Separate the wheat from the chaff: Mapping the current and future landscape of web search engines

by

Student: *Silke Bouwman*
Supervisor: *Alexandra Cristina Ramos da Silva Lopes Gunes*
Co-supervisor: *Sérgio Nunes*

A thesis submitted in conformity with the requirements
for the degree of Master in Innovation & Technological Entrepreneurship

Faculty of Engineering & Faculty of Economics
University of Porto

Separate the wheat from the chaff: Mapping the current and future landscape of web search engines

Silke Bouwman

Master in Innovation & Technological Entrepreneurship

Faculty of Engineering (FEUP) & Faculty of Economics (FEP)
University of Porto

2019

Abstract

The literature in the field of Information Retrieval (IR) is considerably expansive and therefore dense with trends, innovations and challenges (categories of factors), however by no means descriptive on the importance of these factors. Therefore we zoomed-in on an important sub-domain of IR, web search engines, to evaluate the importance of these factors according to the perspectives of academic- and industry-experts in the field of search. A preliminary list of 155 factors (29 trends, 34 innovations and 92 challenges) was found in the literature, by analyzing summarizing IR papers (e.g. conventions, systematic reviews, surveys) within a timeframe of 5 years. Using the Delphi-methodology, two panels of identified experts participated in a three round survey research, validating and complementing the preliminary list of factors, selecting the most important factors, and in the third survey ranking those in terms of importance, resulting in ranked lists of these factors per category and component (crawling, indexing, ranking, evaluation and web search engines in a holistic view). By using this procedure the initial list of 155 factors was reduced to only 45 factors ranked in terms of importance in the perspectives of the panels and form a map of the current environment and future direction.

Industry-experts do not identify as many factors as academic experts, however industry-experts do have a higher level of consensus on the importance of the factors. Industry-experts identifies fully commercialized innovations, while academic-experts do identify innovations in an early stage of development as well. Nevertheless, both panels strongly agree on the importance of NLP as innovation. Trends concerning machine learning rank higher within industry. The trend of conversational search is highly supported by both panels as well as clients, measured by indicative survey results, indicating a promising direction for web search engines. Academic experts highly support the trend of semantic query processing, however they also see it as the biggest challenge within ranking.

Keywords: IR, Web Search Engines, Trends, Innovations, Challenges,

Acknowledgments

Here I would like to take the opportunity to express my gratitude to the persons who helped greatly in the last couple of months. Not only for their direct support for my dissertation, but also for their personal support.

First of all, I would like to thank my supervisor, professor Alexandra Lopes, for her ongoing supports, which never stops to amaze me. I set my expectations of having a good supervisor, and Alexandra keeps surprising me by going the extra mile to help me. Even though the topic was somewhat out of her field of expertise, I never had any doubts on her competency as a supervisor. She helped me greatly with the part I expected to struggle most with: a consistent methodology.

Next, I would like to thank my co-supervisor professor Sérgio Nunes together with his PhD student José Devezas, who despite their busy schedules were able to guide me within the field of IR and ask critical questions putting me to work. I was able to get up to speed due to the guidance I received, especially in the beginning of the dissertation.

My guidance did not limit itself to the university, since my research was conducted during a 5 month internship at Incentro. The company made me feel very welcome even before I started working, by sending flowers. Every day at the office was a day well spent, with tons of learnings, fun and occasionally games of Fifa. With great pride I can share I will join the team of Incentronauts after graduating. Although I have many persons to thank within Incentro, there is one person I would like to thank especially: Jorick van Hees, my supervisor within the company, who has played a big role with guiding me within the organization. Additionally, he frequently came with ideas and feedback, which was highly appreciated.

Lastly, I would like to thank all the participants of the research. They were willing and able to participate in multiple surveys, sacrificing their precious time. I hope I have helped raise their awareness about interesting developments within the IR landscape.

Table of Contents

Acknowledgments v

Table of Contents vii

List of Tables ix

List of Figures xi

List of Appendices xiii

List of Abbreviations xv

1 Introduction 1

1.1 Background 1

1.2 Crawling 3

1.3 Indexing 4

1.4 Ranking 4

1.5 Evaluation 5

1.6 Motivation of study 6

1.7 Research objectives 6

2 Literature review 7

2.1 Methods 7

2.1.1 Eligibility criteria 7

2.1.2 Information sources 8

2.1.3 Search 8

2.1.4 Study selection 8

2.1.5 Data collection process 8

2.1.6 Data items 9

2.2 Results 9

2.2.1 Study selection 9

2.2.2 Study characteristics 9

2.2.3 Identified trends, innovations and challenges in the literature 11
2.2.3.1 Crawling 11
2.2.3.2 Indexing 16
2.2.3.3 Ranking 17
2.2.3.4 Evaluation 22
2.2.3.5 General 25
2.2.4 Limitations 34
3 Methodology 35
3.1 Motivation 35
3.2 The Delphi methodology 36
3.2.1 The process 36
4 Results 41
4.1 Analysis of the survey rounds 41
4.1.1 Analysis of the 1st survey results 41
4.1.2 Analysis of the 2nd survey results 43
4.1.3 Analysis of the 3rd survey results 43
4.2 Overall results 48
5 Discussion & conclusions 49
References 51
Appendices 59

List of Tables

Table 1 - Query results 8

Table 2 - Paper selection characteristics 10

Table 3 - Extracted trends, challenges and innovations of crawling 14

Table 4 - Extracted trends, challenges and innovations of indexing 17

Table 5 - Extracted trends, challenges and innovations of ranking 21

Table 6 - Extracted trends, challenges and innovations of evaluation 24

Table 7 - Extracted trends, challenges and innovations of web search engines 29

Table 8 - Level of engagement survey round 1 42

Table 9 - Kendall's W 44

List of Figures

Figure 1 - Web search engine process (source: Kathuria et al., 2016) 2

Figure 2 - Process of a web crawler (source: Kumar et al., 2017) 3

Figure 3 - Flow diagram paper selection 9

Figure 4 - Flowchart sampling procedure 38

Figure 5 - Trends, ranked in terms of importance... 45

Figure 6 - Innovations, ranked in terms of importance 46

Figure 7 - Challenges, ranked in terms of importance 47

List of Appendices

Appendix 1: Template survey 1 academics & industry 59

Appendix 2: Template survey 1 clients 61

Appendix 3: Template survey 2 academics & industry 63

Appendix 4: Template survey 3 academics 65

Appendix 5: Template survey 3 industry 67

Appendix 6: 1st Survey results analysis, clients 69

Appendix 7: 1st Survey results analysis, academic- & industry-experts 71

Appendix 8: 2nd Survey results analysis, academic- & industry-experts 79

Appendix 9: Analysis of all three survey rounds compared to each other 85

List of Abbreviations

NLP	Natural Language Processing
IR	Information Retrieval
URL	Uniform Resource Locator
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
IT	Information Technology
QA	Question Answering
DNS	Domain Name System
DOM	Document Object Model
FPGA	Field-Programmable Gate Array
GPU	Graphical Processing Unit
TREC	Text Retrieval Conference
NTCIR	NII Testbeds and Community for Information access Research
GPGPU	General-Purpose computing on Graphics Processing Units
IOT	Internet Of Things
POS-tagging	Part Of Speech tagging
SEO	Search Engine Optimization

1 Introduction

This chapter provides an overview of the concepts of web search engines. The goal is not to provide a fully comprehensive overview of each concept. Instead, it provides an overview of the process of web search engines and its most important concepts. Finally, Section 1.7 discusses the motivation for the study and Section 1.8 the research objectives.

1.1 Background

What started as a humble journey with just a few computers at the Advanced Research Projects Agency (ARPA) connected to the first versions of the internet in 1969, has grown to an information big bang by the usage of the public. This has led to a big pile of heterogeneous data, which needs to be searchable in order to create value. As such the first web search engine was introduced in 1990 by the name of Archie (Kathuria et al., 2016). The industry has been evolving enormously ever since with names as Google and Yahoo among others to enter the market. Nowadays search engines have become a part of our day to day life.

Finding information is like finding a needle in a haystack and therefore there is a real need for efficient and effective search engines. The difficulty in achieving this is the size, diversity and dynamicity of the world wide web (Cambazoglu & Baeza-Yates, 2015). Continuous efforts are made to enhance the current solutions to achieve more relevant results in a more efficient manner.

Web search engines consists out of three major components, which together represent the web search engine (Brin & Page, 1998). The first component is the crawler, which visits all desired website pages and downloads its contents. Secondly, the indexing components necessary for indexing the downloaded pages in order to transform the page into a homogenous format and store it in the index. The third component is the ranking algorithm in which the user's search query is matched to the indexed documents and ranked in order of relevance. Lastly, an independent process, the evaluation, for determining the actual relevance of the visited proposed web pages. Evaluation can be seen as an independent process, since it has no direct effect on the core process of the web search engine.

There are two types of search engines: crawler based and human powered based. Crawler based web search engines are algorithm based in which human intervention is excluded. Human powered web search engines are indexed by humans and are not capable to process large volumes of documents. Therefore, a hybrid is a more common approach, since humans can alter the results for highly frequent queries in order to achieve higher relevance and a crawler can efficiently deliver results for all possible queries.

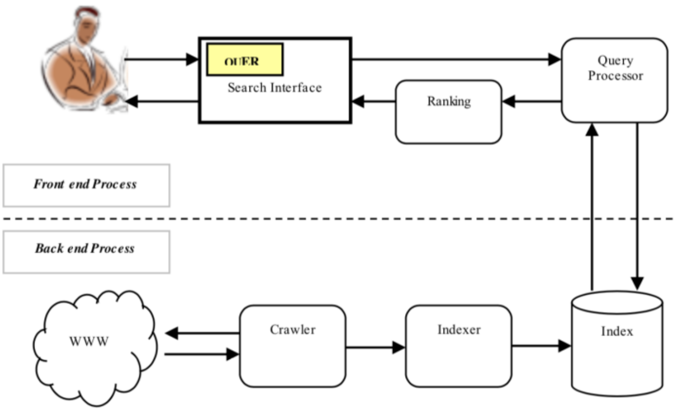


Figure 1 - Web search engine process (source: Kathuria et al., 2016)

Figure 1 shows a schematic overview of the web search engine process. First the crawler traverses the web and downloads the desired pages, see Section 1.2. These pages get indexed by the indexer and stored in an index. The index is the database web search engines store and retrieve their data from, see Section 1.3. When the user then submits a query to the web search engine, the query first gets processed by a query processor, Section 1.4 elaborates on this process. The modified query gets sent to the index and retrieves the webpages with corresponding characteristics in the form of keywords conventionally. Followed by a ranking algorithm, which ranks the retrieved results in order of relevance to the user based on the modified query.

In the following sections the components of web search engines will be discussed in greater detail. Furthermore, the motivation for the study will be discussed in addition with the research objectives.

1.2 Crawling

The crawler, also called spider, robots or wanderers is the supplier of information to the index. According to the literature around 40% of the total internet traffic and consumptions of bandwidth is due to web crawlers (Badawi et al., 2013). The crawler traverses the internet to gather new information to process. The traditional crawler starts from a base-URL or multiple seed URLs, it downloads the URL and collects all connected hyperlinks and URLs. Next, it prioritizes the found URL's and visits the next one in line to repeat the process, see Figure 2. The base-URL is the main domain of which the contents are to be extracted. As such, some crawlers are bound by the base-URL and only crawl the extensions of it.

Size and variety are the main challenges of the crawler, since web content is constantly added and updated, the crawler has a tradeoff between coverage and freshness. To measure the effectiveness of crawlers, regular metrics used are coverage, freshness, page importance and throughput.

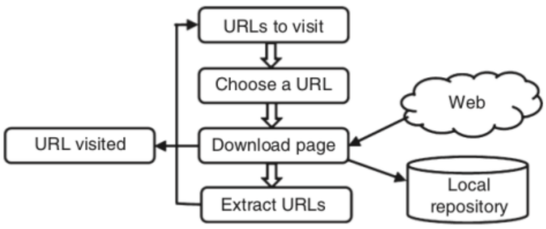


Figure 2 - Process of a web crawler (source: Kumar et al., 2017)

Kumar et al. (2017) describes five different types of crawlers: universal crawlers, preferential crawlers, hidden web crawlers, mobile crawlers and continuous/incremental crawlers. The first, universal crawlers or broad crawlers, are not limited to a base URL and are simply parsing all the information it finds. Preferential crawlers do not crawl all links they encounter rather the user submits a condition or topic of interest that guides the preferential crawler. Preferential crawlers themselves can be divided into three categories(Kumar et al., 2017). 1) Focused crawlers give priority to those URLs in the process of crawling, in which probability of finding information of user's interest is high. 2) Topical crawlers are restricted to a certain topic (e.g. football). 3) Lastly forum crawlers, which are solely used for crawling forums. The third type of crawlers, hidden web crawlers, are used to traverse the hidden web, since a significant amount of information cannot be found by simply following hyperlinks. Mobile web crawlers make use of server-side crawling in which the selection and filtering takes place, resulting in a reduction of network load caused by the traditional web crawler (Hammer & Fiedler, 2000). The last type, the continuous/incremental crawlers, are used for maintaining the indexed web pages, guaranteeing freshness of the pages.

1.3 Indexing

Indexing is the following step after crawling, in which the downloaded web pages are transformed in a format, called the index, to use later on in the ranking process and is in turn the main information that gets fed into the rank algorithm(s). The indexer catalogs and organizes detailed data on the nature of the content and the topical relevance of each web page, a map with the links between each page, the clickable anchor text of any links, and additional data e.g. if the links are ads, where they are located on the page and so forth

Currently the most commonly used index structure is the inverted index model (Lester et al., 2006; Mahapatra & Biswas, 2011; Seo et al., 2003). Inverted indexes associate each term/word with the document in which it is present. Then, when the user queries those terms it will retrieve all the documents containing those terms. Since full text search systems became available to the public, and the use of natural language, the queries contain more noise (Tekli et al., 2019). As such queries get formulated which are syntactically different due to non-expert usage. Therefore, the results have lower relevance (Tekli et al., 2019).

One way to tackle this issue is using semantic query processing. Semantic query processing extracts the associative and contextual nature of queries, and can answer more fuzzy and wide open questions. Search is one of the most used features in the Internet (Pewinternet, 2019), and it is evolving in ways that can give users more meaningful data than before. Including the use of semantic search, which is one of the hottest fields in recent years that have gained attraction.

1.4 Ranking

The process of ranking is one of the most computational expensive processes of the web search engine and continuous efforts are made to increase its efficiency (Catena & Tonellotto, 2015). During ranking, the user’s query is processed, the query processing can contain several steps e.g.

adding synonyms, translating, stemming. Next, the processed query is sent to the index, returning results that are subsequently ranked based on a set of algorithms.

A well-known ranking algorithm is PageRank (Page et al., 1998), which ranks the results based on the authority score of the pages, which take the number and status of incoming links into account. Currently commercial web search engines use a great variety of ranking algorithms in order to increase the relevance of the returned results.

1.5 Evaluation

How to conduct IR system evaluation has been an active area of research for the past 50 years or so, and the subject of much discussion and debate (Robertson, 2008; Saracevic, 1995). Evaluating the performance of an IR system is an important part of developing an effective, efficient and usable search engine (Robertson, 2008; Saracevic, 1995). Saracevis (1995) distinguished six levels to evaluate IR systems:

1. The engineering level deals with aspects of the technology, e.g. computer hardware and networks including metrics as reliability, errors, failures and faults.
2. The input level, evaluates the contents of the systems using metrics as freshness and coverage of the document selection.
3. The processing level deals with the performance of the output of the system and its interaction. Assessing the performance of the algorithms for indexing and retrieval.
4. The output level deals with the interaction with the system and output(s) obtained to evaluate search interactions, feedback and outputs.
5. The use and user level evaluate how well the search engine supports the search task of the user in the wider context of information seeking behavior.
6. The social level deals with the impact on the environment and could assess aspects as productivity, effects on decision making and socio-cognitive relevance.

The conventional focus of evaluating search engines lies in the first three levels. For the third level, one can use online or offline evaluation techniques to assess the relevance of the web search engine's results, each with their own pros and cons. Online evaluation methods offer a fast, cheap and more scalable alternative to the traditional offline evaluation methods (Hofmann, Li & Radlinski, 2016). However offline evaluation methods are based on editorial judgement and therefore arguably more accurate. Typical online evaluation metrics include click based metrics such as CTR (click through rate), UCTR (binary value representing click) (Chuklin et al., 2013) and PLC (number of clicks divided by the position of the lowest click) as well as dwell time-based metrics such as query dwell time, average of click dwell time (Jiang et al., 2015) and so on (Chen et al., 2017). The mentioned metrics sketch a clear view on the performance of the algorithms, however will not be discussed in great detail due to the limited scope of the dissertation. Typical offline metrics include average precision, normalized discounted cumulative gain and rank-biased precision (Deng & Shi, 2016). Two major risks, while conducting offline evaluation are the potential bias due to a small and incomplete dataset and the difference between user's assessments and assessor's judgements (Liu et al., 2015).

1.6 Motivation of study

There is a considerable amount of academic literature concerning all components of web search engines and even more of Information Retrieval, however the information is scattered. There is a lack of academic documents describing the web search engine landscape in a holistic view. In addition, the relationship between various panels is not described. Therefore, the motivation of this study is to first sketch a general view of the components of web search engines and identify the corresponding trends, innovations and challenges in the literature. Followed by gaining a deeper understanding on the different perspectives of clients (of search solutions), academic experts and industry experts. This dissertation is by no means an extensive systematic review of all the components of web search engines and should not be treated as such. Nevertheless, it offers a general introduction to web search engines by identifying the state of the art and future directions.

In addition, this study is commissioned by Incentro, an IT-company based in the Netherlands, to support their resource allocations for innovations in the field. Incentro facilitated the study in the form of a 5-month internship.

1.7 Research objectives

The main research objective is to identify and assess the importance of trends, innovations and challenges in the field of web search engines as a whole, as well as for its four main processes (crawling, indexing, ranking, evaluating). Furthermore, to evaluate the relationship between the perspectives of industry and academics experts in the field of web search engines.

2 Literature review

This chapter provides an overview of current trends, innovations and challenges in the field of web search engines found in the literature. Section 2.1 details the methodology of the literature review, followed by Section 2.2 describing the found results. Based on the large amount of identified factors, the goal is not to describe all identified factors (trends, innovations and challenges) in detail, but to generate a generic overview.

2.1 Methods

For the literature review, the PRISMA guidelines were followed (Liberati, 2009), a commonly, however not exclusively, method used for conducting systematic literature reviews. PRISMA is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses and was originally developed in the health sciences but offers a series of guidelines that have progressively inspired other fields.

"*Systematic reviews should build on a protocol that describes the rationale, hypothesis, and planned methods of the review; few reviews report whether a protocol exists. Detailed, well-described protocols can facilitate the understanding and appraisal of the review methods, as well as the detection of modifications to methods and selective reporting in completed reviews"* (Shamseer, 2016).

The PRISMA guidelines will be used, since the systematic approach fits the means of this research. However, it is slightly changed to fit the current application better.

2.1.1 Eligibility criteria

In order to increase the relevance and reduce the scope of the review, several eligibility criteria are established. First of all, since the number of publications in the field of IR/search engines is too extensive for a MSc dissertation project, this research is limited to summarized information sources (surveys, systematic reviews, books and conventions). In addition, the publication has to be published in 2014 or later due to the fast-changing landscape of IR. Lastly, the publication needs to be in the field of computer science and must be written in English in order to be eligible for this study.

2.1.2 Information sources

The information sources used for the systematic literature review are Scopus and papers suggested by supervisors. The last search has been conducted on February 8th, 2019.

2.1.3 Search

For the query formulation, multiple criteria have been taken into account to be as broad, but simultaneously as specific, as possible. The eligibility criteria are processed in the query, in addition to the selection of words needed to be present in the title, abstract or keywords.

Table 1 - Query results

Query	Number of results
(TITLE-ABS-KEY ((search AND engines) OR crawling OR indexing OR ranking OR (search AND engine AND evaluation) OR ir OR (information AND retrieval))) AND (((trend AND innovation) OR (innovation AND challenge) OR (trend AND challenge))) AND ((systematic AND review) OR survey OR convention OR architecture OR meta-analysis) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014)) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (LANGUAGE , "English"))	2627

2.1.4 Study selection

In order to filter out the noise, a study selection procedure has been used with the following steps:

1. Remove repeated articles.
2. Determine compatibility with the articles' title and the research topic.
3. Determine the congruency of each article's abstract with the research topic.
4. Determine the research topic's inclusion or appearance in the body text of the article.

2.1.5 Data collection process

The main objective for the data collection process is to summarize and identify the trends, innovations or challenges mentioned in the selected publications. These factors were divided in five fields: crawling, indexing, ranking, evaluation and general, since the selected paper could cover web search engines as a whole or solely a component of it.

2.1.6 Data items

For the data collection a Microsoft Excel sheet was used to extract the desired information. The variables for data extraction were: Author(s), Year of publication, Title, Publication Journal, Keywords, Category, Methods, Geography, Concepts, Crawling{Trends, Innovation, Challenges}, Indexing{Trends, Innovation, Challenges}, Ranking{Trends, Innovation, Challenges}, Evaluation{Trends, Innovation, Challenges}, General{Trends, Innovation, Challenges}, Question & Objective, Potential bias and Observations. Resulting in a quite extensive data extraction for the continuation of the systematic review.

2.2 Results

The systematic review was conducted as laid-out in Section 2.1. The results are described in this section. Section 2.2.1 describes the results of the selection procedure. Section 2.2.2 describes the characteristics of the found publications, and finally Section 2.2.3 lists the found factors and describes some factors in greater detail.

2.2.1 Study selection

For the study selection 2.627 found documents were screened. The documents were filtered and selected based on the selection procedure described in Section 2.1.4. The selection procedure is illustrated by a flow diagram in Figure 3.

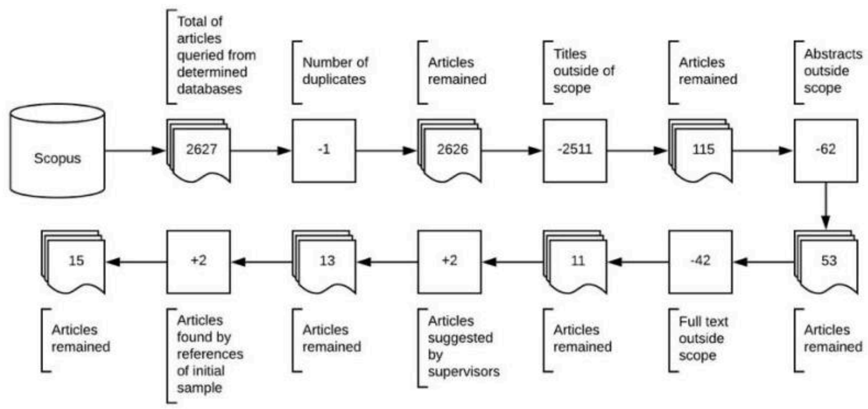


Figure 3 - Flow diagram paper selection

2.2.2 Study characteristics

The extracted characteristics of the publications can be found in Table 2.

