S.: **Serach Computing**. In (Ceri, S.; Brambilla, M. Eds.): Search Computing. Challenges and Directions. Springer Berlin Heidelberg; Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010; pp. 3–10.

[CHLP+04] Chang, K. C.-C.; He, B.; Li, C.; Patel, M.; Zhang, Z.: **Structured databases on the web**. In ACM SIGMOD Record, 2004, 33; p. 61.

[Com95] Comer, D. E.: **The Internet book. Everything you need to know about computer networking and how the Internet works**. Prentice-Hall International, Englewood Cliffs N.J, 1995.

[DSV12] Dillon, St.; Stahl, F.; Vossen, G.: **Towards The Web in Your Pocket: Curated Data as a Service**. Submitted for publication, 2012.

[Dop09] Dopichaj, P.: **Ranking-Verfahren für Web-Suchmaschinen**. In (Lewandowski, D. Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009.

[FLMR+06] Farahat, A.; LoFaro, T.; Miller, J. C.; Rae, G.; Ward, L. A.: **Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization**. In SIAM Journal on Scientific Computing, 2006, 27; p. 1181.

[FS01] Feldman, S.; Sherman, C.: **The High Cost of Not Finding Information Information**. An IDC White Paper, 2001.

[FKMD+05] Fox, S.; Karnawat, K.; Mydland, M.; Dumais, S.; White, T.: **Evaluating implicit measures to improve web search**. In ACM Transactions on Information Systems, 2005, 23; pp. 147–168.

[Gil09] Giles, J.: “**Knowledge engine” is unveiled / Got a burning question? Ask Alpha**. In The New Scientist, 2009, 202; pp. 18–19.

[Gil93] Gilster, P.: **The Internet navigator**. Wiley, New York, 1993.

[GBR09] Griesbaum, J.; Bekavac, B.; Ritterberg, M.: **Typologie der Suchdienster im Internet**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009.

[Jer10] Jer Lang Hong: **Deep web data extraction**: 2010 IEEE international conference on systems, man man and cybernetics. IEEE Press Books, New York, 2010; pp. 3420–3427.

[Kle99] Kleinberg, J. M.: **Authoritative sources in a hyperlinked environment**. In Journal of the ACM, 1999, 46; pp. 604–632.

[LM06] Langville, A. N.; Meyer, C. D.: **Google's PageRank and beyond. The science of search engine rankings**. Princeton Univ. Press, Princeton [u.a.], 2006.

[LM00] Lempel, R.; Moran, S.: **The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effec**t. In ACM Transactions on Information Systems, 2000, 19; pp. 387–401.

[Lev10] Levene, M.: **An Introduction to Search Engines and Web Navigation**: Wiley, Hoboken, 2010.

[LSDJ06] Lew, M. S.; Sebe, N.; Djeraba, C.; Jain, R.: **Content-based multimedia information retrieval: State of the art and challenges**. In ACM Trans. Multimedia Comput. Commun. Appl, 2006, 2; pp. 1-19.

[Lew09] Lewandowski, D.: **Spezialsuchmaschinen**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009; pp. 53–69.

[LLD11] Liu, G.; Liu, K.; Dang, Y.-y.: **Research on discovering Deep Web entries based ontopic crawling and ontology**: Conference on Electrical and Control Engineering (ICECE), 2011 International, 2011; pp. 2488–2490.

[MSHP09] Maaß, C.; Skusa, A.; Heß, A.; Pietsch, G.: **Der Markt für Internet-Suchmaschinen**. In (Lewandowski, D.. Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009.

[MKKG+08] Madhavan, J.; Ko, D.; Kot, L.; Ganapathy, V.; Rasmussen, A.; Halevy, A.: **Google’s Deep Web crawl**. In Proc. VLDB, 2008, 1; pp. 1241-1252.

[MV98] Masermann, U.; Vossen, G.: **Suchmaschinen und Anfragen im World-Wide Web**; Informatik-Spektrum 21, 1998; pp. 9-15.

[MV00] Masermann, U., Vossen, G.: **SISQL: Schema-Independent Database Querying (on and off the Web)**; Proc. 4th International Conference on Database Engineering and Applications (IDEAS) 2000, Yokohama, Japan, IEEE Computer Society Press; pp. 55-64.

[MSLV12] Muschalle, A.; Stahl, F.; Löser, A.; Vossen, G.: **Pricing Approaches for Data Markets**; to appear in 6th International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE), Istanbul, Turkey, 2012.

[MO07] Musser, J.; O'Reilly, T.: **Web 2.0. Principles and best practices**. O'Reilly Media, Sebastopol, CA, 2007.

[OLC11] Otto, B.; Lee, Y. W.; Caballero, I.: **Information and data quality in networked business**. In Electronic Markets, 2011, 21; pp. 79–81.

[PBMW99] Page, L.; Brin, S.; Motwani, R.; Winograd, T.: **The PageRank Citation Ranking: Bringing Order to the Web**. In STANFORD INFOLAB, 1999; p. 17.

[Qui09] Quirmbach, S.: **Universal Search. Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009; pp. 220–248.

[Raj09] Rajaraman, A.: **Kosmix: Exploring the Deep Web using Taxonomies and Categorization**. In: IEEE Data Engineering Bulletin, 2009, 32; p. 12.

[RB09] Riemer, K.; Brüggemann, F.: **Personalisierung der Internetsuche**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009; pp. 148–169.

[RBRB+05] Rusbridge, C.; Burnhill P.; Ross S.; Buneman P.; Giaretta D.; Lyon L.; Atkinson M.: **The digital curation centre: A vision for digital curation**. In Local to Global Data Interoperability - Challenges and Technologies, 2005.