Table 2 - Paper selection characteristics

Author(s) & Year	Category	Methods	Concepts covered
Kathuria et al., (2016)	General	Systematic review	Human powered directories, hybrid search engine, meta search engine, vertical search engine
Moulahi et al., (2015)	General	Systematic review	Time-aware IR
Culpepper et al., (2018)	General, crawling, indexing, ranking, evaluation	Convention	Generated information objects, personal information access, conversational information seeking, FACT IR, IR for decision making, evaluation, learnable information retrieval, next-gen efficiency, IOT search engine
Kumar et al., (2017)	Crawling	Systematic review	Types of crawlers
Morales & Melgar, (2017)	Ranking	Systematic review	Semantic search engines
Calumby et al., (2016)	General, ranking, evaluation	Systematic review	Learn to rank, relevance feedback, interactive systems
Elbedweihy et al., (2015)	General, ranking, evaluation	Systematic review	Semantics
Soares & Parreiras, (2018)	General, evaluation	Systematic review	QA technologies
Shen et al., (2015)	General, Ranking	Survey	Entity linking techniques
Huurdeman & Kamps, (2014)	General	Systematic review	User interface multiphase
Asikri et al., (2017)	General	Systematic review	Semantic web mining
Jayanthi & Rathi, (2014)	General	Systematic review	Personalized search
Tamine & Daoud, (2018)	Evaluation	Systematic review	Evaluation considering contextual information
Altinel & Ganiz, (2018)	General, ranking	Survey	Semantics
Balog, (2018)	General, indexing, ranking	-	Entity oriented search, semantics

2.2.3 Identified trends, innovations and challenges in the literature

All the selected papers have been analyzed in order to find consensus across the papers concerning trends, innovations and challenges. It is common they mention factors in multiple components, however, few were truly addressing all.

Culpepper et al. (2018) is considered the most vital paper due to the fact it is built upon the expertise of the most prestigious researchers in the field of IR. The paper is constructed by the consensus of a convention discussing the trends and challenges currently present in the field. Furthermore, the recency of the paper is an important factor for its relevance. Although web search engines do not stretch as far as IR, the two are intertwined, therefore, most of the identified trends, challenges and innovations will have some applicability on web search engines.

In the sections below, the extracted trends, innovations and challenges of each component and web search engines in a holistic perspective are listed in tables. Additionally, the most relevant, based on source and consensus are discussed in detail.

2.2.3.1 Crawling

Hidden/deep web search

The development of deep or hidden web search engines is an identified trend (Kathuria et al., 2016; Kumar et al., 2017). The deep- or hidden-web is the part of the world wide web, which is not indexed. It is estimated that the size of the deep web is 500 times larger than the indexed web (Sumit & Muktawat, 2010), additionally its heterogenous nature and dynamicity makes it hard to crawl (Singh & Sharma, 2013). Therefore it is the challenge to explore the hidden web. Kathuria et al., (2016) lists five reasons why web pages cannot be found by traditional web search engines:

1. Due to dynamic pages, non-linked web pages, pages that require registration or webpages whose content vary as per access rights and contexts.
2. Websites prohibiting search engines to index the pages in their robot.txt file or the use of captcha.
3. Webpages consisting of textual formats, which are not conventionally readable by web search engines.
4. Web contents intentionally kept invisible to the standard internet.

5. Archived versions of webpages, which are irrelevant to web search engine, since they strive to keep the pages fresh.

Only open corpus sources can be targeted for crawling data present behind the login page comes under the category of hidden web (Kumar et al., 2017). Kumar et al., 2017 describes five approaches to access and index the hidden web:

1. *Keyword query-based approach*: The crawler searches for search interfaces on the webpages. If it found the search interface it will input a single- or multi-attribute query in order to access the data hidden behind the search interface. However, the challenge of this approach is to insert meaningful, relevant and efficient queries. To increase these factors the input can be based on the user's query or on the indexed domain.
2. *Form-based approach*: A category of crawlers specialized in detecting forms on the webpages and the type of form. These types can range from all sorts as polling forms, login forms, search forms, message posting forms, email forms, subscription forms, etcetera. One of the ways of detecting the type of form is using machine learning, which after detecting, inputs relevant data. This way the data behind the form can be extracted. Another way is using ontologies for identifying the type of form and input relevant data.
3. *Revisit policy and incremental approach*: This method will input a query in the search interface and extract the results. Based on a probabilistic time period, the process is repeated, this time with the query optimized based on the extracted information of the previous round. Resulting in more relevant and fresh webpages.
4. *Attribute and label extraction approach*: Form labels and attributes are the prime requirements for many techniques that aim at retrieving data from hidden databases (Kumar et al., 2017). The attributes and labels play a major role in how the hidden database can be accessed. Some forms will have indefinite possibilities due to white box inputs, while some others will have finite possibilities due to drop-down lists for example. By extracting this knowledge, it can be leveraged when the user inputs a query by inputting the relevant parts of the query in the forms to access the hidden data.
5. *Labeled value set-based approach*: The labels from forms are extracted and categorized. Followed by matching these labels with label-value pairs extracted from a knowledge database, in order to input data in the labels on the webpage.

6. *Domain- or topic-specific approach*: After identifying the forms it is the challenge to input relevant and efficient queries. Therefore the domain- or topic-specific approach determines which subjects or domains can be queried on which webpages, based on the indexed data of the webpage.

Onion search is a type of deep web search; however, it is the part of the internet with a negative intent for illegal activities. These activities are hard to trace due to opacity between information seeker and provider due to a pseudo top level domain (TLD). The TLD in case of onion sites is not an actual DNS root, but an access mechanism provided through a proxy server (Kathuria et al., 2016). The addresses are automatically generated based on a public key when the hidden service is created. Due to the illegal activities, the interest from government has risen to a level that serious attempts are made to trace down the activities. One of these efforts is the Defense Advanced Research Projects Agency (DARPA). They developed a search engine, Memexa, in an attempt to penetrate the deep web.

Dealing with non-uniform structures

As there is no universal norm on how to build websites, the web is built of dynamic and inconsistent data structures. Due to lack of uniformity, collecting data becomes difficult. The problem is amplified when the crawler has to deal with semi-structured or unstructured data (Abiteboul 1997; Kathuria et al., 2016; Kumar et al., 2017, Morales & Melgar 2017*)*. Therefore, it is a challenge to find effective and efficient ways to cope with this issue.

Mobile crawlers

Centralized data access, centralized webpage filtering and centralized indexing are some of the limitations of traditional web crawling (Kumar et al., 2017). However, the limitations can be effectively handled by the use of mobile web crawlers, enabling server-side executable code. Instead of using the traditional “data to code”, mobile web crawlers use “code to data” in which the web crawler transfers itself to the web server. Resulting in the advantages of localized data access, remote webpage selection, remote webpage filtering and compressions that are highly efficient regarding resources (Kumar et al., 2017).

The application framework of the mobile crawler, responsible for the creation and management, can access the server on remote. The virtual environment of the server manages various policies, resulting in permission to only execute particular code. This way the mobile crawler consists out of a web search engine agent and a slave. The slave is positioned on the server side determining which contents to download. Additionally, it can detect when a webpage is updated due to the webpage size and pushes these pages to the search engine agent. As such multi-agent systems can collaborate in order to prioritize the data needed to be downloaded. Furthermore, it is possible to index the data already on the server side, resulting in a reduction of bandwidth and load on remote server. Therefore, all stakeholders benefit from this system (e.g. web search engine, site server, user).

For making mobile crawlers a real success, some security issues have to be addressed (Kumar et al., 2017). These security issues fall into four different categories: host to migrant attacks, migrant to host attacks, migrant to migrant attacks and third-party attacks. Encryption could possibly be one of the ways to tackle these issues (Dixit & Sharma, 2011; Upadhyay et al., 2012; Pahal, 2010)

Table 3 - Extracted trends, challenges and innovations of crawling

Trends	**Innovations**	**Challenges**
*C1. Deep web search - (*Kathuria *et al., 2016)*	1. Onion search (Darpa) - (Kathuria et al., 2016) 2. Deep web search engines - (Kathuria et al., 2016)	1. Opacity between information seeker and provider - (Kathuria et al., 2016) 2. To effectively deal with invisible or deep web - (Kathuria et al., 2016)
C2. Focused crawler - (Kumar et al., 2017)	1. Focused web crawler based on link, text and URL - (Kumar et al., 2017) 2. Focused web crawler based on context, graph, decision tree and DOM - (Kumar et al., 2017) 3. Topic-specific focused crawler - (Kumar et al., 2017) 4. Application-based focused crawler - (Kumar et al., 2017) 5. Semantic crawling-based focused crawler - (Kumar et al., 2017) 6. Learnable focused crawler - (Kumar et al., 2017) 7. Parallel and distributed focused crawler - (Kumar et al., 2017) 8. Language classification-based focused crawler - (Kumar et al., 2017) 9. Location and geographical-based	1. Measuring metrics like recall for focused crawlers due to size - (Kumar et al., 2017)

Trends	Innovations	Challenges
	focused crawler - (Kumar et al., 2017)	
C3. Hidden web crawler - (Kumar et al., 2017)	1. Label value set based retrieval - (Kumar et al., 2017) 2. Form based retrieval - (Kumar et al., 2017) 3. Keyword query-based retrieval - (Kumar et al., 2017) 4. Document object model retrieval - (Kumar et al., 2017) 5. Revisit and incremental approach - (Kumar et al., 2017) 6. Attribute and label extraction approach - (Kumar et al., 2017) 7. Domain or topic specific approach - (Kumar et al., 2017)	1. Explore hidden web - (Kathuria et al., 2016) 2. Dealing with non-HTML search interfaces - (Kumar et al., 2017) 3. Search interface detection and extraction - (Kumar et al., 2017) 4. Generalization of labeled value set - (Kumar et al., 2017) 5. Determining coverage in hidden web crawlers - (Kumar et al., 2017)
C4. Topical crawler - (Kumar et al., 2017)		
C5. Mobile crawler - (Kumar et al., 2017; Morales & Melgar, 2017)	1. Freshness and revisit policy-based mobile crawler - (Kumar et al., 2017) 2. Agent based mobile crawler - (Kumar et al., 2017) 3. Ontology-based mobile crawler - (Kumar et al., 2017) 4. SPHINX, a Java based toolkit - (Kumar et al., 2017)	1. Security issues - (Kumar et al., 2017) 2. The need for user feedback - (Kumar et al., 2017) 3. Dealing with client-side scripts - (Kumar et al., 2017) 4. Server-side push methods - (Kumar et al., 2017)
C6. Forum crawler - (Kumar et al., 2017)		
C. General factors		1. Effectively deal with the web spam i.e. web pages that only exist to mislead search engines as well as the users to certain website - (Kathuria et al., 2016) 2. To deal with the multiple replicas of webpages (Kathuria et al., 2016) 3. To deal with the unstructured or vaguely structured contents - (Kathuria et al., 2016; Kumar et al., 2017) 4. Scale and revisit, the trade-off between freshness and coverage - (Kumar et al., 2017) 5. Crawling multimedia - (Kumar et al., 2017) 6. Open source web crawler - (Kumar et al., 2017) 7. Lack of crawling standards, e.g. bandwidth usage restrictions - (Kumar et al., 2017)

2.2.3.2 Indexing

Standardization

With the transformation from the tradition web search to entity-oriented search (see section 2.2.3.5 for description of entity oriented search), there is a need for standardization of index techniques. As the inverted index of the traditional search engines is the standard, and not efficiently applicable to entity-oriented search, there is a need for an index norm in entity-oriented search. Currently multiple standards and schemas are prevalent to markup different types of information on webpages, making it difficult for a webmaster to choose one.

One attempt to create a norm is Schema.org, a collaborative effort of Google, Bing and Yahoo (Tort & Olivé, 2015) to achieve faster and more relevant search using a structured data markup schema that helps in recognizing entities (people, events, attributes, etcetera). Schema.org is, however, not a standard body like W3C, but a website providing the schema and markup supported by major search engines. The on-page markup helps the search engines to better understand the information of web pages and produce richer search results. Therefore it is beneficial for all stakeholders (webmaster, search engine, user).

Culpepper et al., (2018) describes the challenge and need to come up with a generic enough representation of heterogenous data sources and formats. Efforts are being made to this end in the form of data integration and common vocabularies such as schema.org and Dublin Core. However, once the data has been represented in a common format, one must think about suitable ways of querying it and interfaces to expose to users.

New ways of indexing

The data to be indexed is commonly dynamic and heterogenous and the challenge is to find a one fits all solution to index the data, however this is nearly impossible to the various applications of IR. Therefore, novel ways of indexing must be developed to cope with the various applications and data forms. As Kumar et al. (2017) describes the need for new index structures for deep web applications. Moreover Culpepper et al. (2018) also mentions the need for new index structures as an index structure able to accommodate downstream fairness and transparency operations. Furthermore, the introduction of knowledge graphs demands for new index approaches. For efficiency increases, index structures can be revisited. Lastly, the expressed need for automatic ways to create optimized index structures for a given Multi-stage Search Systems (MSS).

Smaller, faster, better

Culpepper et al. (2018) describes various hardware advancements providing new opportunities for IR, e.g. the vector processor and FPGA. Furthermore, the advancement of GPU's, available on all devices (smartphone, laptop, tablet) enabling off-load work from the data center and onto the user's device.

Table 4 - Extracted trends, challenges and innovations of indexing

Trends	Innovations	Challenges
*I1. Entity oriented search, using knowledge bases - (*Kathuria *et al., 2016)*	1. Schema.org for standardization of storing entities, relations among them and actions for web page authors - (Kathuria et al., 2017)	1. Standardization of storing of information about entities - (Kathuria et al., 2017)
I2. Hidden web search - (Kumar et al., 2017)		1. Indexing technique for hidden data - (Kumar et al., 2017)
I3. Smaller, Faster, Better - (Culpepper et al., 2018)		1. Efficiency and management support systems - (Culpepper et al., 2018) 2. New ways of indexing - (Culpepper et al., 2018) 3. Challenging the current setup - (Culpepper et al., 2018) 4. Change of hardware setup - (Culpepper et al., 2018)

2.2.3.3 Ranking

Learn to rank

The advancement of machine learning has led to breakthroughs in various domains, e.g. computer vision, NLP. However, a naive utilization of deep learning approaches developed for other areas may not be a good fit for IR problems (Culpepper et al., 2018). Additionally, the high existing baselines in IR, by years of tweaking conventional methods, make it more difficult for machine learning to compete (Culpepper et al., 2018). Nonetheless, there is a trend in research concerning machine learning for IR and has promising directions, since deep learning methods are highly versatile and adaptable, allowing usage for multimodal data inputs, heterogenous data views and can be trained jointly over multiple tasks simultaneously (Culpepper et al., 2018).

Reinforcement learning

Reinforcement learning is an area of machine learning of which the prediction model ought to maximize some notion of cumulative reward. This is a continuous process in which the reward could be an action of the user interacting with the system. Reinforcement learning is considered as one of three machine learning paradigms, together with supervised learning and unsupervised learning

Since information access is often situated in an interactive search task, the ability to perform intelligent sequential decision-making is a fundamental — yet underexplored — area of information retrieval (Culpepper et al., 2018). Recent advances in reinforcement learning suggests that the technique is ready to be applied to IR. This would require, however, substantial research in various domains (Culpepper et al., 2018). Reinforcement learning techniques can be used for combining multiple feature modalities or even multiple learning strategies such as active learning and exploration/exploitation (Calumby et al., 2017). Therefore, the importance to further explore learning boosting alternatives such as diversity promotion for handling ambiguous, multi-intent and underspecified queries.

Machine reasoning

Machine reasoning is the generation of conclusions from the available knowledge, e.g. inference engines, theorem provers, classifiers. There has been significant progress on machine reasoning in the context of tasks such as text understanding and reasoning, and dialogue state tracking, focused on memory architectures for selectively capturing dialogue/document context as needed for long distance inference (Culpepper et al., 2018). Attempts are being made to integrate knowledge graphs in NLP, with possible applications as tracking “state” in multi-turn IR, smoothing document- and term-level, the ability to make predictions across a session, interpreting complex search requests, supporting question answering and implementing domain specific IR (Culpepper et al., 2018). Moreover, Morales & Melgar (2017) mentions the use of ontologies to enhance machine reasoning.

Data representation

Since deep learning techniques work with multi-modal input (e.g. text, images, video's), the representation is not directly suitable for IR (Culpepper et al., 2018). Therefore, new ways of representations need to be developed specific for IR. Additionally extending the input by connecting other sources of information could be valuable in more effective ways than a hard filter (Culpepper et al., 2018).

Data efficiency

In the academic community, high quality data is limited and vital for training the machine learning model. Resulting in various challenges to overcome lack thereof in order to increase the effectivity and efficiency of current machine learning models. There are vast amounts of useable data in the industry, however this would incur substantial risks to intellectual property and user privacy.

One of the challenges that follows out of limited data is to train robust and accurate models using small collections. Some call this a cold start, as the model starts with very limited data. To avoid this problem, Calumby et al. (2016) mentions the possibility of using new learning algorithms like manifold learning, improving multi-modal fusion methods, and making better use of implicit feedback. Another solution could be to automatically generate training data from scavenged public data, developing new techniques to expand current labeled datasets or community efforts to build labeled datasets that are a magnitude larger than existing ones (Culpepper et al., 2018). Lastly using trained models for new tasks to avoid retraining a new model.

Another challenge is to deal with noisy and incomplete data, making it hard to train a model on. The lack of qualitative data demands for a more robust global model effective for data-poor domains. One of the focus points is to be more data efficient to use robust models in a fast fashion.

End-to-end learning

Certain complex IR problems might be learnable in an end-to-end manner, empowered by the neural network revolution in IR (Culpepper et al., 2018). This would include the complete process from query to the presentation of the results, or from query to the output of the answer, all realized using labeled data.

Explainability

By using machine learning models, it is a challenge to make the model explainable. However, it is a necessary part for model improvement. Current error analysis techniques and model explainability need to be improved, in order to find errors in training sets that cause problems on output levels, to understand which models work better in which scenarios and to justify the results to users and system designers (Culpepper et al., 2018).

Semantic query processing

Semantic query processing is the advancement in understanding the user's query, instead of simply connecting the user's keywords to indexed words of documents. The underlying concept of semantics is based upon the semantic similarity being taken over documents, words, terms, sentences and entities. One example of the query processing is Google's RankBrain, which optimizes the user's query by using machine learning. Other examples include Powerset, Hakia, Google's Hummingbird (Napgal et al., 2016) and Word2Vec (Altinel & Ganiz, 2018), all focused on understanding the query beyond its keywords. One major difference with key-based search is semantic methods take relational words into account which are vital to the meaning of the query, e.g. by, for, in.

Guided user interface

Culpepper et al. (2018) mentions research is needed to investigate ways that search systems can provide information (beyond ranked lists and underlying document) that will aid the searcher in evaluating and contextualizing results. Developing new user interface tools increasing the user's understanding, learning and ability to make more informed decisions. As such Calumby et al. (2016) mentions the possibility to interact with the user interface by explanation. Enabling a new way of interaction. Furthermore, the possibility to customize the user interface by downloadable plug-ins (Morales & Melgar 2017). Although, introducing new search interfaces is not an easy task since it poses extra onus on users by relearning the interface (Culpepper et al., 2018).

Table 5 - Extracted trends, challenges and innovations of ranking

Trends	Innovations	Challenges
*R1. Semantic query understanding - (*Kathuria *et al., 2016)*	1. NLP for query understanding - (Kathuria et al., 2016) 2. SCINET, new semantic search engine - (Kathuria et al., 2016) 3. Option based search - (Kathuria et al., 2016) 4. Machine learning for query understanding, Rankbrain - (Kathuria et al., 2016) 5. Word2vec - (Altinel & Ganiz, 2018)	1. Accurate semantic query understanding - (Kathuria et al., 2016)
R2. Learnable information retrieval - (Culpepper et al., 2018)	1. End-to-End learning - (Culpepper et al., 2018) 2. Reasoning components - (Morales & Melgar, 2017) 3. Deep learning-based text classification - (Altinel & Ganiz, 2018)	1. Achieving good accuracy with a few training samples - (Calumby et al., 2016) 2. Reducing labelling effort and attenuating the cold start problem - (Calumby et al., 2016) 3. Differentiating positive and negative sample treatments on the learning process for their different representativeness in relation to the real data distribution - (Calumby et al., 2016) 4. Integrating advanced procedures for handling complex queries - (Calumby et al., 2016) 5. Using reinforcement learning methods for combining multiple feature modalities or even multiple learning strategies, such as active learning and exploration/exploitation - (Calumby et al., 2016; Culpepper et al., 2018) 6. Accessing massive amounts of labelled data - (Altinel & Ganiz, 2018) 7. Model interpretation - (Altinel & Ganiz, 2018) 8. Data efficiency, limited academic data - (Culpepper et al., 2018) 9. Data representation - (Culpepper et al., 2018) 10. Machine reasoning - (Culpepper et al., 2018) 11. Explainability - (Culpepper et al., 2018) 12. Delivering models which can compete with the high existing baselines - (Culpepper et al., 2018) 13. The development of effectiveness measures that are more informative and better suited for learn-to rank methods - (Calumby et al., 2016)

Trends	Innovations	Challenges
R3. Question answering (focused on multi-modality) - (Calumby et al., 2016)		1. To go beyond the list of possible relevant web pages and to focus on providing an exact answer - (Kathuria et al., 2016)
R4. Interaction by explanation, explain how the results were generated - (Calumby et al., 2016)		1. Optimal user interface - (Calumby et al., 2016) 2. Analysing user behaviour impacts on search tasks - (Calumby et al., 2016)
R5. Exploring additional sources to improve accuracy, additional image collections or knowledge sources - (Calumby et al., 2016)		1. Tags and comments exploration for improving the system - (Calumby et al., 2016)
R6. Social interaction for system's optimization through collaborative filtering - (Calumby et al., 2016)		1. Leveraging long-term learning and collaborative retrieval for effectiveness and efficiency improvement - (Calumby et al., 2016)
R. General factors		1. Avoiding memory recall/smart cache management - (Kathuria et al., 2016) 2. Convert misspelled queries in relevant results - (Shen et al., 2015)

2.2.3.4 Evaluation

New evaluation metrics

For most applications the traditional evaluation metrics as precision, recall and F1-measure are used, however due to the new advancements in IR and the desire for more accurate metrics for system evaluation, new evaluation metrics are needed. As Culpepper et al. (2018) mentions the current evaluation metrics are focused on naïve user interactions. There is a growing concern, metrics based on such interactions are not capturing the important information from users. Additionally, new search interfaces as smartwatches and conversational search engines don't allow for traditional interaction metrics. Therefore, it is necessary to develop more sophisticated logging methods, a richer understanding of context and user session, and technologies allowing eye tracking that could be used to determine how users are reacting and benefitting from an online system (Culpepper et al., 2018). Calumby et al. (2016) adds to this the need to develop better log analysis methods, click models and user models considering reformulation understanding, stopping criteria and erroneous feedback simulation.

Evaluation metrics are normally used to evaluate a specific part of the system, however, according to Culpepper et al. (2018), there is a need for more insightful and richer explanation of IR system performance, which does not only allow us to account for why we observe given performance: e.g. failure analysis, but also how the separate parts perform, the interaction between them and how factors external to the system are impacting the overall performance.

Furthermore, the need for evaluation metrics tailored to learn-to-rank strategies containing more informative metrics (Calumby et al., 2016). As conventional methods focus more on the system's performance and not primarily on whether the user is met in its informational needs. This becomes increasingly important with entity-oriented search, since the comprehension of queries and documents increases significantly. The user's experience and satisfaction with this information seeking process is influenced by many aspects, including query format, performance of the search system and presentation of the results (Elbedweihy et al., 2015). Therefore, a more hybrid approach should be developed to satisfy both (system and user) needs. According to Elbedweihy et al. (2015), these evaluation initiatives to tackle these issues should be administered by a well-respected organization, which would create and distribute datasets, organize campaigns (both system- and user-oriented) and report results to the community.

Counterfactual analysis

Counterfactual analysis is a tool from casual reasoning that allows the study of how users respond to a change in the retrieval system. Drawing on the system's interaction log, one can offline "replay" the user's session with a slightly different IR-system, reweighting their interactions according to their likelihood of being recorded under the changed system (Culpepper et al., 2018). Counterfactual analysis is therefore a hybrid between online and offline, by gathering online data and manipulating it in an offline fashion.

Conducting real-user studies

Conducting real-user studies always have been hard tasks and often neglected. Gathering subjects, logistics and time-intensive experiments require a substantial budget, however Calumby et al. 2016) stresses the importance of a centralized organization conducting such studies. As comparing real user studies with lab-based analysis does not only allow for an accurate system's performance, but also validate current modelling approaches. Additionally, using ordinal scales instead of binary relevance scores, allowing a more real-life approach (Elbedweihy et al., 2015).

Table 6 - Extracted trends, challenges and innovations of evaluation

Trends	**Innovations**	**Challenges**
*E1. Generated information objects - (*Kathuria *et al., 2016)*	1. NTCIR one-click - (Culpepper et al., 2018) 2. TREC complex answer retrieval track - (Culpepper et al., 2018)	1. Shift to more goal/user-oriented way of evaluation metric - (Culpepper et al., 2018; Elbedweihy et al., 2015)
E2. Mainly using precision, recall and F1-measure - (Kumar et al., 2017; Soares & Parreiras, 2018; Shen et al., 2015)	1. Using F1 and accuracy for evaluating QA systems - (Soares & Parreiras, 2018) 2. Using precision, accuracy and F1for evaluating entity linking - (Shen et al., 2015)	
E3. Crowdsourcing for labelling - (Elbedweihy et al., 2015)		1. Gathering relevance assessments efficiently - (Elbedweihy et al., 2015)
E4. Personalized search - (Tamine & Daoud, 2018; Culpepper et al., 2018)	1. TREC contextual suggestion track - (Tamine & Daoud, 2018) 2. NTCIR12 lifelog semantic access task - (Tamine & Daoud, 2018)	1. More advanced user modelling (Calumby et al., 2016) 2. Context dynamicity - (Tamine & Daoud, 2018) 3. Data privacy, privacy preserving evaluation guidelines - (Tamine & Daoud, 2018) 4. Lack of data in multi-stage search systems such as query reformulation and contextual knowledge in evaluating a realistic dynamic IR framework - (Tamine & Daoud, 2018) 5. Evaluation of large-scale outcomes using limited samples of safe data - (Tamine & Daoud, 2018)
E5. Living lab evaluation of dynamic IR - (Tamine & Daoud, 2018)		1. Conducting real-user studies - (Calumby et al., 2016; Elbedweihy et al., 2015) 2. Evaluating the performance with real-life settings - (Calumby et al., 2016) 3. Legal and ethical issues - (Tamine & Daoud, 2018) 4. Reproducibility - (Tamine & Daoud, 2018)
E. General factors	1. Explicit feedback on an ordinal scale from 1-7 - (Elbedweihy et al., 2015)	1. Counterfactual analysis - (Culpepper et al., 2018; Elbedweihy et al., 2015) 2. Define the axiometrics of online evaluation metrics - (Culpepper et al., 2018) 3. New online metrics from new interactions -

Trends	Innovations	Challenges
		(Culpepper et al., 2018; Calumby et al., 2016; Elbedweihy et al., 2015) 4. Decomposable ways of predicting the performance of IR systems - (Culpepper et al., 2018; Elbedweihy et al., 2015) 5. Developing better benchmarks - (Calumby et al., 2016) 6. The proposal of better interactivity cost functions to evaluate - (Calumby et al., 2016) 7. Using grade relevance assessments as a way to improve ground-truth quality and maximize feedback information - (Calumby et al., 2016) 8. Evaluating search strategies and their user effort on retrieval sessions - (Calumby et al., 2016) 9. The development of better log analysis methods - (Calumby et al., 2016) 10. Stopping criteria and erroneous feedback simulation - (Calumby et al., 2016) 11. The need for an evaluation initiative administered by a well-respected organization for distributing datasets, organise campaigns and report results - (Elbedweihy et al., 2015) 12. Using various datasets to evaluate the system - (Elbedweihy et al., 2015) 13. Updated datasets - (Elbedweihy et al., 2015) 14. Using different levels of difficulty of queries when evaluating the system - (Elbedweihy et al., 2015) 15. Repeatability and reliability of evaluations - (Elbedweihy et al., 2015) 16. Reusability of data - (Elbedweihy et al., 2015)

2.2.3.5 General

Entity oriented search

Semantic and entity-oriented search go beyond the traditional keyword-based search. It introduces a new level of consciousness of queries and documents, enabling more relevant search. Search has transformed from finding words “strings”: to finding entities “things”, due to this fact search engines have the ability to directly answer queries if it contains a question. Additionally, the search engine can show multi-modal entities related to the search query.

Entity-oriented search is based upon pre-defined relationships between entities. As such the webpages are extracted, followed by constructing a structured database (the index) linked with the web of concepts, containing the relationship and categories of entities. Resulting in a better understanding of the underlying relationships between entities, properties and documents. Elbedweihy et al. (2015), however, mentions not only to use sources as triples (web of concepts), but additionally semantic data from different domains on the web of data. The web of concepts consists of a knowledge graph, examples of these are Google's Knowledge Graph and Microsoft's Satori (Kathuria et al., 2016).

Some of the challenges faced of entity-oriented search is the continuous existence and maintenance of ontologies (Morales & Melgar, 2017) and the representations thereof (Culpepper et al., 2018). Additionally, the challenges it brings to accurately link entities, especially in the case of name variations and entity ambiguity (Shen et al., 2015).

Fairness, accountability, confidentiality and transparency

Web search engines are about connecting information to information seekers; however, this process is not eliminated from human intervention, resulting in a potential bias. Since people rely more and more on web search engines, it becomes increasingly important that the system should be fair (e.g. it does not discriminate across people), accountable (e.g. a system should be reliable and able to justify its actions), confidential (e.g. confidential information must be inaccessible from people without access to this information), and transparent (e.g. a system should be able to explain why results are returned). These problems are related with the fundamentals of web search engines, including information representation, information reliability, information retrievability and access, evaluation and others. While, traditionally, the IR community has been focused on building systems that support a variety of applications and needs; it is becoming imperative that we focus as much on the human, social, and economic impact of these systems as we do on the underlying algorithms and systems (Culpepper et al., 2018). Therefore, one of the challenges is to eliminate or reduce bias of the system (Moulahi et al., 2015; Culpepper et al., 2018)

IR for supporting knowledge goals and decision-making

Web search engines should support complex evolving or long-term (Kathuria et al., 2016) information seeking goals such as acquiring broad knowledge, either for the user's sake or to make an informed decision (Culpepper et al., 2018). Such support will require understanding what information is needed to accomplish the goal, scaffolding search sessions toward the goal, providing broader context as information is gathered, identifying and flagging misleading or confusing information, and compensating for bias in both information and users (Culpepper et al., 2018). For realizing this, advancements on fundamental aspects of the web search engines should be made as algorithms, interfaces and evaluation methods (Elbedweihy et al., 2015) that support these goals. To support the user in these processes, a growing understanding of cognitive processes of humans is needed.

Currently the knowledge goals and decision-making IR systems fall short and the users need more support in this area (Culpepper et al., 2018). To illustrate the shortcomings, one could observe the 2016 presidential election cycle in US politics, where a large contribution of the misinformation was due to search systems (and related algorithms, such as ranking algorithms employed by social media systems). Search systems should incorporate ways to aid the searchers in evaluating and contextualizing search results, additionally they should guide the user through a learning of decision-making process, for example by using ontologies (Morales & Melgar., 2017). Lastly, Kathuria et al. (2016) mentions the challenge to enable search during a period in which new information is directly linked to the information seeker to guide the user in its knowledge needs.

Personalized search

Information created by, connected to, or consumed by an individual resides across a great number of separate information silos: personal devices (e.g. laptops, smartphones, smartwatches); the web; personal or enterprise file shares; messaging systems and social media; and systems from external parties including medical doctors, bank records, employer and government records, and many others (Culpepper et al., 2018). In order to access this heterogenous and cross-device data we rely on the underlying systems to determine the location of the data and permissions to obtain it. Web search engines become increasingly personalized, enabling specialized information for the information seeker’s needs.

One of the challenges, however, is to understand and anticipate the user’s needs. Other challenges are the identification and abstraction, task representation, personalization, privacy, security and trust, architecture and application, constructing a personal knowledge graph, linking and disambiguation across silos (Culpepper et al., 2018). Kathuria et al. (2016) adds on demand anticipation, customization and personalization to this list.

Smaller, faster and better

The expansion of the internet is diverse and rapid, constantly new webpages are edited and created. Search engines should constantly be scavenging the web guaranteeing coverage, effectivity and efficiency. In addition, dealing with noisy, low quality and contradictory contents continuously being uploaded to the web (Kathuria et al., 2016) and handling scale (Morales & Melgar, 2017; Calumby et al., 2015). Therefore, the need is to constantly innovate making smaller, faster and better web search engines. The current trend is to incorporate new (ranking) algorithms as machine learning to increase the effectivity of the web search engine. However, increasing its efficiency by using machine learning is still one of the current challenges (Culpepper et al., 2018). New architectures are tested and pushing the limits of web search engines.

Question answering paradigm

Due to the advancements of NLP and entity recognition, queries containing a question can be answered directly. This technology works best on questions based on real world relationships as: how many citizens does Amsterdam have? In order to retrieve a relevant answer, the system must first recognize the query as a question, categorize and transform the question in a query, secondly generating a list of potential answers by using the keywords in the query and finally select and present the most relevant answer (Malik et al., 2013; Bhoir and Potey, 2014; Neves and Leser, 2015).

One of the innovations in this field is the QA system of IBM Watson (Soares & Parreiras, 2018; Shen et al., 2015). A possible direction for the future could be the multi-modality of QA systems, e.g. answering a query by displaying an image, video, 3D model or audio file containing a snippet of the original file displaying the answer.

Multi-stage information seeking/guided user interface

Searching can vary from a simple lookup to multifaceted and complex searches, during sessions spanning minutes, seconds, hours or even days (Huurdeman & Kamps, 2014). The more complex search tasks might take multiple phases and involve learning and construction. Therefore a one size fits all solution might not be an optimal solution. The user’s information needs might not be satisfied in the case of more complex search tasks and therefore Huurdeman & Kamps (2014) describes the need for a system taking multi-stage into account.

During a complex search session, the redundancy increases, and the uniqueness reduces, due to the increased knowledge of the subject by the user. Therefore users should not be overwhelmed in the beginning and given more complex information later in the search session. These requirements can be embodied in the user interface for controlling the feature groups consisting out of input (search box, categories, clusters, faceted metadata, social metadata), control (related searches, corrections, sorting, filters, grouping), informational (result display, text snippets, deep links, thumb-nails, immediate feedback, visualizations), and personalizable (recent searches, item tray)(Ruthven & Kelly, 2011).

Table 7 - Extracted trends, challenges and innovations of web search engines

Trends	Innovations	Challenges
*WSE1. Continuous and social search - (*Kathuria *et al., 2016)*		1. Search during a period - (Kathuria et al., 2016)
WSE2. Personalized information retrieval - (Culpepper et al., 2018; Jayanthi & Rathi, 2014)		1. Ranking and retrieval - (Culpepper et al., 2018) 2. Computing over aggregate personal data - (Culpepper et al., 2018) 3. Learning to rank from personalized data - (Culpepper et al., 2018) 4. To offer on demand anticipation, customization and personalization - (Kathuria et al., 2016) 5. Understanding the needs of the information seeker - (Culpepper et al., 2018) 6. Task representation, identification & abstraction - (Culpepper et al., 2018) 7. Index and schema - (Culpepper et al., 2018) 8. Privacy, security and trust - (Culpepper et al., 2018) 9. Linking and disambiguation across different data silos - (Culpepper et al., 2018) 10. Architectures and applications - (Culpepper et al., 2018)

Trends	Innovations	Challenges
		11. User modelling and representation methods - (Jayanthi & Rathi, 2014) 12. Building a complex user profile with minimal user interaction - (Jayanthi & Rathi, 2014) 13. Better accuracy of interest prediction - (Jayanthi & Rathi, 2014) 14. Building profiles of ubiquitous nature which sense the environmental contexts automatically and profiles with social inputs such as common interest group, social networks, etc - (Jayanthi & Rathi, 2014)
WSE3. Fairness, accountability, confidentiality and transparency in IR - (Culpepper et al., 2018)		1. Reducing bias - (Calumby et al., 2016; Moulahi et al., 2015; Culpepper et al., 2018) 2. Fair information retrieval - (Culpepper et al., 2018) 3. Accountable information retrieval - (Culpepper et al., 2018) 4. Confidential information retrieval - (Culpepper et al., 2018) 5. Transparent information retrieval - (Culpepper et al., 2018)
WSE4. Time and stream analysis - (Moulahi et al., 2015)	1. Google news timeline - (Moulahi et al., 2015) 2. Newsmap - (Moulahi et al., 2015)	1. Using various data sources and evaluation there of - (Moulahi et al., 2015) 2. Presentation of time sensitive results - (Moulahi et al., 2015) 3. Determining the ranking model - (Moulahi et al., 2015) 4. Novel methods for more time sensitive search results - (Moulahi et al., 2015) 5. Presenting results from various times of interest - (Moulahi et al., 2015) 6. Comparing retrieval models due to different architectures - (Moulahi et al., 2015) 7. Operational and shared data repository - (Moulahi et al., 2015) 8. Topic detection and tracking - (Moulahi et al., 2015) 9. Novelty detection - (Moulahi et al., 2015)

Trends	Innovations	Challenges
WSE5. Conversational information seeking - (Culpepper et al., 2018)		1. Personalization - (Culpepper et al., 2018) 2. Finding information - (Culpepper et al., 2018) 3. Engagement with the information seeker - (Culpepper et al., 2018) 4. Domain generality and specificity - (Culpepper et al., 2018) 5. Failure modes - (Culpepper et al., 2018) 6. Multi-modal conversations - (Culpepper et al., 2018) 7. Cross-device conversations - (Culpepper et al., 2018) 8. Collaborative information seeking - (Culpepper et al., 2018) 9. Evaluation of conversational IR systems - (Culpepper et al., 2018)
WSE6. IR for supporting knowledge goals and decision making - (Culpepper et al., 2018)		1. Understanding cognitive aspects of users that are relevant to their information seeking - (Culpepper et al., 2018) 2. Investigating ways how search systems can provide information (beyond ranked lists and underlying documents) that will aid the searcher in evaluating and contextualizing search results - (Culpepper et al., 2018) 3. Exploring ways that allow search systems to help users move through a learning or decision-making process - (Culpepper et al., 2018) 4. Overcoming challenges in evaluating how well systems support users in learning and decision making - (Culpepper et al., 2018)
WSE7. Generated information objects(GIO) - (Culpepper et al., 2018)		1. Temporal summarization - (Moulahi et al., 2015) 2. Knowledge graph representation in GIO's - (Culpepper et al., 2018) 3. Adversarial GIO's - (Culpepper et al., 2018) 4. Merging of heterogeneous GIO's - (Culpepper et al., 2018) 5. Deriving explanations from GIO's - (Culpepper et al., 2018) 6. Context and personalization - (Culpepper et al., 2018)

Trends	Innovations	Challenges
WSE8. Next-gen efficiency challenges: smaller, faster, better - (Culpepper et al., 2018)		1. Using machine learning for efficiency improvements - (Culpepper et al., 2018) 2. New applications for IR - (Culpepper et al., 2018)
WSE9. Changes in hardware: CPU with vector processors, FPGA's, GPGPU - (Culpepper et al., 2018)		
WSE10. IR for an IOT world - (Culpepper et al., 2018)		1. Integrate the "syntactic" level of rough sensor data within a broader concept of information - (Culpepper et al., 2018)
WSE11. Cross device search - (Culpepper et al., 2018)		1. Result presentation on mobile devices (dynamic user exploration) - (Moulahi et al., 2015)
WSE12. Question answering systems (focused on multi-modality) - (Soares & Parreiras., 2018; Calumby et al., 2016)	1. POS tagging and named entity recognition for QA systems - (Soares & Parreiras., 2018) 2. IBM Watson - (Soares & Parreiras, 2018; Shen et al., 2015)	1. To go beyond the list of possible relevant web pages and to focus on providing an exact answer - (Kathuria et al., 2016)
WSE13. Multi-stage information seeking/guided user interface - (Huurdeman & Kamps., 2014; Morales & Melgar, 2017)		1. To support simultaneously the generic overview of topics and enabling specialists groups to drill down to their exclusively relevant items. - (Kathuria et al., 2016) 2. Lack of data in multi-stage search systems such as query reformulation and contextual knowledge in evaluating a realistic dynamic IR framework - (Tamine & Daoud, 2018) 3. Polypresentation in dynamic IR evaluation frameworks - (Tamine & Daoud, 2018) 4. Use of ontologies for suggesting or guiding the user's search - (Morales & Melgar, 2017)
WSE14. Entity oriented search - (Asikri et al., 2017; Elbedweihy et al.,	1. Semantic web mining using machine learning techniques to create and improve the semantic web - (Asikri et al., 2017) 2. Web mining - (Asikri et al., 2017)	1. Knowledge graph representation - (Culpepper et al., 2018) 2. Precision on results - (Morales & Melgar, 2017) 3. Existence and maintenance of

Trends	Innovations	Challenges
2015; Jayanthi & Rathi, 2014; Altinel & Ganiz, 2018; Shen et al., 2015; Kathuria et al., 2016)	3. Swoogle - (Elbedweihy et al., 2015) 4. Sindice - (Elbedweihy et al., 2015) 5. NLP-reduce - (Elbedweihy et al., 2015) 6. Querix - (Elbedweihy et al., 2015) 7. PowerAqua - (Elbedweihy et al., 2015) 8. Freya - (Elbedweihy et al., 2015) 9. K-search - (Elbedweihy et al., 2015) 10. Smeagol - (Elbedweihy et al., 2015) 11. Sig.ma - (Elbedweihy et al., 2015) 12. VisisNav - (Elbedweihy et al., 2015) 13. Wordnet - (Jayanthi & Rathi, 2014; Altinel & Ganiz, 2018) 14. DBpedia, a multilingual knowledge base constructed by extracting Wikipedia's 4 million entities - (Elbedweihy et al., 2015; Shen et al., 2015) 15. Yago, an open knowledge base with 10 million entities - (Shen et al., 2015) 16. Freebase, a large online knowledge base containg 43 million entities - (Shen et al., 2015) 17. Knowitall - (Shen et al., 2015) 18. ReadTheWeb - (Shen et al., 2015) 19. Probase - (Shen et al., 2015) 20. Wikilinks corpus, a large scale entity linking dataset with 40 million mentions over 10 million web pages - (Shen et al., 2015; Jayanthi & Rathi, 2014; Altinel & Ganiz, 2018) 21. Dog4Dag, an ontology generation plugin for protege 4.1 and OBOEdit - (Asikri et al., 2017) 22. Ontostudio - (Asikri et al., 2017) 23. OWL ontology languages - (Asikri et al., 2017) 24. Wikitory - (Altinel & Ganiz, 2018) 25. Satori, knowledge graph - (Kathuria et al., 2016) 26. Knowledge graphs - (Kathuria et al., 2016) 27. Using (Domain) Ontologies - (Morales & Melgar, 2017) 28. Extended learnability - (Khadija et al., 2015) 29. Entity recognition tools: Stanford NER, OpenNLP, LingPipe - (Shen et al., 2015) 30. Domain knowledge based text classification - (Altinel & Ganiz,	ontologies - (Morales & Melgar, 2017) 4. Usability - (Morales & Melgar, 2017) 5. Evolution of knowledge bases as new documents appear - (Morales & Melgar, 2017) 6. Turning unstructured data into machine-understandable data using semantic web tools - (Asikri et al., 2017) 7. Scalability of the algorithms due to data size to mine semantic web content and structure mining (Asikri et al., 2017) 8. Distribution of the data instead of one central database (Asikri et al., 2017) 9. Entity oriented search - (Kathuria et al., 2016) 10. Using additional data sources for the semantic web - (Elbedweihy et al., 2105) 11. Name variations and entity ambiguity in entity linking - (Shen et al., 2015) 12. Combining named entity recognition and entity linking - (Shen et al., 2015) 13. Record linkage also called duplicate detection, entity matching and referend conciliation - (Shen et al., 2015) 14. Cope with un-linkable entities - (Shen et al., 2015) 15. Entity linking for different sorts of textual formats, e.g. tables, tweets - (Shen et al., 2015) 16. Availability of a knowledge base for a specific language - (Altinel & Ganiz, 2018) 17. Processing complexity of a large external knowledge base - (Altinel & Ganiz, 2018) 18. Complexity of computations to extract latent semantics - (Altinel & Ganiz, 2018) 19. Standardization of evaluation methods - (Elbedweihy et al., 2015)

Trends	Innovations	Challenges
	2018) 31. Corpus based text classification - (Altinel & Ganiz, 2018) 32. Word/character sequence enhanced approaches text classification - (Altinel & Ganiz, 2018) 33. Linguistic enriched approaches text classification - (Altinel & Ganiz, 2018)	
WSE. General factors		1. To effectively deal with noisy, low quality, unreliable and contradictory contents continuously being uploaded on the web - (Kathuria et al., 2016; Morales & Melgar, 2017) 1. Lack of sentimental search engine - (Kumar et al., 2017) 2. Use of natural language - (Morales & Melgar, 2017) 3. Handle large amounts of data - (Morales & Melgar, 2017; Calumby et al., 2016)

2.2.4 Limitations

Due to the limited time scope a fraction of the available literature has been used for this literature review. However, the objective was to use summarizing papers in order to explore as much of the knowledge in a broad spectrum. This introduces however a latency in the research found, since experiments first must be conducted before a summarizing paper even can be made. Additionally, only papers were used in a timeframe of five years, which could possibly lead to some gaps in the research, however it is unlikely due to the extended scope of the summarizing papers. Since some of the papers contained most of the found factors, therefore the final list could be slightly biased, since not all factors are validated by multiple sources. Lastly, the extracted information was forced in the format of trends, innovations and challenges, in addition to five categories. This introduces an interpretation processing step, possibly leading to incorrectness of the interpretation.