[Sto03] Stock, W. G.: **Weltregionen des Internet: Digitale Informationen im WWW und via WWW**. In Password, 2003; pp. 26_28.

[SC11] Stoffle, C. J.; Cuillier, C.: **Living the Future: Introduction**. In Journal of Library Administration, 2011, 51; pp. 595–598.

[VH07] Vossen, G.; Hagemann, S.: **Unleashing Web 2.0. From concepts to creativity**. Elsevier; M. Kaufmann publ., Burlington (Mass.), 2007.

[Wie09] Wiechert, T.: **Datenschutz bei Suchmaschinen**. In (Lewandowski, D., Hrsg.): Handbuch Internet-Suchmaschinen. Nutzerorientierung in Wissenschaft und Praxis. AKA, Akad. Verl.-Ges., Heidelberg, 2009; pp. 285–300.

[WFH11] Witten, I. H.; Frank, E.; Hall, M. A.: **Data mining. Practical machine learning tools and techniques**. Morgan Kaufmann, Burlington, MA, 2011.

[XCZY+10] Xian, X.; Cui, Z.; Zhao, P.; Yang, Y.; Zhang, G.: **Utility Maximization Model for Deep Web Source Selection and Integration**. In Journal of Computers, 2010, 5.

Web References

[1] *About Wolfram|Alpha*. http://www.wolframalpha.com/about.html. Accessed 12th June 2012.

[2] *SES San Jose: Universal and Blended Search*. http://www.toprankblog.com/2008/08/ses-san-jose-universal-and-blended-search/. Accessed 12th June 2012.

[3] *Snapshot of a web directory on the first webserver*. http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/bySubject/Overview.html. Accessed 12th June 2012.

[4] *The website of the world's first-ever web server*. http://info.cern.ch/. Accessed 12th June 2012.

[5] *The WWW Virtual Library*. *About the Virtual Library*. http://vlib.org/AboutVL. Accessed 12th June 2012.

[6] *yahoo! Meilensteine*. http://yahoo.enpress.de/Meilensteine.aspx#1994. Accessed 12th June 2012.

[7] 2010. *The Economist: The Data Deluge - Businesses, governments and society are only starting to tap its vast potential*. http://www.economist.com/node/15579717. Accessed 12th June 2012.

[8] "Cody". 2011. *Is Google Manipulating Search Results In Their Favor?* http://www.frooglegeek.com/?p=337. Accessed 12th June 2012.

[9] Berners-Lee, T. 1989. *Information Management: A proposal*. http://www.w3.org/History/1989/proposal.html. Accessed 12th June 2012.

[10] David Bailey. 2007. *An Insider’s View Of Google Universal Search*. http://searchengineland.com/an-insiders-view-of-google-universal-search-12059. Accessed 12th June 2012.

[11] Efrati, A. *Google Gives Search a Refresh*. http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html. Accessed 12th June 2012.

[12] Fox, V. 2011. *Schema.org: Google, Bing & Yahoo Unite To Make Search Listings Richer Through Structured Data*. http://searchengineland.com/schema-org-google-bing-yahoo-unite-79554. Accessed 12th June 2012.

[13] Google. *Google Basics*. http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70897. Accessed 12th June 2012.

[14] Google. *Google Begins Move to Universal Search*. http://www.google.com/intl/en/press/pressrel/universalsearch_20070516.html. Accessed 12th June 2012.

[15] Huffman, S. *Search quality highlights: new monthly series on algorithm changes*. http://insidesearch.blogspot.com/2011/12/search-quality-highlights-new-monthly.html. Accessed 12th June 2012.

[16] Jasra, M. *Interview: Yahoo's Larry Cornett on Universal Search*. http://www.searchengineguide.com/manoj-jasra/interview-yahoos-larry-cornett-on-univer.php. Accessed 12th June 2012.

[17] Jon Mitchell. 2011. *A Year of Tweaks to Google Search: Are You "Fed Up?"*. http://www.readwriteweb.com/archives/a_year_of_tweaks_to_google_search_are_you_fed_up.php. Accessed 12th June 2012.

[18] Jon Mitchell. 2011. *Google+ Is Going To Mess Up The Internet*. http://www.readwriteweb.com/archives/google_is_going_to_mess_up_the_internet.php. Accessed 12th June 2012.

[19] Jon Mitchell. 2012. *How Google Search Really Works. (Interview with Ben Gomes).* http://www.readwriteweb.com/archives/interview_changing_engines_mid-flight_qa_with_goog.php. Accessed 12th June 2012.

[20] Oxford Dictionaries. *Oxford Dictionaries: Curation*. http://oxforddictionaries.com/definition/curate--2?q=curation. Accessed 12th June 2012.

[21] Smarty, A. 2009. *Let's Try to Find All 200 Parameters in Google Algorithm*. http://www.searchenginejournal.com/200-parameters-in-google-algorithm/15457/. Accessed 12th June 2012.

[22] Sullivan, D. 2010. *Schmidt: Listing Google's 200 Ranking Factors Would Reveal Business Secrets*. http://searchengineland.com/schmidt-listing-googles-200-ranking-factors-would-reveal-business-secrets-51065. Accessed 12th June 2012.

[23] Tim Berners-Lee. 2009. *Linked Data*. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 12th June 2012.

[24] W3C. *Resource Description Framework (RDF)*. http://www.w3.org/RDF/. Accessed 12th June 2012.

[25] Wright, A. 2009. *Exploring a 'Deep Web' That Google Can't Grasp*. http://www.nytimes.com/2009/02/23/technology/internet/23search.html?_r=1&ref=business. Accessed 12th June 2012.