3 Methodology

This chapter lays the foundation of the research by describing the methodology. Moreover, it argues why the specific methodology was chosen for the research objective. As such Section 3.1 discusses the motivation for the chosen methodology, followed by Section 3.2 describing the research methodology, and how it has been implemented in this research.

3.1 Motivation

For the validation, supplementation and filtration of the found list of trends, innovations and challenges in the literature, expert knowledge was used. Moreover, to evaluate the relationship between industry and academics, previously not described in the field of web search engines.

The methodology used for this is the Delphi methodology motivated by the desired consensus between industry- and academic-experts. Reaching consensus in an area of uncertainty or lack of empirical evidence is the main advantage of this methodology (Delbecq et al. 1975, Dawson & Barker 1995, Murphy et al. 1998).

The method can be applied to problems that do not lend themselves to precise analytical techniques but rather could benefit from the subjective judgments of individuals on a collective basis (Adler & Ziglio, 1996) and to focus their collective human intelligence on the problem at hand (Linstone & Turloff, 1975). It is consequently used for forecasting and issue identification/prioritization and applicable to these research questions.

In addition, it has been described as a quick (Everett 1993), cheap (Jones et al. 1992) and relatively efficient way to combine the knowledge and abilities of a group of experts (Lindeman 1975). The technique has been widely used in business, industry and health care research with various methodological interpretations and modifications (Powell, 2002*)*. Skulmoski et al., (2007) gives a list of some applications in information system research for graduate research and the examples share certain characteristics with the research at hand.

The creative process benefits by providing the participants anonymity and a structured approach (Rowe et al., 1991*)*. However, the lack of accountability could lead to hasty jumps to conclusions and withdrawal of information as well, although the same holds true for other data collection methods. The risk here is minimized by using experts and the continuation of the research.

Because the technique is intended to correct for lack of conclusive data by drawing on, and sharing, the knowledge and experience of experts (Fink et al., 1991), it should not be open to the same validation criteria as hard science. In addition, Murphy et al. (1998) notes that the Delphi methodology, and other consensus development methodologies should not be viewed as a scientific method of creating new knowledge, but rather a process for making the best use of the available data, may that be scientific or the collective knowledge of experts.

3.2 The Delphi methodology

The Delphi method is an iterative process to collect and distill the anonymous judgements of experts using a series of data collection and analysis techniques interspersed with feedback (Skulmoski, 2007). The anonymity makes the participants feel free to express their judgements without the social pressure from others in the group. Decisions are evaluated on their merit, rather than who has proposed the idea (Skulmoski, 2007). Iteration is to refine the participant's view in the light of the progress of the groups work. The controlled feedback is the tool to enable refinement of the views based on feedback of other participants. Since the output data will be quantifiable, it opens the opportunity for quantitative analysis and interpretation of the data.

3.2.1 The process

The Delphi technique is in essence a series of sequential questionnaires or 'rounds', interspersed by controlled feedback, that seek to gain the most reliable consensus of opinion of a group of experts (Linstone & Turoff 1975). Three expert panels will be established with a group size of 10-18 persons per panel. These expert panels consist out of academic experts, industry experts (software developers, business developers, etcetera) and finally customers of web search engines (site search appliances: Segona, Mindbreeze and Google Search). This will enable to distill not only the identification of trends, innovations and challenges, but also the relationship between the groups. The data collection will consist out of three surveys in order to reach consensus. Every survey will be pilot tested with the supervisors of this research to guarantee quality and completeness. The invitations to the survey will be sent through email and data will be captured through Google Forms [1]and Kwiksurveys[2]. This data can then by analyzed by using the statistical programming language R.

Expert selection & process of sending out the surveys

Experts were selected for three panels: academic experts, industry experts and clients. The academic experts were chosen based on their presence in the papers read for the literature review, qualifying them as expert in the field of web search engines. Out of this, an initial list of 72 experts were identified and were reached out to by email. After they did not respond within a week of the email containing the invitation for the survey, a reminder was sent to increase the chances of participation. After a week of still no response, a final attempt was made by supervisor Alexandra Lopes requesting collaboration. After this request was denied the experts were removed from the list and the respondents continued to the following round.

Industry experts were chosen first internally within Incentro, based on their experience with the search solutions the company delivers. Additionally, people outside of Incentro were reached out to, working for prestigious companies who deliver search solutions, e.g. Elastic, Microsoft and Google. Resulting in a total list of 23 industry experts of which 13 experts of Incentro. They were reached out to by a combination of using the message function of LinkedIn, the internal Google chat of Incentro, and email. After they did not respond, a reminder was sent in order to remind them of the research. If necessary, the experts within Incentro were contacted multiple times, since there were short communication lines.

Clients were chosen based on their involvement with the search implementation of Incentro's services into their software, this ranges from product owner to SEO specialist. They were reached out by email through contacts within Incentro. The total list of selected clients consisted out of 20 people, from 7 different companies. After they did not respond on the first mail,

---

[1] https://www.google.nl/intl/en/forms/about/

[2] https://kwiksurveys.com/

another was sent, however not more than two to prevent “spam” to the client. For the flowchart of the sampling procedure, please see Figure 4.

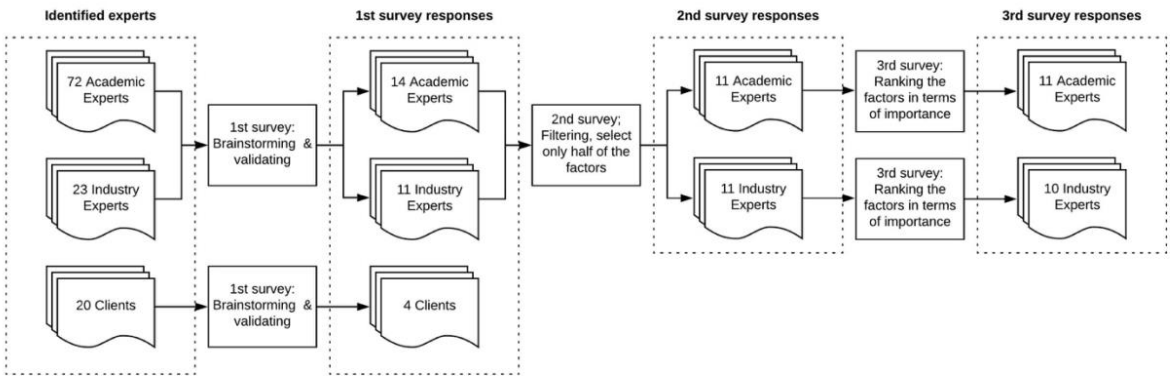


Figure 4 - Flowchart sampling procedure

First survey – Brainstorming & validating

The literature review unveiled a total of 262 factors (34 trends, 75 innovations and 153 challenges), however this was too extensive to present to the participants. Therefore, similar factors were unified. Moreover, some factors were removed on the basis of discussions with supervisors, classifying some factors as irrelevant.

The objective of the first survey is to brainstorm and validate relevant factors. Therefore the experts can come up with trends, innovations and challenges. Nevertheless, the 155 preliminary identified factors in the literature will be presented in order to stimulate discussion and possibility directly. The processes will be semi-guided by the data format and explanation, but the objective is to keep this brainstorming flexible and open for creativity. As Rowe et al. (1991) suggest that a structured or predetermined first survey will lead to sloppy execution.

The, in Google Forms made, survey for the industry- and academic-experts is built out of 3 categories: trends, innovations and challenges. In these categories the following 5 components are structured: crawling, indexing, ranking, evaluation and web search engines in a holistic view. Resulting in 5 times the question to select and add the items known to the participants per category, totaling in 15 listings of factors to select and 15 boxes to add additional missing factors. The order of factors in the presentation will be done in a randomized order to reduce bias. Additionally, 6 questions capturing the demographics of the participant. For the survey template please see Appendix 1.

The survey for the clients is made in an easier and compacter fashion, since they favor features and business value over trends, innovations and challenges. Therefore, the found factors have been connected to 27 features/statements to be selected if desired. Additionally, demographics were captured as well. For the client survey template, please see Appendix 2.

Second survey – Validation of lists

The second survey is based on the first one. The data of the first survey is removed of duplicates and unified terminology. Furthermore, several factors are removed, based on the level of agreement on each factor. The threshold for this is filter is determined to be 25%, filtering out less relevant factors.

Followed by sending the complete new merged list to all experts to await selection of the identified trends, innovations and challenges. The format is similar to the first survey; however the objective of the survey differs. Of each category (trends, innovations and challenges) and component (crawling, indexing, ranking, evaluation, web search engines) half of the factors can be selected, e.g. if there are 6 crawling trends then the participants may select a maximum of 3 factors, please see Appendix 3 for the template of the survey. Participants are also able to select no options, in order to avoid noise (forced choices). Uneven amounts of factors are rounded upwards. In the case the levels of agreement are a tie within a panel, there will be looked at the other panel to favor the one with the highest average level of agreement. On the rare occasion this also ends up in a tie, there will be an evaluation to keep or remove all factors in question for the next round based on convenience for the participant. Based on this, a new selection can be made based on the most chosen factors.

Third survey – Ranking of the list

If the results of the second survey vary in the selected factors across the panels, the third survey will be customized for each panel, resulting in two slightly different surveys. As such, the experts will be divided into two panels: academic- & industry-experts. The list of validated factors from survey 2 will be send out to every panel in order to rank the selected factors. The template differs greatly, since Google Forms did not offer the opportunity to rank items in an aesthetic appealing manner, therefore the services of Kwiksurveys have been chosen for the continuation of the research. This survey presents all the validated factor from survey 2 per category and component, similar to survey 1 & 2, in a randomized manner to be ranked against each other. Resulting in a ranked list based on importance of the factors within each category and component. For the templates of the third surveys, please see Appendix 4 & 5.

The returned lists are then analyzed, assessing the consensus by using Kendall's W, used to statistically measure the concordance within each group. The measure is between 0 and 1, with the former indicating there is no concordance, while the latter is absolute concordance. In addition. the statistical significance of each outcome was calculated.

Lastly, in each round there was an opportunity to give feedback or discussion points. The feedback between the rounds can widen knowledge and stimulate new ideas and be highly motivating (Pill, 1971) and educational (Stokes, 1997) for the participants.

4 Results

This chapter is devoted to listing and describing the results of the survey, by analyzing the results of every survey round in a brief manner. Followed by the overall findings of the accumulative sum of the survey rounds and the conclusions that be drawn of them. Section 4.1 describes the results of rounds with some isolated analysis, while Section 4.2 takes all three survey rounds into account to draw conclusions on.

4.1 Analysis of the survey rounds

All survey rounds have been analyzed and the results are described below in an atomistic manner, not considering the previous rounds of the survey.

4.1.1 Analysis of the 1st survey results

The first round consisted out of 2 panels, 14 academic- and 11 industry-experts (all from Incentro) totaling 25 participants. The third panel, consisting out of 4 clients of web search engines solutions, proved to be of inadequate number of participants to continue the panel. The reason for this low number of participants is the lack of gained benefits for the client in a direct sense combined with the low quantity list of selected experts. Efforts were made to counter this; however, it was not possible to get the threshold of 10 participants within the specified time period. Results of the client panel (survey results: Appendix 6) can be viewed as indicative, however should not be taken as a definitive measure.

The other two panels, consisting out of 25 participants, were confronted with 155 preliminary identified factors found in the literature. The level of agreement per factor was calculated by the following formula:

$$Level\ of\ agreement = 100 * \frac{Number\ of\ times\ a\ factor\ was\ selected}{Number\ of\ participants}$$

With this level of agreement, the factors could be compared and trimmed based on the threshold level of 25%. Resulting in removing 83 factors due to the level of agreement threshold. Additionally, participants contributed 15 more factors. Resulting in a total list of 87 relevant factors to be evaluated in the next round.

There seems to be a significant difference in the selection behavior of academics and industry experts, with academics being more likely to find factors relevant (37% vs 17%), see Table 8. When looking at the generated graphs in Appendix 7 and Table 8, one could address most of the difference to the innovations and challenges, since the engagement with these two categories of industry experts were especially lower. One explanation for the low engagement of industry experts in innovation could be because all innovations were found in academic literature and are in an early stage of development, while industry has it eyes on commercialized solutions available on a larger scale. Industry seems to lean more towards solutions launched by trusted firms of size.

Furthermore, since all industry experts were coming from Incentro and are implementing search solutions on a relatively small scale, they could address technical and ethical challenges in a more nuanced manner. Another possible explanation for the lower engagement of industry experts in challenges is the fact of the diverse backgrounds and professions of participants, which varies from business developer and board member to sales manager and software developer, as such business-related functions will feel less connected to technical challenges and vice versa. Lastly, the industry looks at challenges with a commercial view, as every solution to a challenge should be commercially rewarded, however, not all challenges can be directly linked to a beneficial economical outcome. Innovations should add business value to the product, if the innovation fails to accomplish this, the business has little interest pursuing the challenge, which could explain the lower levels of engagement.

Table 8 - Level of engagement survey round 1

	Level of engagement		
Category	Academics	Industry	**Multiplier**
Trends	53,4 %	38,2 %	1,40
Innovations	32,4 %	11,8 %	2,75
Challenges	33,2 %	15,0 %	2,21

4.1.2 Analysis of the 2nd survey results

For the second round of the surveys the number of participants decreased from 25 to 22, containing an equal amount of 11 participants in each group (industry- & academic-experts). The newly acquired list of 87 factors as result from the first survey was presented to the participants in order to select half of the factors. The objective of this survey round was to end up with only half of the factors (45) in order to rank them in the next round.

The panel results of the two panels had the majority of the factors congruent, although some varied, resulting in a factor list per panel, differentiating the groups for the consecutive rounds, see Appendix 8. Consequently, for the following rounds the different panels will receive different surveys, since one ranked list as result can not reflect the perspectives of multiple groups.

Since the category of indexing trends only contained two factors, "using knowledge graphs" has been selected as most important factor and will not continue to the consecutive rounds.

4.1.3 Analysis of the 3rd survey results

The third survey consisted of 45 factors, obtained from the results of the second survey. The results consisted out of 21 observations, 11 academic participants and 10 industry participants. The 45 factors were presented to the participants divided by the categories in order to rank them in terms of importance. The objective was to end up with 15 ranked lists per panel (trends, innovations and challenges per component).

The consensus within panels are assessed with Kendall's W, a statistical method for measuring consensus in groups. The formula for calculating Kendall's W is:

$$W = \frac{12S}{m^2(n^3 - n)}$$

With S as the sum of squared deviations, m as the number of judges and n as the total number of objects being ranked. W is the level of concordance, used to assess agreement between different raters, and ranges from 0 to 1. Zero indicating there is no agreement at all between the raters, while 1 is perfect agreement.

Industry has a higher level of agreement than academics as can be seen in Table 9. When breaking down the levels of agreement, one can observe the level of concordance of innovations is significantly higher within industry. Possibly explained by the tools they use within the company and channels they consume information through.

Table 9 - Kendall's W

Factor	W Industry	W Academics
Total overall	0,413***	0,321***
Trends	0,324***	0,339***
• Crawling	0,270*	0.008
• Ranking	0,182	0,178*
• Evaluation	0,210	0,074
• WSE	0,070	0,157
Innovations	0,331***	0,138
Challenges	0,393***	0,310***
• Crawling	0,170	0,157
• Indexing	0,040	0,074
• Ranking	0,576***	0,127
• Evaluation	0,000	0,207
• WSE	0,286**	0,077

$* = p < 0.1$, $** = p < 0.05$, $* * * = p < 0.01$

The ranking results of the 3rd survey are illustrated in dot charts with mean and error-bar plots in order to view the factor’s rank, mean and its standard deviation. The error bar allows to display the consensus within the panel, a smaller error-bar indicates a higher level of concordance. These plots show both the academic’s ranking as the one of the industry. Three plots have been generated: trends, innovations and challenges.

Trends

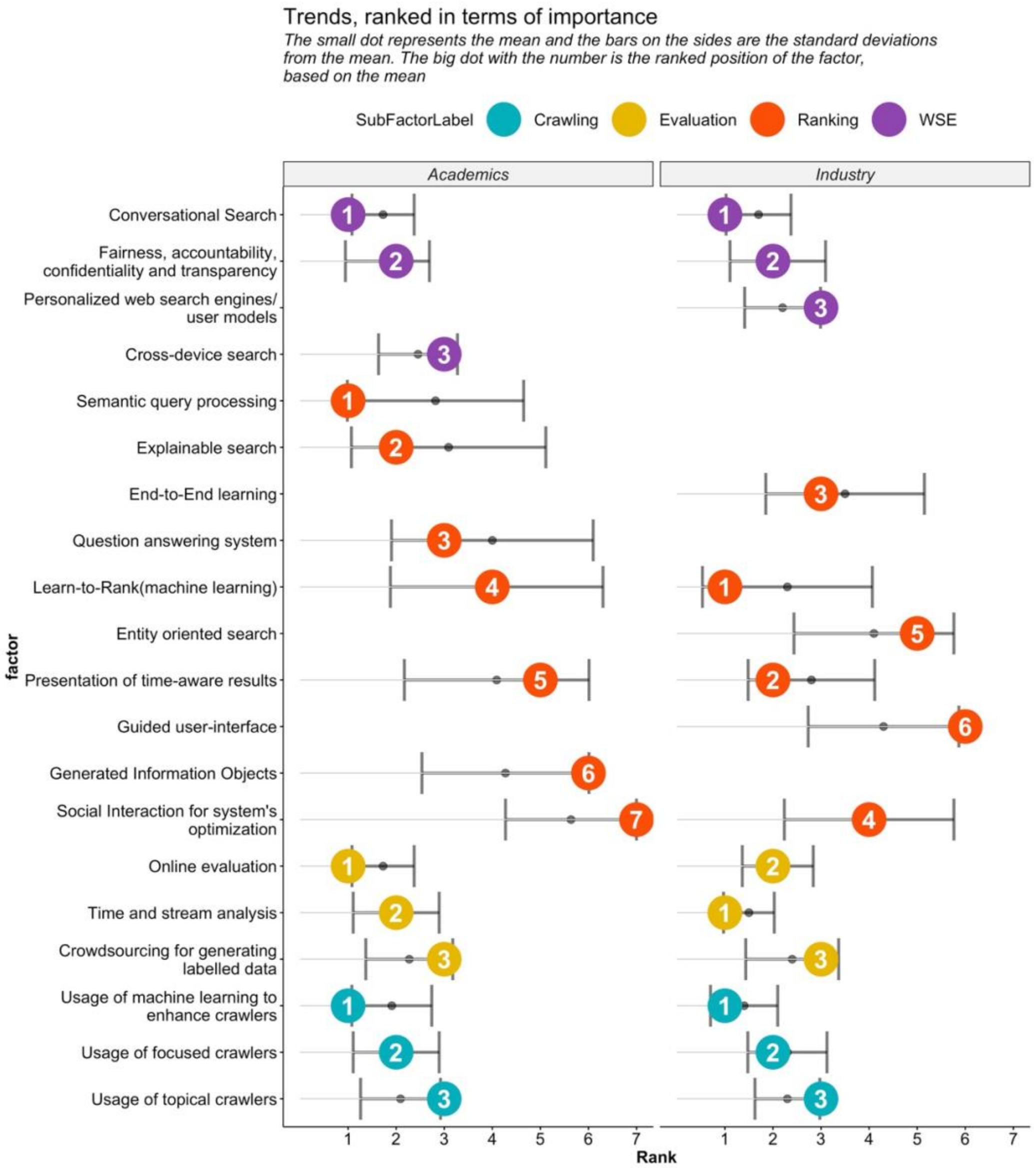


Figure 5 - Trends, ranked in terms of importance

As can be seen in figure 4, the trend factors are displayed with their corresponding rank. One can observe that crawling is quite similar, however, the others (indexing, ranking, evaluation and web search engines) vary across the panels. Some even have different factors in their ranked list. Both academics as industry experts seem to strongly disagree on the factors within ranking, as semantic query processing is not even listed within the industry list and learn-to-rank, the most important trend within ranking according to industry experts, is placed 4th by academics.

Conclusively, industry is laying the emphasis more on the importance of machine learning. Another interesting observation one could make is the lack of entity-oriented search within the ranked list of academics, which can be seen as controversially, since many of the found papers in the literature are published about entity-oriented search.

Innovations

The selected factors from the second survey vary since only 4 of the 7 factors are present in both panels. The other three do not occur in the other panel. Both sides do strongly agree natural language processing is the most important innovations currently, especially the industry advocates for this statement with a lower standard deviation see Figure 5. The support for NLP for query understanding is the highest of all factors. However, the other factors do not seem to correspond across the panels and show little consensus.

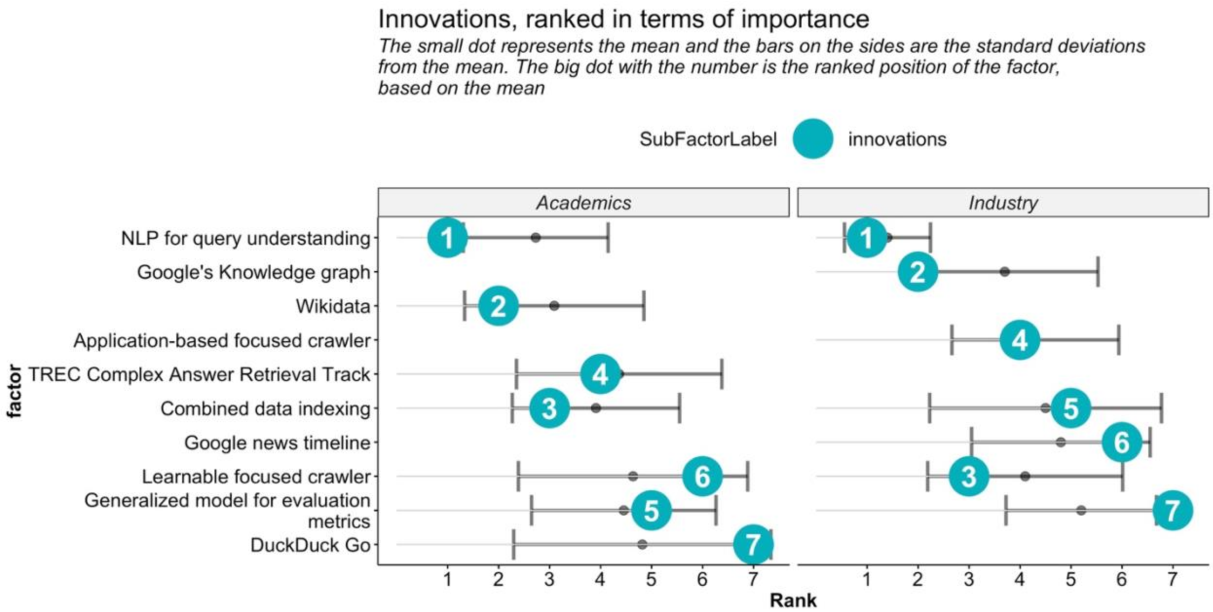


Figure 6 - Innovations, ranked in terms of importance

Challenges

Half of the challenges in crawling seem to follow the same pattern, left aside the factors which are not corresponding between the panels, however academics favor “crawling closed platforms” more than the industry does, see Figure 6. The panels strongly disagree on the challenges in ranking, there only seems to be an agreement on the importance of “semantic query understanding”. For the challenges in web search engines, academics put an emphasize on more ethical challenges as fairness, transparency and privacy, security and trust in personalized search, while the industry ranks factors related to technical challenges higher.

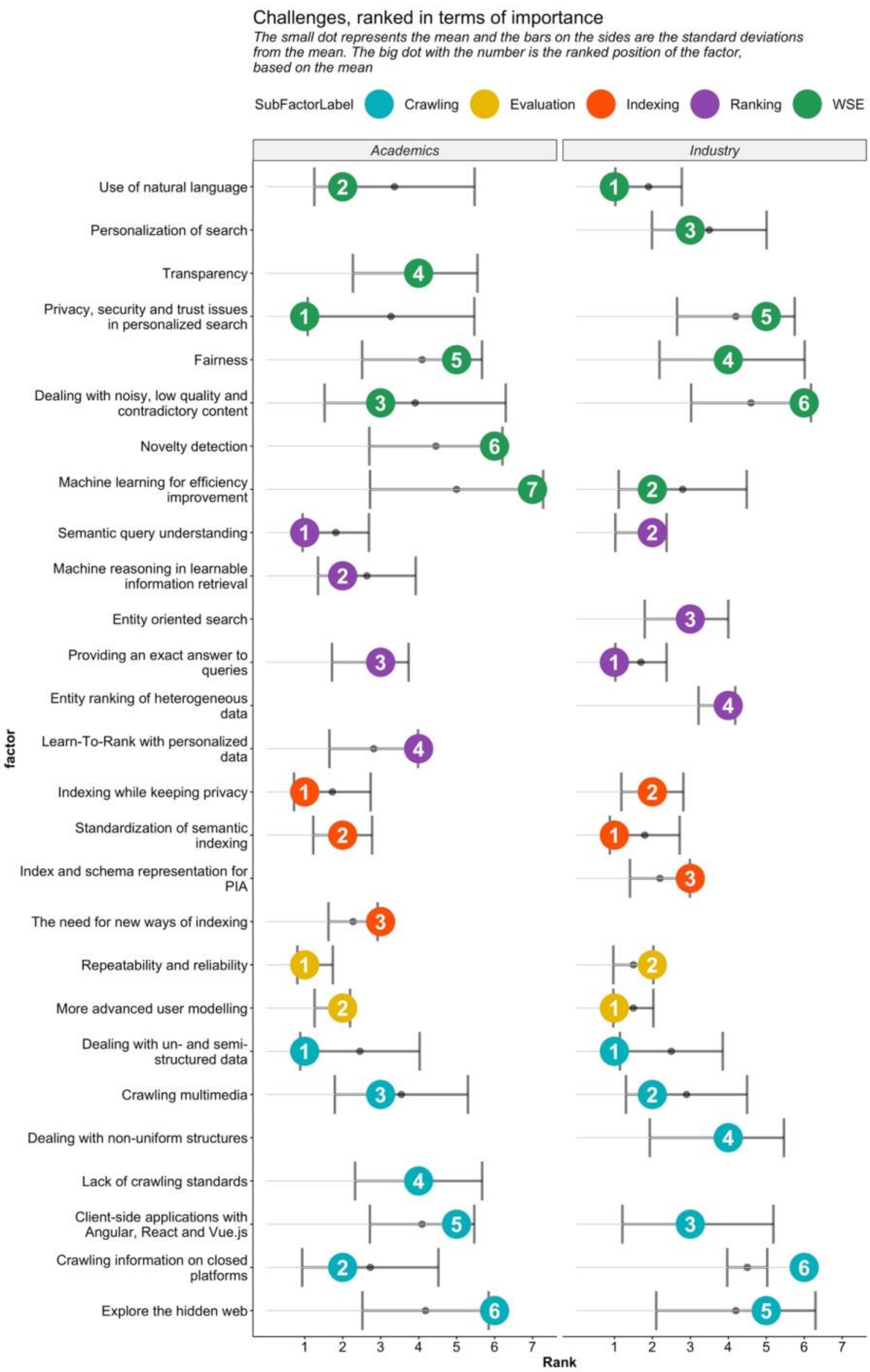


Figure 7 - Challenges, ranked in terms of importance

4.2 Overall results

Next to the analysis of the three survey rounds, another was made taking all three surveys together into account, making it possible to analyze the progress of the factors along the surveys. The results of the three survey rounds have been normalized in order to make them comparable with each other. The results of the third survey have also been reversed (1 – normalized value), to make them compatible with the other two surveys in terms of readability. Using these normalized values, plots were generated to create a view of the progress of the popularity of factors. These are divided between trends (crawling, indexing, ranking, evaluation and web search engines), innovations and challenges (crawling, indexing, ranking, evaluation and web search engines). The generated plots can be found in Appendix 9.

Personalized web search engines/user models was a factor recognized by more than 90% of the industry participants in the first and second survey, however only had a normalized mean of 0,4 as a result in the third survey. More than 85% (highest support in the first survey) of the academic participants chose personalized web search engines and application-based focused crawler in the first survey, while it did not even make it till the third survey. Therefore the conclusion can be made, many participants recognize the factor, while the importance of it is not that significant. Moreover, academics had high levels of agreement on the challenge of transparency (86%) and the trend of using crowdsourcing for generating labelled data (64%) in the first survey, while their level of importance lacked in the third survey (0.52 and 0.36 respectively). Making these four the biggest fluctuations between the first and third survey in terms of normalized means.

The factors with the highest normalized values in the third survey by industry are: NLP for query understanding (innovation), use of natural language (challenge), and usage of machine learning to enhance crawlers (trend). While the highest factors by academics are: NLP for query understanding (innovation), semantic query processing (trend), and repeatability and reliability (challenge).

5 Discussion & conclusions

No studies could be found describing the different viewing points of various panels on web search engines. Therefore this study has been conducted and can be a catalyst to further investigate the relationship between the panels, perhaps by adding other panels as end users and clients of web search solutions. All resulting in a better understanding of the current landscape and to stimulate cooperation and collaboration between the different stakeholders to increase the value search engines deliver to billions of users.

In the present dissertation, we pursued a broad introductory study covering different perspectives on the trends, innovations and challenges of web search engines using the Delphi methodology. By constructing three expert panels (academics, industry and clients) and conducting three survey rounds, the perspectives on the importance of the different aspects of web search engines were distilled. Using the PRISM method, the literature review of 15 articles identified 155 preliminary trends, innovations and challenges. These were presented to the participants in the consecutive survey rounds, each survey with its objective. The first to validate and add factors, the second to select only half of the factors, and finally the third to rank the chosen factors. Resulting in a final list of 45 factors ranked by terms of importance per category, facetted by panel.

From the results it can be concluded academics and industry do share some perspectives of web search engine component's trends, innovations and challenges. However, there was also some disagreement on certain perspectives. As such industry has its view more towards fully commercialized solutions, while academics also recognizes the innovations in an early stage of development. The trends of machine learning are ranked high on importance by the industry, while academics seem to favor a wider spectrum of trends, although academics also highly support the use of machine learning to enhance crawlers. Moreover, academics agreed on the importance of semantic query processing, even though they also view it as one of the most important challenges. When comparing all three panels, they support the trend of conversational search and the trend of fairness, accountability, confidentiality and transparency, however industry in a more nuanced manner.

Conversational search has increased in popularity due to advancements in NLP recognized by both panels, however, simultaneously, use of natural language is still seen as one of the most important challenges. Both panels agree on the challenge of crawling un- and semi-structured data. Industry also sees challenges related to entity-oriented search as the most important: semantic query understanding, entity-oriented search and providing an exact answer to queries. It seems odd to see the industry supporting entity-oriented search related issues more than their academic counterpart, as entity-oriented search is still in its infancy and not often implemented in commercialized solutions.

Overall the industry does not identify (above the threshold level of agreement of 25%) as many challenges as academics do, which could be due to the fact industry consists of people with different functions all concerned with challenges in their own working environment, e.g. business developer, software developer, marketing & sales manager. One class, which industry identifies itself less with are the more ethical related challenges, which could be the result of operation on a relatively small and narrow scale.

The panel of clients seems to advocate for incremental innovations, however conversational search is a desired feature (75%). While the client's results can only be taken as an indicative value due to size of sample size, one can observe the three panels do align on the importance of conversational search. Indicating an interesting direction for the future of web search engines. However, in order to gain a deeper understanding of the IR landscape from a holistic viewpoint, additional research would be needed, involving more clients, a more diverse group of industry experts and possibly end-users in the research.

Since this research focused on identifying the importance of factors within their component, a more high-level conclusion is missing. As such, future research could continue by evaluating the importance of all factors compared with each other in a category (trends and challenges).

References

1. Abiteboul, S. (1997). Querying semi-structured data. In: International Conference on Database Theory. Berlin and Heidelberg: Springer; 1997, 1–18

2. Adler, M. & Ziglio. E. (1996). Gazing into the oracle: The Delphi Method and its application to social policy and public health. London: Jessica Kingsley Publishers.

3. Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), 1129-1153. doi:10.1016/j.ipm.2018.08.001

4. Asikri, M. E., Krit, S., Chaib, H., Kabrane, M., Ouadani, H., Karimi, K., Elbousty, H. (2017). Mining the Web for learning ontologies: State of art and critical review. 2017 International Conference on Engineering & MIS (ICEMIS). doi:10.1109/icemis.2017.8273103

5. Badawi, M., Mohamed, A., Hussein, A. and Gheith, M. (2013). Maintaining the search engine freshness using mobile agent. Egyptian Informatics Journal, 14(1), pp.27-36.

6. Balog, K. (2019). Etity-Oriented Search. S.l.: SPRINGER.

7. Bhoir, V., Potey, M., 2014. Question answering system: a heuristic approach. 2014 Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT). IEEE, pp. 165–170. https://doi.org/10.1109/icadiwt.2014.6814704.

8. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks And ISDN Systems, 30(1-7), 107-117. doi: 10.1016/s0169-7552(98)00110-x

9. Calumby, R., Gonçalves, M., & Torres, R. (2016). On interactive learning-to-rank for IR: Overview, recent advances, challenges, and directions. Neurocomputing, 208, 3-24. doi: 10.1016/j.neucom.2016.03.084

10. Cambazoglu, B. and Baeza-Yates, R. (2015). Scalability Challenges in Web Search Engines. Synthesis Lectures on Information Concepts, Retrieval, and Services, 7(6), pp.1-138.

11. Catena, M., Tonellotto, N. (2017). Energy-Efficient Query Processing in Web Search Engines. In EEE Transactions on Knowledge and Data Engineering PP(99):1-1

12. Chen, Y., Zhou, K., Liu, Y., Zhang, M., Ma, S. (2017) Meta-evaluation of online and offline web search evaluation metrics. In: SIGIR '17: 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 7-11 August 2017, Shinjuku, Tokyo, Japan

13. Chuklin, A., de Rijke, M. and Sedyukov, P. (2013). Click model-based information retrieval metrics. SIGIR '13 Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.493-502.

14. Culpepper, J. S., Diaz, F., & Smucker, M. D. (2018). Research Frontiers in Information Retrieval. ACM SIGIR Forum, 52(1), 34-90. doi:10.1145/3274784.3274788

15. Dawson S. & Barker J. (1995) Hospice and palliative care: a Delphi survey of occupational therapists roles and training needs. Australian Occupational Therapy Journal of 42, 119–127.

16. Delbecq, A.L., Van de Ven, A.H., Gustafson, D.H. (1975). Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes, Scott, Foresman and Company, Glenview, Illinois, 1975.

17. Deng, A., and Shi, X. (2016). Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. In SIKDD'16. ACM.

18. Dixit, A., Sharma, A.K. (2011). Security system for migrating crawlers. In: 2011 International Conference of the Comput. Intell. Commun. Networks, IEEE, 2011, 667–671. https://doi.org/10.1109/CICN.2011.145.

19. Elbedweihy, K., Wrigley, S., Clough, P., & Ciravegna, F. (2015). An Overview of Semantic Search Evaluation Initiatives. SSRN Electronic Journal. doi: 10.2139/ssrn.3199177

20. Everett, A. (1993) Piercing the veil of the future: a review of the Delphi method of research. Professional Nurse 9, 181–185.

21. Fink , A., Kosecoff, J., Chassin, M. & Brook, R. (1991) Consensus Methods: Characteristics and Guidelines for Use RAND, Santa Monica, California.

22. Hammer, J. and Fieldler, J. (2000). Using mobile crawlers to search the Web efficiently. International Journal of Computer and Information Science, 1, pp.36-58.

23. Hofmann, K., Li, L., & Radlinski, F. (2016). Online Evaluation for Information Retrieval. Foundations And Trends In Information Retrieval, 10(1), 1-117. doi: 10.1561/1500000051

24. Huurdeman, H. C., & Kamps, J. (2014). From Multistage Information-Seeking Models to Multistage Search Systems. IIiX '14 Proceedings of the 5th Information Interaction in Context Symposium, 145-154. doi:10.1145/2637002.2637020

25. Jayanthi, J., & RATHI, S., DR. (2014). Personalized Web Search Methods – A Complete Review. Journal of Theoretical and Applied Information Technology, 62(3).

26. Jiang, J., Hassan Awadallah, A., Shi, A. and W White, R. (2015). Understanding and predicting graded search satisfaction. In WSDM'15. ACM, 57–66.

27. Jones, J.M.G., Sanderson, C.F.B. & Black, N.A. (1992) What will happen to the quality of care with fewer junior doctors? A Delphi study of consultant physicians' views. Journal of the Royal College of Physicians London 26, 36–40.

28. Kathuria, M., Nagpal, C., & Duhan, N. (2016). Journey of Web Search Engines: Milestones, Challenges & Innovations. International Journal Of Information Technology And Computer Science, 8(12), 47-58. doi: 10.5815/ijitcs.2016.12.06

29. Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of Web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery, 7(6), e1218. doi: 10.1002/widm.1218

30. Lester, N., Zobel, J. and Williams, H. (2006). Efficient online index maintenance for contiguous inverted lists. Information Processing & Management, 42(4), pp.916-933.

31. Liberati, A. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. Annals of Internal Medicine, 151(4), p.W.

32. Lindeman, C. (1975) Delphi survey of priorities in clinical nursing research. Nursing Research 24, 434–441.

33. Linstone, H. & Turloff, M. (1975). The Delphi method: Techniques and applications. London, UK: Addi- son-Wesley.

34. Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., and Zhu, X. (2015). Different users, different opinions: Predicting search satisfaction with mouse movement information. In SIGIR'15. ACM, 493–502.

35. Mahapatra, A. and Biswas, S. (2011). Inverted indexes: Types and techniques. IJCSI International Journal of Computer Science Issues, 8(4).

36. Malik, N., Sharan, A., Biswas, P., (2013). Domain knowledge enriched framework for restricted domain question answering system. 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, pp. 1–7. https://doi.org/10.1109/iccic.2013.6724163.

37. Morales, J., & Melgar, A. (2017). Research on Proposals and Trends in the Architectures of Semantic Search Engines: A Systematic Literature Review. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. doi:10.15439/2017f88

38. Moulahi, B., Tamine, L. and Yahia, S. (2016). When time meets information retrieval: Past proposals, current plans and future trends. Journal of Information Science, 42(6), pp.725-747.

39. Murphy, M.K., Black ,N., Lamping, D.L., McKee, C.M., Sanderson, C.F.B., Askham, J. (1998) Consensus development methods and their use in clinical guideline development. Health Technology Assessment 2(3).

40. Neves, M., Leser, U., 2015. Question answering for biology. Methods 74, 36–46. https://doi.org/10.1016/j.ymeth.2014.10.023.

41. Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web.

42. Pahal, N. (2010). Security on mobile agent based crawler (SMABC). Int J Comput Appl 2010, 1:5–11.

43. Pewinternet. (2019). Search and email still top the list of most popular online activities. [online] Available at: http://www.pewinternet.org/2011/ 08/09/search- and- email- still- top- the- list- of- most- popular- online- activities/ [Accessed 16 Jan. 2019].

44. Pill, J. (1971) The Delphi method: substance, context, a critique and an annotated bibliography. Socio-Economic Planning and Science 5 , 57–71.

45. Powell, C. (2003). The Delphi technique: Myths and realities. Journal of Advanced Nursing, 41(4), 376-382. doi:10.1046/j.1365-2648.2003.02537.x

46. Robertson, S. (2008). On the history of evaluation in IR. Journal of Information Science, 34(4), pp.439-456.

47. Rowe, G., Wright, G. & Bolger, F. (1991) Delphi: a re-evaluation of research and theory. Technical Forecasting Social Change 39, 235– 251.

48. Ruthven, I. and Kelly, I. (2011). Interactive Information Seeking, Behaviour and Retrieval. Facet, 2011.

49. Saracevic, T. (1995), Evaluation of evaluation in information retrieval. In Proceedings of SIGIR, pp.138-146

50. Seo, C., Lee, S. and Kim, H. (2003). An efficient inverted index technique for XML documents using RDBMS. Information and Software Technology, 45(1), pp.11-22.

51. Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., A Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. (2016). BMJ, p.i4086.

52. Shen, W., Wang, J., & Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2), 443-460. doi:10.1109/tkde.2014.2327028

53. Singh, L. and Sharma, D. (2013). An approach for accessing data from hidden web using intelligent agent technology. 2013 3rd IEEE International Advance Computing Conference (IACC).

54. Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi Method for Graduate Research. Journal of Information Technology Education: Research, 6, 001-021. doi:10.28945/199

55. Soares, M. A., & Parreiras, F. S. (2018). A literature review on question answering techniques, paradigms and systems. Journal of King Saud University - Computer and Information Sciences. doi:10.1016/j.jksuci.2018.08.005

56. Stokes, F. (1997) Using the Delphi technique in planning of a research project on the occupational therapists' role in enabling people to make vocational choices following illness/injury. British Journal of Occupational Therapy 60, 263–267.

57. Sumit, B., Muktawat, H.S., (2010). Deep web.

58. Tamine, L., & Daoud, M. (2018). Evaluation in Contextual Information Retrieval. ACM Computing Surveys, 51(4), 1-36. doi:10.1145/3204940

59. Tekli, J., Chbeir, R., Traina, A. and Traina, C. (2019). SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data. Knowledge-Based Systems, 164, pp.378-403.

60. Tort, A., Olivé, A. (2015). An approach to website schema.org design. In Data & Knowledge Engineering. Volume 99, September 2015, Pages 3-16

61. Upadhyay, V., Balwan, J., Shankar, G., Amritpal, A. (2012). Security approach for mobile agent based crawler. In: Advances in Computer Science, Engineering & Applications. Wyld DC, Zizka J, Nagamalai D, eds. Berlin and Heidelberg: Springer; 2012, 119-123. https://doi. org/10.1007/978-3-642-30111-7_12

Appendices

Appendix 1: Template survey 1 academics & industry

Identifying trends, innovations and challenges

First of all, thank you for participating in this research. Your opinion is highly valued.

The main objective of this survey is to brainstorm and complement the identified relevant factors of web search engines. I identified various factors in the literature and interviews, however I need your expertise to validate these findings and especially add additional ones, which I did not think of. Your contribution will have an impact on the research and your input is important.

You could look at this survey as a brainstorm session in which experts from various domains try to find consensus on which trends, innovations and challenges are relevant to web search engines.

The preliminary found factors were divided between trends, innovations and challenges. Factors can be identified on component level (Crawling, indexing, Ranking, Evaluation) or web search engines as a whole.

I would appreciate if you could SELECT and ADD all factors relevant to web search engines in your opinion. If the factor is not known to you or the definition of it, then don't select it.

If you think a factor(trend, innovation or challenge) is missing, please complement the list in the section below the checkboxes.

Feedback and discussion is very welcome as well.

*Vereist

I would appreciate if you could SELECT and ADD all the relevant TRENDS known to you

1. **Trends in Crawling**

   Vink alle toepasselijke opties aan.

   - [ ] Usage of forum crawlers
   - [ ] Usage of mobile crawlers - selection and filtration of webpages are performed on server side rather than on search engine side
   - [ ] Usage of "hidden web" crawlers - e.g. crawling the web behind interfaces, forms and logins
   - [ ] Usage of topical crawlers - crawling based on topical relevance
   - [ ] Usage of focused crawlers - focused crawler technique gives priority to those URLs in the process of crawling, in which probability of finding information of user's interest is high (e.g. based on link, text and URL)
   - [ ] Usage of deep web crawlers(Hidden Web with bad intent)

2. **Additional trends**

3. **Trends in Indexing**
Vink alle toepasselijke opties aan.

- [ ] Usage of hybrid directories - using multiple data sources and not only the crawled data
- [ ] Using knowledge graphs - A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.

4. **Additional trends**

____
____
____
____
____

5. **Trends in Ranking**
Vink alle toepasselijke opties aan.

- [ ] Social Interaction for system's optimization - exploring social web for making new connections between images and tags
- [ ] Semantic query processing - understanding the user's query based on a knowledge base rather than matching documents directly on search terms
- [ ] Question answering system
- [ ] Entity oriented search - finding "things" instead of "strings"
- [ ] Presentation of time-aware results (e.g. football games that are playing now instead of the results of the previous years)
- [ ] End-to-End learning - running a machine learning model on a task in a end-to-end fashion e.g. input a query output documents
- [ ] Generated Information Objects - summarised content generated by a machine using various sources
- [ ] Learn-to-Rank(machine learning) - automatically construct a ranking model using training data applicable to new documents
- [ ] Guided user-interface
- [ ] Multi-modality - input query or presenting results from various data formats e.g. text, audio, video

6. **Additional trends**

____
____
____
____
____

7. **Trends in Evaluation**
Vink alle toepasselijke opties aan.

- [ ] Online evaluation - using online interaction loggings as way of evaluating relevance of the results
- [ ] Evaluation of Generated Information Objects - summarised content generated by a machine using various sources
- [ ] Crowdsourcing for generating labelled data
- [ ] Time and stream analysis - evaluation of real-time web search engines e.g. news
- [ ] Using eye-tracking for evaluation

8. **Additional trends**

____
____
____
____
____

9. **Trends in Web Search Engines**
Vink alle toepasselijke opties aan.

- [ ] Personalized web search engines/user models - e.g. based on prior queries, knowledge, information needs
- [ ] Smaller, Faster and Better web search engines(incremental innovations)
- [ ] Cross-device search - continuation of a search session on another device
- [ ] Web Search Engines for supporting knowledge goals and decision making
- [ ] Fairness, accountability, confidentiality and transparency
- [ ] Conversational Search - increased interaction by using natural language in a conversational manner

10. **Additional trends**

____
____
____
____
____

I would appreciate if you could SELECT and ADD all the relevant INNOVATIONS known to you

11. **Innovations in Crawling**
Vink alle toepasselijke opties aan.

- [ ] Application-based focused crawler - Crawlers that target a particular group e.g. medical, educational
- [ ] Onion Search (DARPA)
- [ ] Topic-specific crawler - topic-specific crawling is a method that crawls webpages according to the user interest
- [ ] Deep web search engine
- [ ] Learnable focused crawler - a trained crawler deciding the relevance of each page in order to assess which ones to crawl
- [ ] Filtered (on server side) crawling

12. **Additional innovations**

13. **Innovations in Indexing**
Vink alle toepasselijke opties aan.

- [ ] Combined data indexing - new representation models for text, entities and their relations
- [ ] Schema.org for semantic indexing standardization
- [ ] Google Wikilinks corpus, a large scale entity linking data set with 40 million disambiguated mentions within over 10 million web pages

14. **Additional trends**

15. **Innovations in Ranking**
Vink alle toepasselijke opties aan.

- [ ] NLP for query understanding
- [ ] Microsoft's Satori for entity oriented search
- [ ] Continuous search - searching over a period
- [ ] Google's Knowledge graph
- [ ] Extended learnability - change in performance over time after initial training
- [ ] Reasoning components - in order to identify relevant documents and suggest related terms to improve future user queries
- [ ] SCINET search engine for personalized search
- [ ] Domain ontologies - using specific ontologies for better query understanding and more relevant results

16. **Additional trends**

17. **Innovations in Evaluation**
Vink alle toepasselijke opties aan.

- [ ] TREC Complex Answer Retrieval Track(for Generated Information Objects)
- [ ] NTCIR One-Click(for Generated Information Objects)
- [ ] Terrier evaluation framework

18. **Additional trends**

19. **Innovations in Web Search Engines**
Vink alle toepasselijke opties aan.

- [ ] Swoogle - semantic web search engine
- [ ] K-Search - natural language view-based interfaces allowing users to explore the search space whilst formulating their queries
- [ ] DBpedia - consists of around 1.8 billion RDF triples in multiple languages
- [ ] Sig.ma - integrating data from different sources to provide rich descriptions about Semantic Web objects
- [ ] Freya - natural language interface in a multiple and heterogeneous domains
- [ ] Sindice - semantic web search engine
- [ ] Querix - natural language interface in a single domain
- [ ] Smeagol - natural language view-based interfaces allowing users to explore the search space whilst formulating their queries
- [ ] Wikidata - a free and open knowledge base that can be read and edited by both humans and machines
- [ ] Newsmap that visually reflects the constantly changing landscape of the Google News aggregator
- [ ] Visinav - integrating data from different sources to provide rich descriptions about Semantic Web objects
- [ ] Google news timeline
- [ ] NLP-reduce - natural language interface in a single domain
- [ ] PowerAqua - natural language interface in a multiple and heterogeneous domains

20. **Additional trends**

I would appreciate if you could SELECT and ADD all the relevant CHALLENGES known to you

21. **Challenges in Crawling**
Vink alle toepasselijke opties aan.

- [ ] Explore the "hidden web"
- [ ] Opacity between information seeker and provider (onion search, the dark web)
- [ ] Effectively deal with web spam
- [ ] Dealing with non-uniform structures(e.g. dealing with client side scripts, dealing with non HTML search interfaces)
- [ ] Scale and revisit
- [ ] Crawling multimedia
- [ ] Open source web crawlers
- [ ] Lack of crawling standards
- [ ] Mobile crawlers(security issues, server side push methods)
- [ ] Dealing with unstructured and structured data (e.g. should collected data fit a schema or be stored in a schema)

22. **Additional challenges**

23. **Challenges in Indexing**
Vink alle toepasselijke opties aan.

- [ ] Standardization of semantic indexing
- [ ] The need for new ways of indexing
- [ ] Challenging the high benchmarks of the current setup in terms of effectiveness and efficiency
- [ ] Configurations of hardware setup
- [ ] Index and schema representation for personal information access
- [ ] Indexing technique for hidden data

24. **Additional challenges**

25. **Challenges in Ranking**
Vink alle toepasselijke opties aan.

- [ ] Semantic query understanding
- [ ] Entity oriented search - finding "things" instead of "strings"
- [ ] Avoiding memory recall - smart cache management
- [ ] Simultaneously supporting a generic overview as enabling specialists' groups to drill down to exclusively relevant items
- [ ] Providing an exact answer to queries
- [ ] Computing over aggregated personal data
- [ ] Learn-To-Rank with personalized data
- [ ] Coping with limited data in Learn-To-Rank
- [ ] Data representation in Learn-To-Rank
- [ ] Machine reasoning in learnable information retrieval
- [ ] Optimal interactive user interface
- [ ] Reducing labelling effort (training dataset)
- [ ] Integrating advanced procedures for handling complex queries
- [ ] Exploring learning boosting alternatives - e.g. diversity promotion for handling ambiguous, multi-intent, overview, or underspecified queries.
- [ ] Using reinforcement learning for combining multiple features modalities/learning strategies
- [ ] Analyze user behavior impact on search tasks, which will produce information for the development of better generalization models and more realistic user models
- [ ] Leverage long-term learning and collaborative retrieval for effectiveness and efficiency
- [ ] Using additional sources for semantic web
- [ ] Entity linking (e.g. name variation, entity ambiguity)
- [ ] Combining entity recognition and entity linking
- [ ] Convert misspelled queries in relevant results
- [ ] Entity ranking of hetereogeneous data (e.g. tweets, tables)

26. **Additional challenges**

27. **Challenges in Evaluation**

Vink alle toepasselijke opties aan.

- [ ] Counterfactual analysis - offline simulation of user interaction interaction with slightly changed IR system
- [ ] Define the axiometrics of online evaluation
- [ ] New online metrics from new online interactions - e.g. eye tracking
- [ ] Decomposable performance prediction of IR systems to evaluate how the components interact, and how factors external to the system also impact overall performance
- [ ] Shift to a goal-oriented way of evaluation metrics for Generated Information Objects
- [ ] Better benchmarks for evaluation
- [ ] More suitable/effective evaluation metrics
- [ ] Conduct real-user experiments
- [ ] More advanced user modelling
- [ ] Metrics better suited for Learn-To-Rank methods
- [ ] Using graded relevance to improve ground-truth quality and maximize feedback information
- [ ] Better log analysis methods
- [ ] User models considering reformulation understanding
- [ ] Standardized evaluation for semantic search engines
- [ ] Evaluating over sessions instead of queries
- [ ] Gathering relevance assessments efficiently
- [ ] Comparing system effectiveness and user utility
- [ ] The need for an evaluation initiative administered by a well- respected organization for distributing datasets, organize campaigns and report results
- [ ] Developing a hybrid approach between system and user oriented evaluation metric
- [ ] User centric evaluation metrics
- [ ] Repeatability and reliability
- [ ] Reusability of datasets for evaluation

28. **Additional challenges**

29. **Challenges in Web Search Engines**

Vink alle toepasselijke opties aan.

- [ ] Effectively dealing with "deep web"
- [ ] On demand anticipation, customization and personalization
- [ ] Dealing with noisy, low quality and contradictory content continuously uploaded to the web
- [ ] Search during a period (continuous search)
- [ ] Novel methods for more time sensitive search results
- [ ] Comparing retrieval models is difficult due to different architectures
- [ ] Presentation of results on mobile devices
- [ ] Topic detection and tracking
- [ ] Novelty detection
- [ ] Ethics elimination or reduction of bias
- [ ] Fairness
- [ ] Accountability
- [ ] Confidentiallity
- [ ] Transparency
- [ ] Understanding cognitive aspects of users relevant to their information seeking
- [ ] Ways to aid searchers in evaluating and contextualizing search results
- [ ] Guidance through learning or decision-making process of the user
- [ ] Knowledge graph representation
- [ ] Merging of heterogenous information source for a generated information object
- [ ] Deriving explanations from generated information objects
- [ ] Context and personalization in generated information objects
- [ ] Machine learning for efficiency improvement
- [ ] Understanding and anticipating needs for personalized search
- [ ] Personalization of search
- [ ] Privacy, security and trust issues in personalized search
- [ ] Architecture and applications of personalized search
- [ ] Lack of sentimental search engine - apart from the keyword match-based search results, the user can be provided with sentiment- and demographic-based results
- [ ] Existence and maintenance of ontologies
- [ ] Evolution of knowledge base as new documents appear
- [ ] Use of natural language
- [ ] Handling scale
- [ ] Multi-stage Search Systems

30. **Additional challenges**

We are almost finished if you just allowed me to collect some additional data for the purpose of demographics

The data collection will respect the legislation of data protection. No data will be shared and solely be used for follow-ups and distinction of the data. Data will not be linked.

31. **Name** *

32. **Gender** *
Markeer slechts één ovaal.
- Male
- Female

33. **Nationality** *

34. **Profession** *
Markeer slechts één ovaal.
- Professor
- PhD-student
- Business Developer
- Search consultant
- Developer
- Sales manager
- Anders:

35. **Organization/company you work for**

36. **E-mail address** *

Nomination of other experts relevant to the research

Thank you for participating in this research. If you know anyone that would be an added value to this research and willing to participate please nominate this person and I will contact this person for his/her input.

37. **Name**

38. **E-mail address**

Appendix 2: Template survey 1 clients

Brainstorm survey desired features web search engines

First of all, thank you for participating in this research. Your opinion is highly valued.

Currently I'm doing an internship at Incentro to graduate for my master in innovation & technological entrepreneurship. In this internship I am doing research into web search engines with as goal to improve the quality of Incentro's services concerning search. This can be in the form of the addition of features, implementation roadmap, new forms of interaction with the system, etc.

Therefor I need your help, please put your visionary glasses on and think about search in the coming future. What features does it have? How is the interaction with the system?

In the following sections a preliminary list of identified features is presented. These are described in a very brief manner, followed by a short explanation or some simple examples.

Please select all features relevant to web search engines in your opinion. If the factor is not known to you or the definition of it, then don't select it.

See this survey as a brainstorm session in which there are already some ideas (to validate), but your input of information is very important as well. Therefore, if you think a feature, concept or idea is missing, please complement the list in the "completing the list" section below the checkboxes.

Feedback and discussion is very welcome as well.

The survey will take up approximately 5 minutes of your time. I would like to thank you once more for participating in this survey.

*Vereist

1. **Select all desired features in web search engines**

Vink alle toepasselijke opties aan.

- [ ] Reduced network load of webcrawlers
- [ ] Results/indexing of pages behind interface - e.g. displaying results only available after login or filling in forms
- [ ] Guided user interface to drill down to the exact thing you are looking for
- [ ] Entity oriented search, finding "things" instead of "strings" - e.g. knowing the entity(person, location, etc) and its relationships to the real world
- [ ] Question answering instead of presenting relevant pages
- [ ] Input query or presenting results from various data formats e.g. not only showing text as a result, but also relevant images or videos
- [ ] Machine generated content as result - e.g. summary of a page
- [ ] Time-aware results - e.g. football match which is playing now instead of results of last year
- [ ] New online metrics from online interactions - e.g. eye tracking
- [ ] Smaller, faster and better web search engines
- [ ] Personalized web search - awareness of your level of knowledge in certain domains by which it presents you documents corresponding with your level of the subject
- [ ] Fairness, accountability, confidentiality and transparency in web search engines
- [ ] Context- and location-aware retrieval - takes the context and location of the user into account for serving which documents are more relevant
- [ ] Cross-device search - continuation of a search session on another device
- [ ] Conversational Search - increased interaction by using natural language in a conversational manner
- [ ] Continuous search - searching over a period of time
- [ ] Explanation why the list of results were retrieved
- [ ] Evaluation framework to determine the relevance of the retrieved results
- [ ] High scalability of web search engine
- [ ] Crawling and indexing multimedia
- [ ] Convert misspelled queries in relevant results
- [ ] Developing a hybrid approach between system and user oriented evaluation metric - e.g not only evaluating the search engine on system performance or on the user performance, but the combination of the two
- [ ] Repeatability and reliability of web search engine
- [ ] Guidance through learning or decision-making process of the user
- [ ] Automatically labelling multimedia with metadata to make it searchable
- [ ] Automatically add synonyms to the query in order to retrieve more diverse results
- [ ] Data insights from the user's search sessions

2. **Complementing the list/feedback**

____

____

____

____

____

We are almost finished if you just allowed me to collect some additional data for the purpose of demographics

The data collection will respect the legislation of data protection. No data will be shared and solely be used for follow-ups and distinction of the data. Data will not be linked.

3. **Name**

   ____

4. **Gender**

   Markeer slechts één ovaal.

   ◯ Male
   ◯ Female

5. **Profession**

   ____

6. **Organization/company you work for**

   ____

7. **Which search service/product Incentro provides are you using?**

   Vink alle toepasselijke opties aan.

   ☐ Segona
   ☐ MindBreeze
   ☐ Google Cloud Search
   ☐ Anders: ________________________________

8. **How satisfied are you with the service Incentro delivers?**

   Markeer slechts één ovaal.

	1	2	3	4	5	
Very dissatisfied	◯	◯	◯	◯	◯	Very satisfied

9. **How satisfied are you with the product Incentro delivers?**

   Markeer slechts één ovaal.

	1	2	3	4	5	
Very dissatisfied	◯	◯	◯	◯	◯	Very satisfied

10. **Feel free to share feedback**

    ____
    ____
    ____
    ____
    ____

11. **E-mail address** *

Nomination

Thank you for participating in this research. If you know anyone that would be an added value to this research and willing to participate please nominate this person and I will contact this person for his/her input.

12. **Name**

13. **E-mail**

Appendix 3: Template survey 2 academics & industry

Identifying trends, innovations and challenges

Thank you for participating in the second round of these survey rounds. Your opinion is valuable for the continuation of this research.

In the previous round we started with 155 factors. The contribution of the participants was the addition of 15 new factors. Moreover, the initial list of 155 factors was filtered on the level of agreement(results of the previous survey), with 25% level of agreement as a threshold level to consider the factor as relevant and keep it in the list. With this threshold 83 factors have been removed from the list. Resulting in total list of 87 remaining relevant factors.

For this survey round the objective is to filter the list even further, by selecting a maximum of half of the listed factors in each section (crawling, indexing, ranking, evaluation, Web Search Engines). For example "Trends in Crawling" has 6 listed factors, therefore one should pick a maximum of 3 factors as the most relevant in this section.

Feedback and discussion is very welcome as well. At the end of the survey there is an option to deliver feedback and/or discussion on the results of the survey.

*Vereist

Thinking about the TRENDS in the development of Web Search Engines that you know about, please select from the lists below the ones you would consider the MOST RELEVANT.

1. **Trends in Crawling (please select up to 3)**
   Vink alle toepasselijke opties aan.

   - [ ] Usage of mobile crawlers - selection and filtration of webpages are performed on server side rather than on search engine side
   - [ ] Usage of focused crawlers - focused crawler technique gives priority to those URLs in the process of crawling, in which probability of finding information of user’s interest is high (e.g. based on link, text and URL)
   - [ ] Usage of topical crawlers - crawling based on topical relevance
   - [ ] Usage of forum crawlers
   - [ ] Usage of machine learning to enhance crawlers
   - [ ] Centralized repository of crawled documents, to make the crawled documents available for everyone and reduce network load

2. **Trends in Indexing (please select 1)**
   Vink alle toepasselijke opties aan.

   - [ ] Usage of hybrid directories - using multiple data sources and not only the crawled data
   - [ ] Using knowledge graphs - a knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.

3. **Trends in Ranking (please select up to 7)**

Vink alle toepasselijke opties aan.

- [ ] Robustness of ranking models
- [ ] Semantic query processing - understanding the user's query based on a knowledge base rather than matching documents directly on search terms
- [ ] Question answering system
- [ ] Explainable search - explain the results of the search
- [ ] Social Interaction for system's optimization - exploring social web for making new connections between images and tags
- [ ] Exploiting structure - e.g. xml
- [ ] Guided user-interface
- [ ] Entity oriented search - finding "things" instead of "strings"
- [ ] End-to-End learning - running a machine learning model on a task in a end-to-end fashion e.g. input a query output documents
- [ ] Generated Information Objects - summarised content generated by a machine using various sources
- [ ] Learn-to-Rank(machine learning) - automatically construct a ranking model using training data applicable to new documents
- [ ] Multi-modality - input query or presenting results from various data formats e.g. text, audio, video
- [ ] Presentation of time-aware results (e.g. football games that are playing now instead of the results of the previous years)

4. **Trends in Evaluation (please select up to 3)**

Vink alle toepasselijke opties aan.

- [ ] Online evaluation - using online interaction loggings as way of evaluating relevance of the results
- [ ] Evaluation of Generated Information Objects - summarised content generated by a machine using various sources
- [ ] Crowdsourcing for generating labelled data
- [ ] Time and stream analysis - evaluation of real-time web search engines e.g. news
- [ ] Using eye-tracking for evaluation

5. **Trends in Web Search Engines (please select up to 4)**

Vink alle toepasselijke opties aan.

- [ ] Personalized web search engines/user models - e.g. based on prior queries, knowledge, information needs
- [ ] Smaller, Faster and Better web search engines(incremental innovations)
- [ ] Cross-device search - continuation of a search session on another device
- [ ] Web Search Engines for supporting knowledge goals and decision making
- [ ] Fairness, accountability, confidentiality and transparency
- [ ] Conversational Search - increased interaction by using natural language in a conversational manner

In this section we will be thinking about the INNOVATIONS in the development of Web Search Engines that you know about. Please select from lists below the ones you would consider the MOST RELEVANT.

6. **Innovations in Crawling (please select 1)**
Vink alle toepasselijke opties aan.

- [ ] Application-based focused crawler - Crawlers that target a particular group e.g. medical, educational
- [ ] Learnable focused crawler - a trained crawler deciding the relevance of each page in order to assess which ones to crawl

7. **Innovations in Indexing (please select 1)**
Vink alle toepasselijke opties aan.

- [ ] Schema.org for semantic indexing standardization
- [ ] Combined data indexing - new representation models for text, entities and their relations

8. **Innovations in Ranking (please select up to 2)**
Vink alle toepasselijke opties aan.

- [ ] NLP for query understanding
- [ ] Google's Knowledge graph
- [ ] Word2Vec for semantic query understanding
- [ ] Continuous search - searching over a period

9. **Innovations in Evaluation (please select 1)**
Vink alle toepasselijke opties aan.

- [ ] Generalized model for evaluation metrics
- [ ] TREC Complex Answer Retrieval Track(for Generated Information Objects)

10. **Innovations in Web Search Engines (please select up to 2)**
Vink alle toepasselijke opties aan.

- [ ] Google news timeline
- [ ] Wikidata - a free and open knowledge base that can be read and edited by both humans and machines
- [ ] DuckDuck Go - privacy-aware search engine

Thinking about the CHALLENGES in the development of Web Search Engines that you know about, please select from lists below the ones you would consider the MOST RELEVANT.

11. **Challenges in Crawling (please select up to 6)**
Vink alle toepasselijke opties aan.

- [ ] Explore the "hidden web"
- [ ] Opacity between information seeker and provider (onion search, the dark web)
- [ ] Effectively deal with web spam
- [ ] Dealing with non-uniform structures(e.g. dealing with client side scripts, dealing with non HTML search interfaces)
- [ ] Crawling multimedia
- [ ] Lack of crawling standards
- [ ] Dealing with unstructured and semi-structured data (e.g. should collected data fit a schema or be stored in a schema)
- [ ] Crawling information on closed platforms - e.g. Facebook, Instagram
- [ ] Client-side applications with Angular, React and Vue.js
- [ ] Crawling open source data (linked open data)
- [ ] Crawling open software repositories

12. **Challenges in Indexing (please select up to 3)**
Vink alle toepasselijke opties aan.

- [ ] Standardization of semantic indexing - representation models for text, entities and their relations
- [ ] The need for new ways of indexing
- [ ] Index and schema representation for personal information access
- [ ] Indexing technique for hidden data
- [ ] Indexing while keeping privacy
- [ ] Rank based indexing

13. **Challenges in Ranking (please select up to 4)**
Vink alle toepasselijke opties aan.

- [ ] Semantic query understanding
- [ ] Entity oriented search - finding "things" instead of "strings"
- [ ] Providing an exact answer to queries
- [ ] Computing over aggregated personal data
- [ ] Learn-To-Rank with personalized data
- [ ] Coping with limited data in Learn-To-Rank
- [ ] Machine reasoning in learnable information retrieval
- [ ] Entity ranking of hetereogeneous data (e.g. tweets, tables)

14. **Challenges in Evaluation (please select up to 2)**
Vink alle toepasselijke opties aan.

- [ ] Better benchmarks for evaluation
- [ ] More advanced user modelling
- [ ] Using graded relevance to improve ground-truth quality and maximize feedback information
- [ ] Repeatability and reliability

15. **Challenges in Web Search Engines (please select up to 7)**

Vink alle toepasselijke opties aan.

- [ ] On demand anticipation, customization and personalization
- [ ] Dealing with noisy, low quality and contradictory content continuously uploaded to the web
- [ ] Search during a period (continuous search)
- [ ] Novelty detection
- [ ] Fairness
- [ ] Accountability
- [ ] Confidentiallity
- [ ] Transparency
- [ ] Machine learning for efficiency improvement
- [ ] Understanding and anticipating needs for personalized search
- [ ] Personalization of search
- [ ] Privacy, security and trust issues in personalized search
- [ ] Use of natural language

16. **Discussion & Feedback**

We are almost finished. I would like you to answer the questions below for the purposes of linking to the previous stage of the survey.

The data collection will respect the legislation of data protection. No data will be shared and will solely be used for follow-ups and classification of data. Data will not be linked outside the scope of these surveys.

17. **Name** *

18. **E-mail address** *

Appendix 4: Template survey 3 academics

Ranking the trends

Thank you for participating in the third round of these survey rounds. Your opinion is valuable for the continuation of this research! The format of this survey will be different due to Google Forms inability to rank items in an aesthetic manner. Therefore I had to switch to the services of KwikSurveys.

In the previous round, we started with 87 factors. You have been asked to select half of the factors in each category, resulting in a remaining list of the most relevant factors.

For this survey round the objective is to rank the list of factors per category, by dragging and dropping the factors ordered on importance in your opinion. As such, the factor ranked first will be seen as the most important, the second as the second most important factor, and so forth. If the factors are in the right order, tick the "finished sorting?" button and continue to the next question.

Feedback and discussion are very welcome as well. At the end of the survey, there is an option to deliver feedback and/or discussion on the results of the survey.

1* Rank the trends in Crawling in terms of importance

- ☐ Usage of topical crawlers - crawling based on topical relevance
- ☐ Usage of machine learning to enhance crawlers
- ☐ Usage of focused crawlers - focused crawler technique gives priority to those URLs in the process of crawling, in which probability of finding information of user’s interest is high (e.g. based on link, text and URL)

2* Rank the trends in Ranking in terms of importance

- ☐ Question answering system
- ☐ Generated Information Objects - summarized content generated by a machine using various sources
- ☐ Learn-to-Rank(machine learning) - automatically construct a ranking model using training data applicable to new documents
- ☐ Explainable search - explain the results of the search
- ☐ Social Interaction for system's optimization - exploring social web for making new connections between images and tags
- ☐ Semantic query processing - understanding the user's query based on a knowledge base rather than matching documents directly on search terms
- ☐ Presentation of time-aware results (e.g. football games that are playing now instead of the results of the previous years)

3* Rank the trends in Evaluation in terms of importance

- [ ] Time and stream analysis - evaluation of real-time web search engines e.g. news
- [ ] Crowdsourcing for generating labelled data
- [ ] Online evaluation - using online interaction loggings as way of evaluating relevance of the results

4* Rank the trends in Web Search Engines in terms of importance

- [ ] Fairness, accountability, confidentiality and transparency
- [ ] Conversational Search - increased interaction by using natural language in a conversational manner
- [ ] Cross-device search - continuation of a search session on another device

5* Rank the innovations in terms of importance

- [ ] Combined data indexing - new representation models for text, entities and their relations
- [ ] Wikidata - a free and open knowledge base that can be read and edited by both humans and machines
- [ ] Learnable focused crawler - a trained crawler deciding the relevance of each page in order to assess which ones to crawl
- [ ] TREC Complex Answer Retrieval Track(for Generated Information Objects)
- [ ] Generalized model for evaluation metrics
- [ ] NLP for query understanding
- [ ] DuckDuck Go - privacy-aware search engine

6* Rank the challenges in Crawling in terms of importance

- [ ] Dealing with unstructured and semi-structured data (e.g. should collected data fit a schema or be stored in a schema)
- [ ] Lack of crawling standards
- [ ] Client-side applications with Angular, React and Vue.js
- [ ] Crawling information on closed platforms - e.g. Facebook, Instagram
- [ ] Crawling multimedia
- [ ] Explore the "hidden web"

7* Rank the challenges in Indexing in terms of importance

- [ ] Standardization of semantic indexing - representation models for text, entities and their relations
- [ ] The need for new ways of indexing
- [ ] Indexing while keeping privacy

8* Rank the challenges in Ranking in terms of importance

- [ ] Providing an exact answer to queries
- [ ] Semantic query understanding
- [ ] Machine reasoning in learnable information retrieval
- [ ] Learn-To-Rank with personalized data

9* Rank the challenges in Evaluation in terms of importance

- [ ] Repeatability and reliability
- [ ] More advanced user modelling

10* Rank the challenges in Web Search Engines in terms of importance

- [ ] Novelty detection
- [ ] Dealing with noisy, low quality and contradictory content continuously uploaded to the web
- [ ] Transparency
- [ ] Use of natural language
- [ ] Fairness
- [ ] Privacy, security and trust issues in personalized search
- [ ] Machine learning for efficiency improvement

11) Comments and thoughts of the participants often give us relevant insights and food for thought, below is a shared comment of the previous round:

"Better understanding of search users, be it based on personalization, NLP, machine learning, etc. is the key in developing the next generation search engine. It's the same evolution from Altavista to Google - and it will be transparent over any data source or device. Combine that with all current trends of assistants and conversational AI's and the future is here. It's also the biggest threat to free choice and diversity - it will funnel us all into 'personal containers' and alienate us from each other and the less favorable content the world has to offer... Just my 2 cents "
- anonymous participant

11 If you could give any opinions, discussion points or remarks, please feel free to do so below.

12* What is your e-mail address?

Appendix 5: Template survey 3 industry

Ranking the trends

Thank you for participating in the third round of these survey rounds. Your opinion is valuable for the continuation of this research! The format of this survey will be different due to Google Forms inability to rank items in an aesthetic manner. Therefore I had to switch to the services of KwikSurveys.

In the previous round, we started with 87 factors. You have been asked to select half of the factors in each category, resulting in a remaining list of the most relevant factors.

For this survey round the objective is to rank the list of factors per category, by dragging and dropping the factors ordered on importance in your opinion. As such, the factor ranked first will be seen as the most important, the second as the second most important factor, and so forth. If the factors are in the right order, tick the "finished sorting?" button and continue to the next question.

Feedback and discussion are very welcome as well. At the end of the survey, there is an option to deliver feedback and/or discussion on the results of the survey.

1* Rank the trends in Crawling in terms of importance

- ☐ Usage of machine learning to enhance crawlers
- ☐ Usage of focused crawlers - focused crawler technique gives priority to those URLs in the process of crawling, in which probability of finding information of user's interest is high (e.g. based on link, text and URL)
- ☐ Usage of topical crawlers - crawling based on topical relevance

2* Rank the trends in Ranking in terms of importance

- ☐ End-to-End learning - running a machine learning model on a task in a end-to-end fashion e.g. input a query output documents
- ☐ Social Interaction for system's optimization - exploring social web for making new connections between images and tags
- ☐ Learn-to-Rank(machine learning) - automatically construct a ranking model using training data applicable to new documents
- ☐ Guided user-interface
- ☐ Entity oriented search - finding "things" instead of "strings"
- ☐ Presentation of time-aware results (e.g. football games that are playing now instead of the results of the previous years)

3* Rank the trends in Evaluation in terms of importance

- [ ] Online evaluation - using online interaction loggings as way of evaluating relevance of the results
- [ ] Crowdsourcing for generating labelled data
- [ ] Time and stream analysis - evaluation of real-time web search engines e.g. news

4* Rank the trends in Web Search Engines in terms of importance

- [ ] Conversational Search - increased interaction by using natural language in a conversational manner
- [ ] Personalized web search engines/user models - e.g. based on prior queries, knowledge, information needs
- [ ] Fairness, accountability, confidentiality and transparency

5* Rank the innovations in terms of importance

- [ ] Learnable focused crawler - a trained crawler deciding the relevance of each page in order to assess which ones to crawl
- [ ] Google's Knowledge graph
- [ ] Combined data indexing - new representation models for text, entities and their relations
- [ ] NLP for query understanding
- [ ] Application-based focused crawler - Crawlers that target a particular group e.g. medical, educational
- [ ] Generalized model for evaluation metrics
- [ ] Google news timeline

6* Rank the challenges in Crawling in terms of importance

- [ ] Explore the "hidden web"
- [ ] Crawling information on closed platforms - e.g. Facebook, Instagram
- [ ] Dealing with non-uniform structures(e.g. dealing with client side scripts, dealing with non HTML search interfaces)
- [ ] Crawling multimedia
- [ ] Dealing with unstructured and semi-structured data (e.g. should collected data fit a schema or be stored in a schema)
- [ ] Client-side applications with Angular, React and Vue.js

7* Rank the challenges in Indexing in terms of importance

- [ ] Index and schema representation for personal information access
- [ ] Indexing while keeping privacy
- [ ] Standardization of semantic indexing - representation models for text, entities and their relations

8* Rank the challenges in Ranking in terms of importance

- [ ] Semantic query understanding
- [ ] Entity ranking of heterogeneous data (e.g. tweets, tables)
- [ ] Entity oriented search - finding "things" instead of "strings"
- [ ] Providing an exact answer to queries

9* Rank the challenges in Evaluation in terms of importance

- [ ] More advanced user modelling
- [ ] Repeatability and reliability

10* Rank the challenges in Web Search Engines in terms of importance

- [ ] Dealing with noisy, low quality and contradictory content continuously uploaded to the web
- [ ] Machine learning for efficiency improvement
- [ ] Personalization of search
- [ ] Use of natural language
- [ ]
- [ ] Fairness
- [ ] Privacy, security and trust issues in personalized search

11) Comments and thoughts of the participants often give us relevant insights and food for thought, below is a shared comment of the previous round:

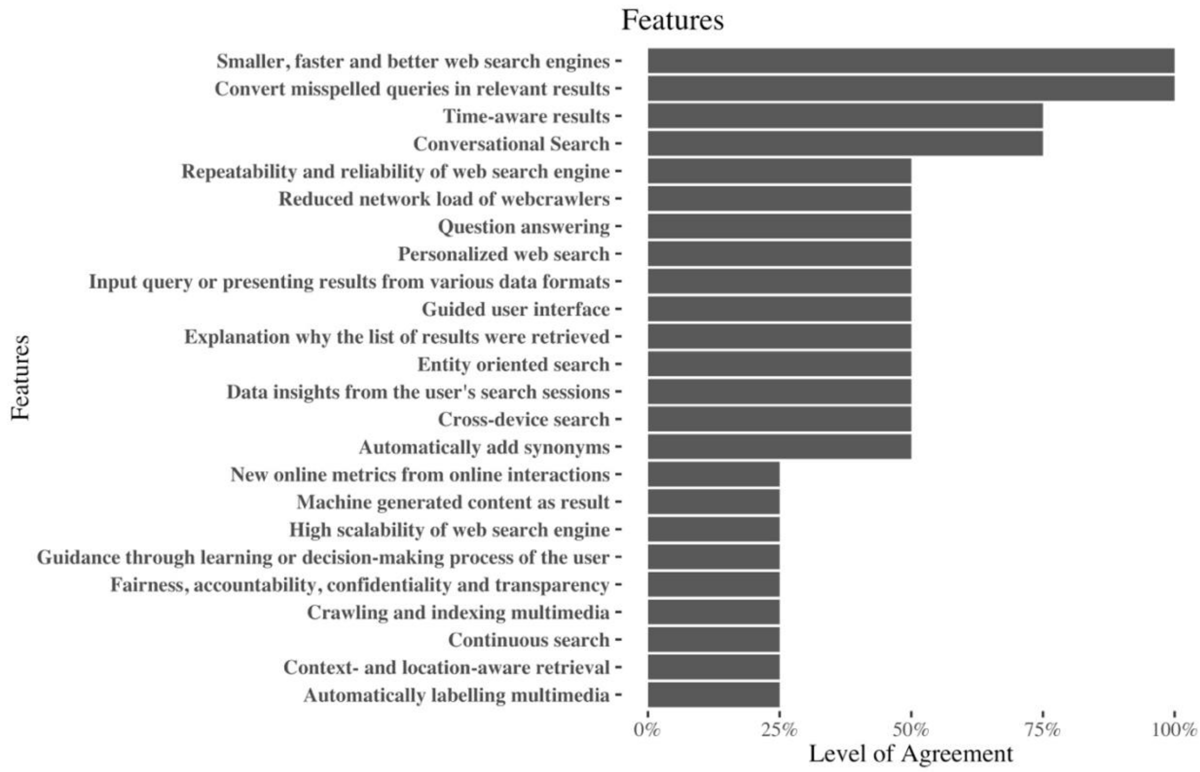
"Better understanding of search users, be it based on personalization, NLP, machine learning, etc. is the key in developing the next generation search engine. It's the same evolution from Altavista to Google - and it will be transparent over any data source or device. Combine that with all current trends of assistants and conversational AI's and the future is here. It's also the biggest threat to free choice and diversity - it will funnel us all into 'personal containers' and alienate us from each other and the less favorable content the world has to offer... Just my 2 cents "
- anonymous participant

11 If you could give any opinions, discussion points or remarks, please feel free to do so below.

12* What is your e-mail address?

Appendix 6: 1st Survey results analysis, clients

In the bar plots that have been made the features with their corresponding level of agreement are listed. Since the sample size only consisted out of 4 participants, the results can only be taken as an indication. All the participants were working for a company which purchased search services from a third party, additionally they were actively involved in the implementation thereof.

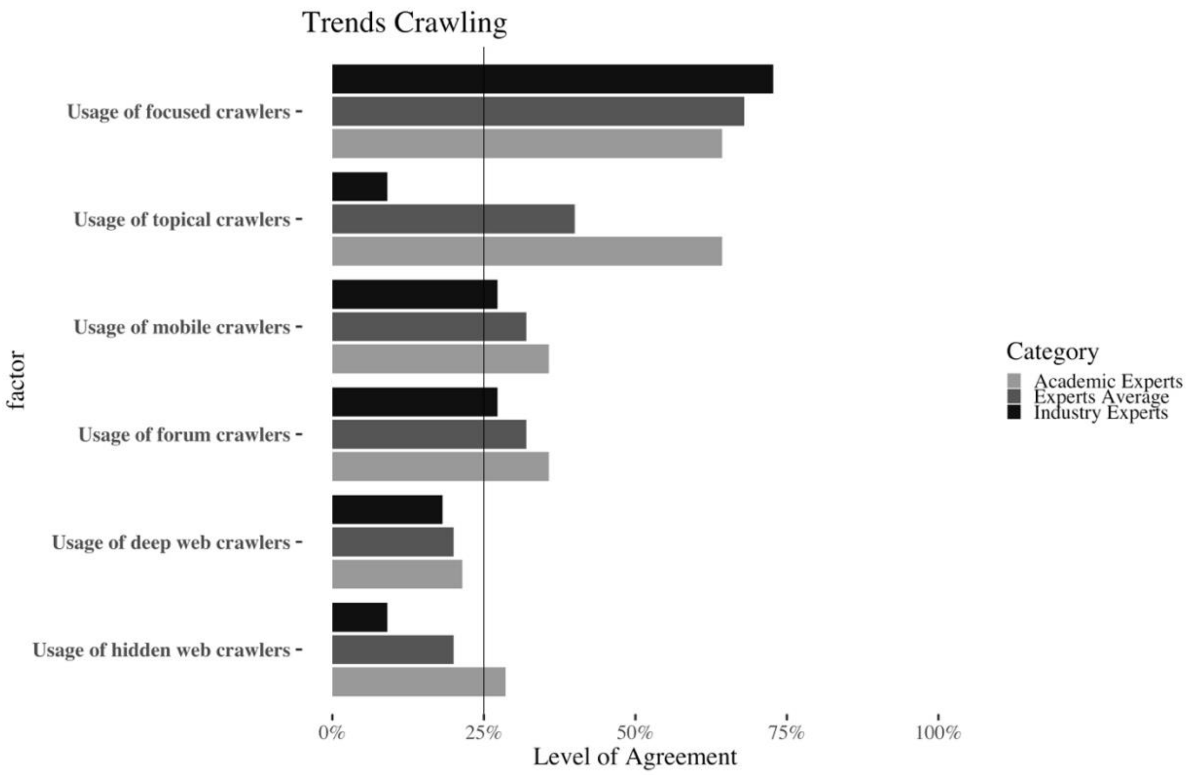


Clients seem to be very practical centered by having a 100% level of agreement in incremental innovations as “smaller, faster and better web search engines” and “convert misspelled queries in relevant results”. However, the more radical feature, conversational search is also listed high with a level of agreement of 75%. Conclusively, when looking at all the features with a level of agreement of 50% or higher, one could argue that 9 out of the chosen 15 features can be described as incremental innovation. Resulting in the hypothesis clients are more centered on the immediate added value in contrast with the long-term vision of a completely redesign of search systems.

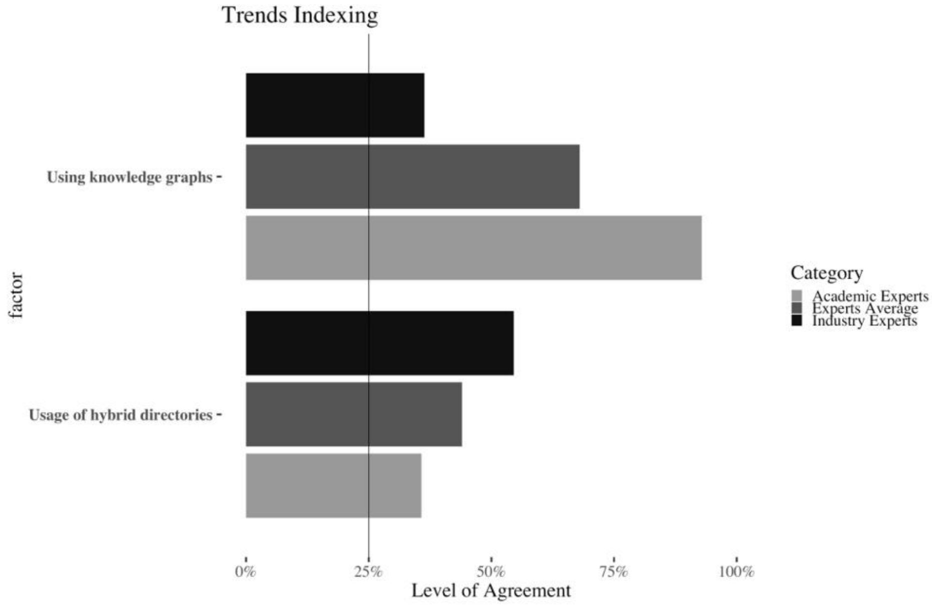
Appendix 7: 1st Survey results analysis, academic- & industry-experts

In the following sections the results of the first survey are discussed. The results are simply visualized by showing bar plots with the level of agreement per factor. These plots are colored by the groups: academics, industry and the average of them both taken together. The two groups had some similarity, however on some points strongly disagreed. Moreover, the level of engagement of academics is significantly higher (37% vs 17%), connecting themselves to more factors compared to the industry experts. All the factors with an expert's average above the threshold of 25% (see thin lines bar plot) are kept for the following round of the survey.

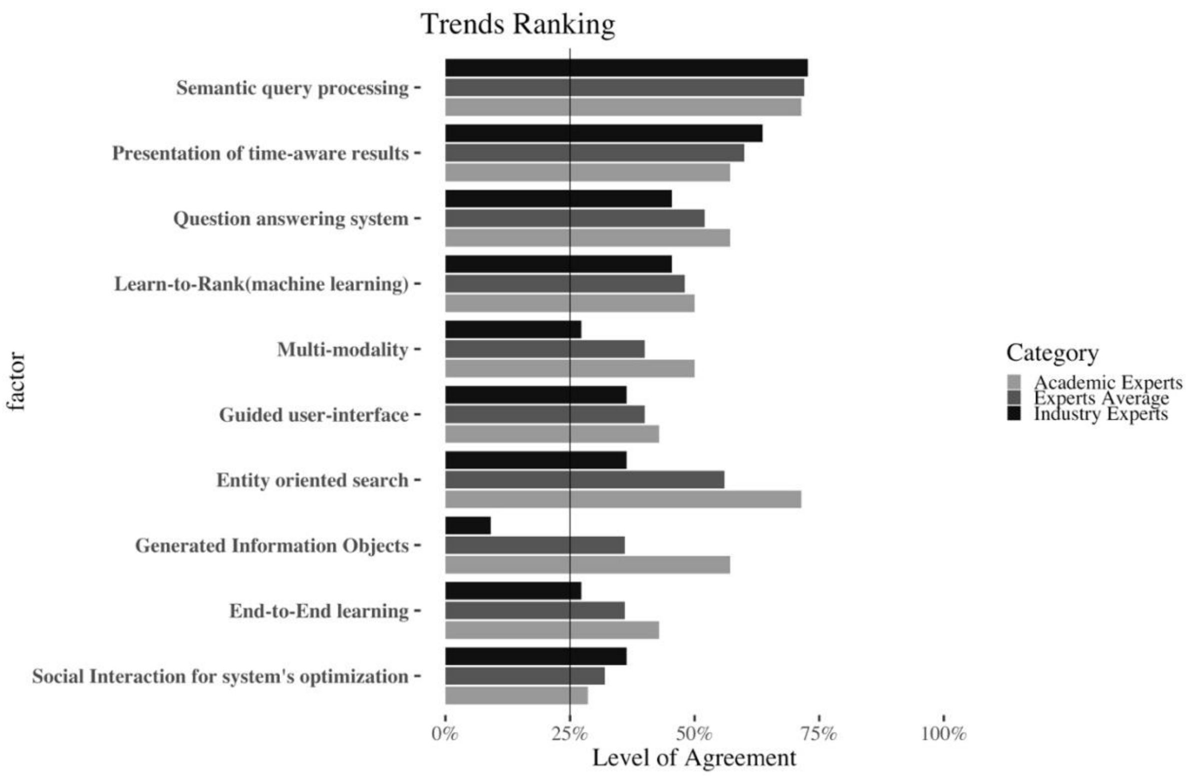
Trends crawling



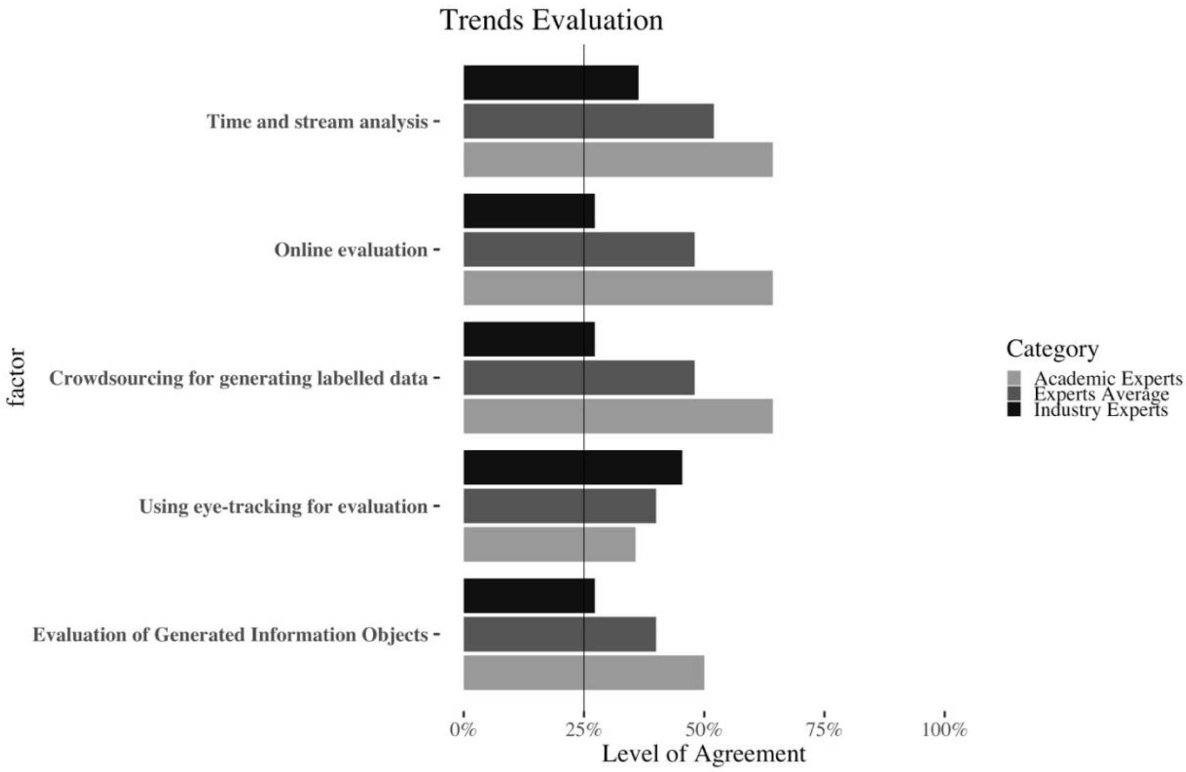
Trends indexing



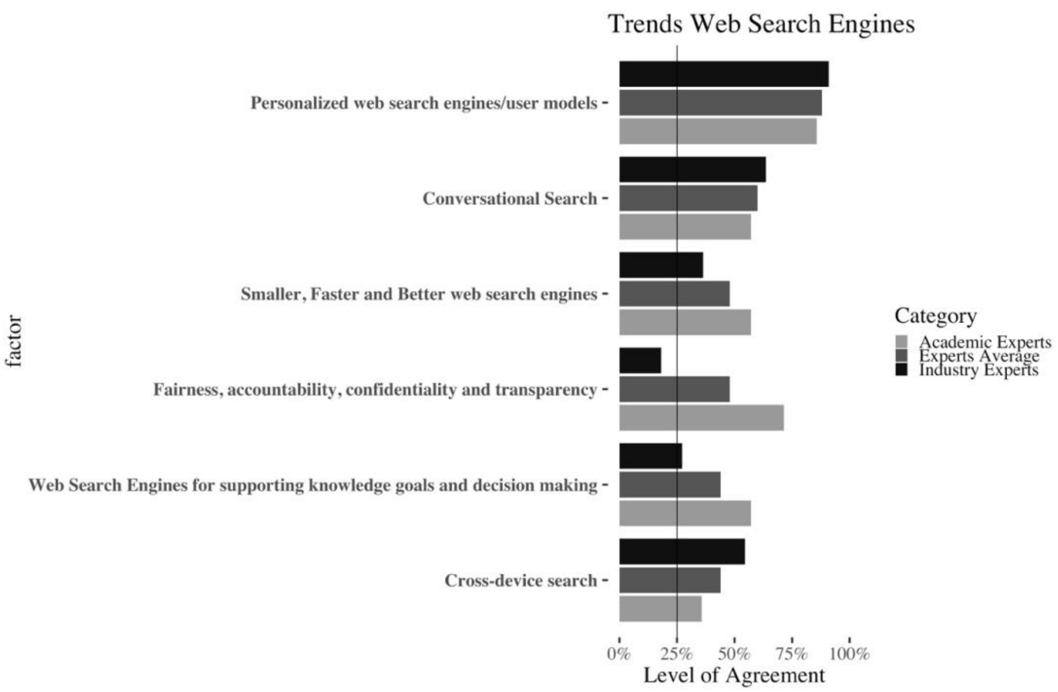
Trends ranking



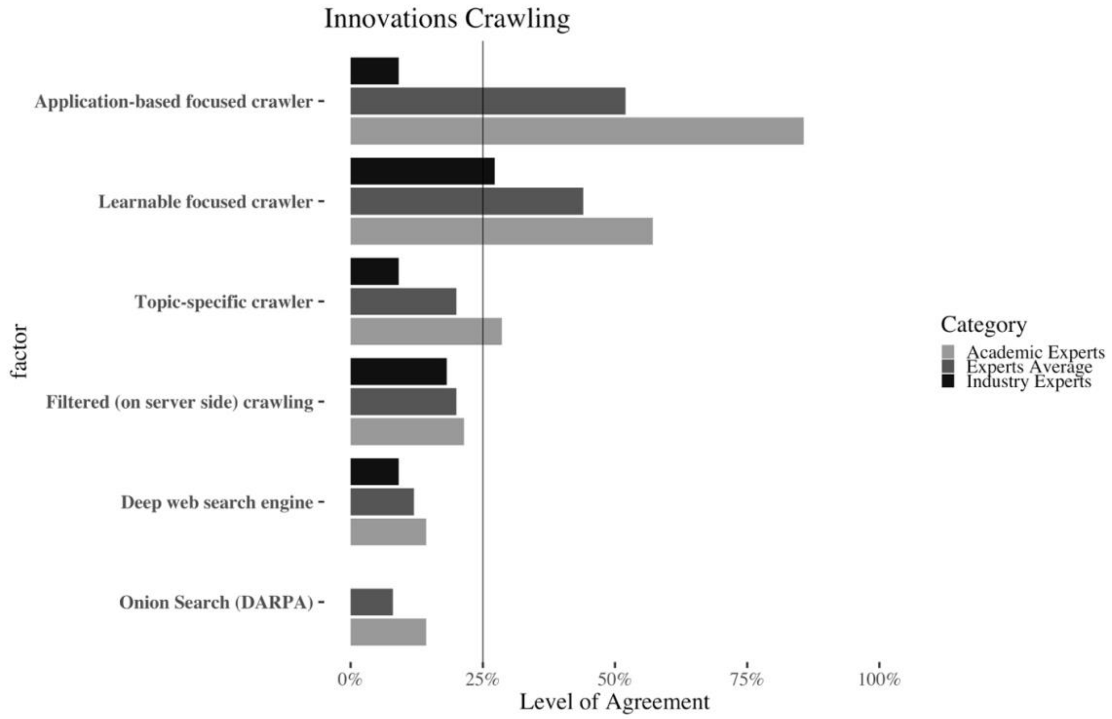
Trends evaluation



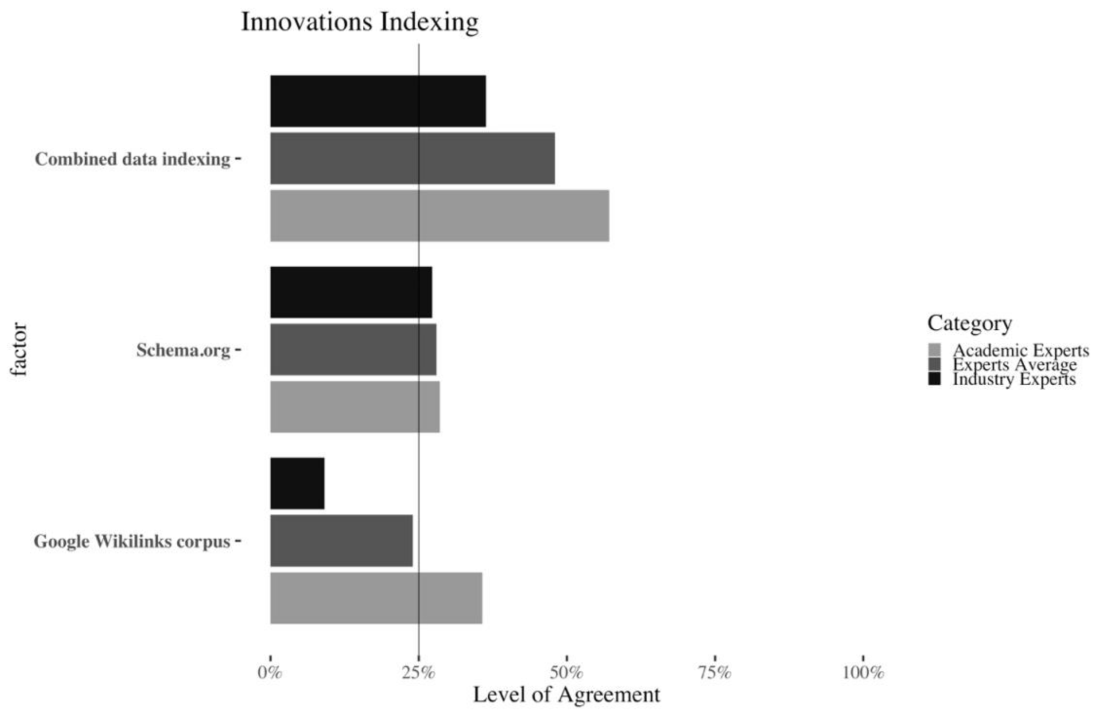
Trends web search engines



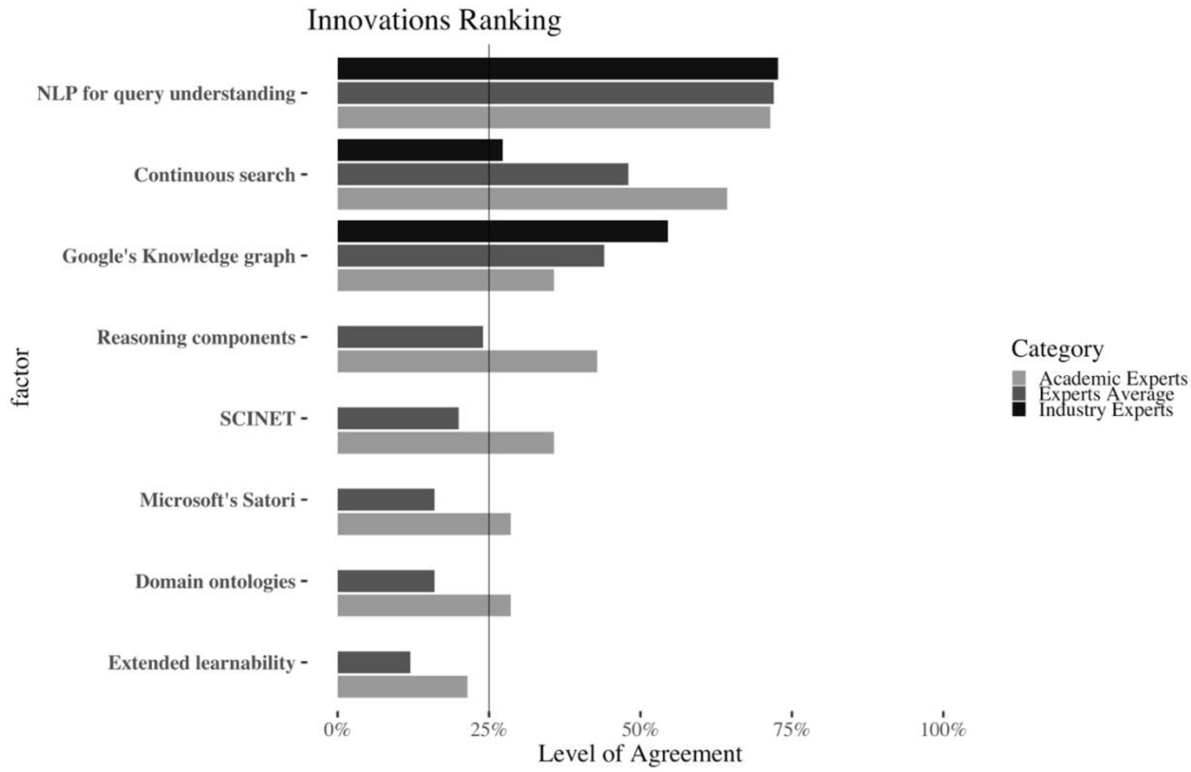
Innovations crawling



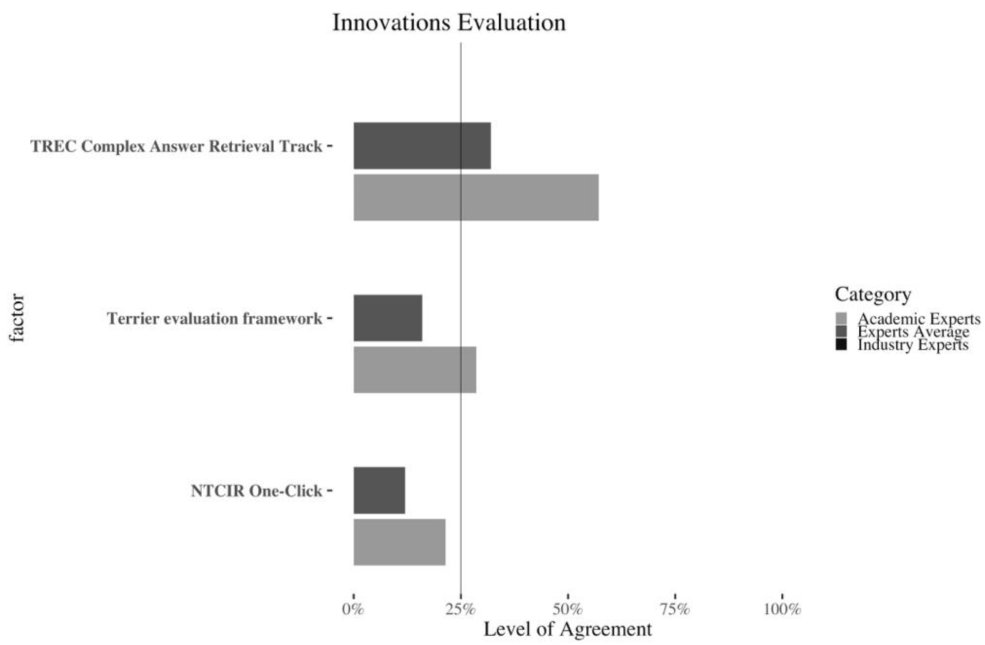
Innovations indexing



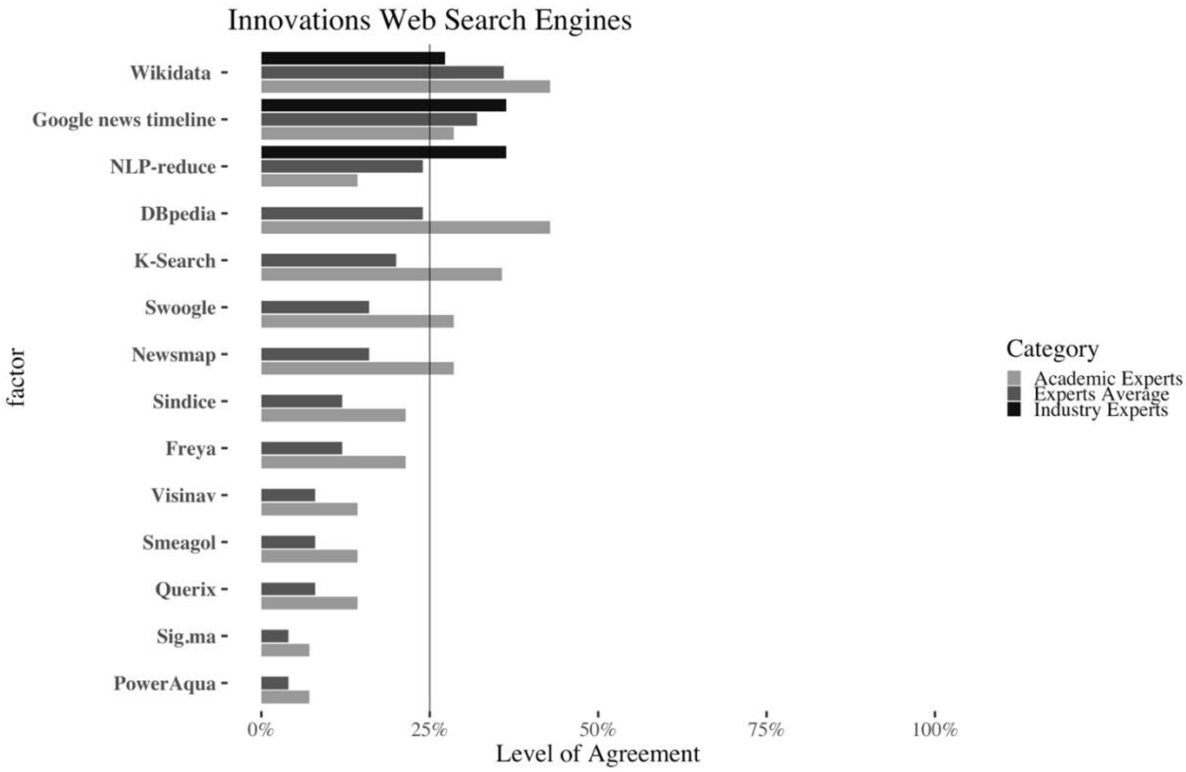
Innovations ranking



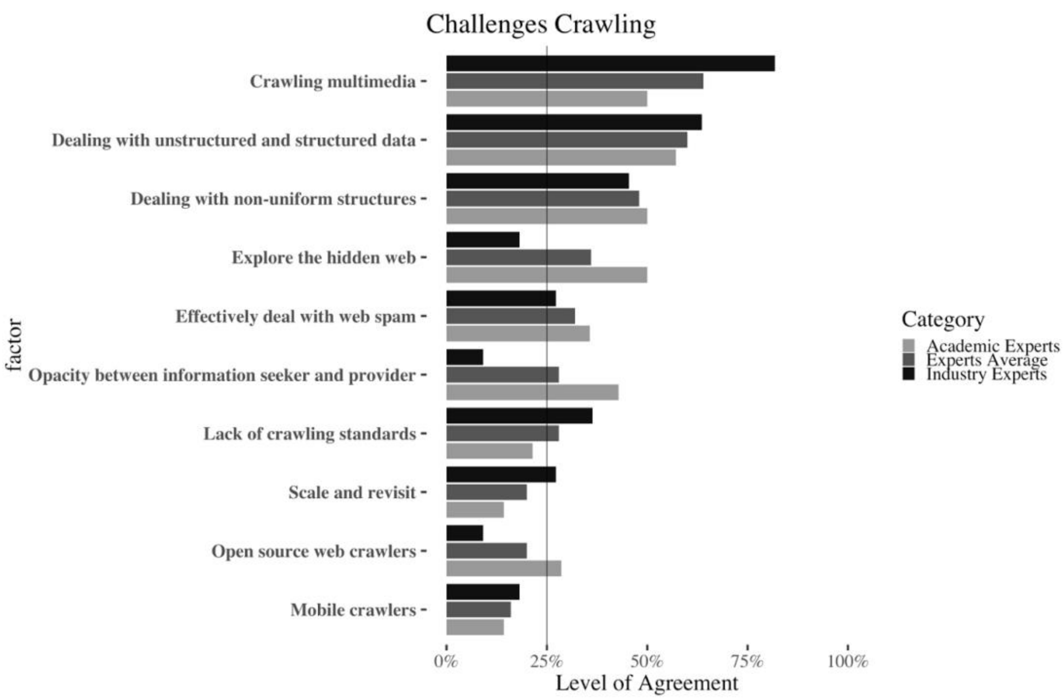
Innovations evaluation



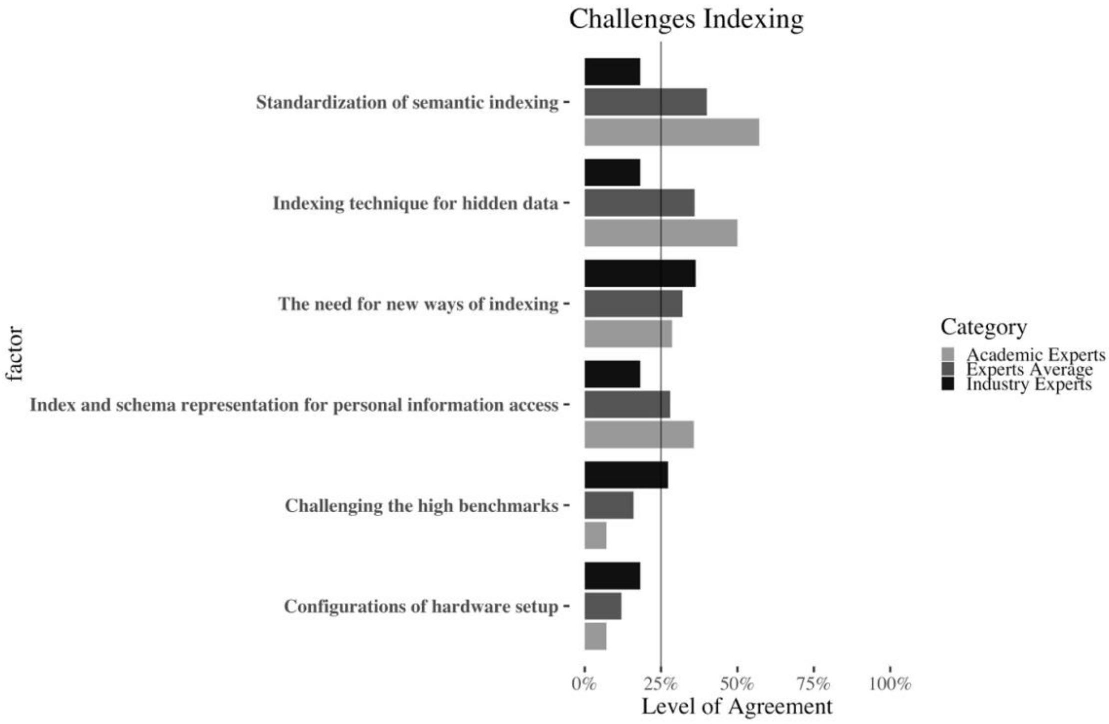
Innovations web search engines



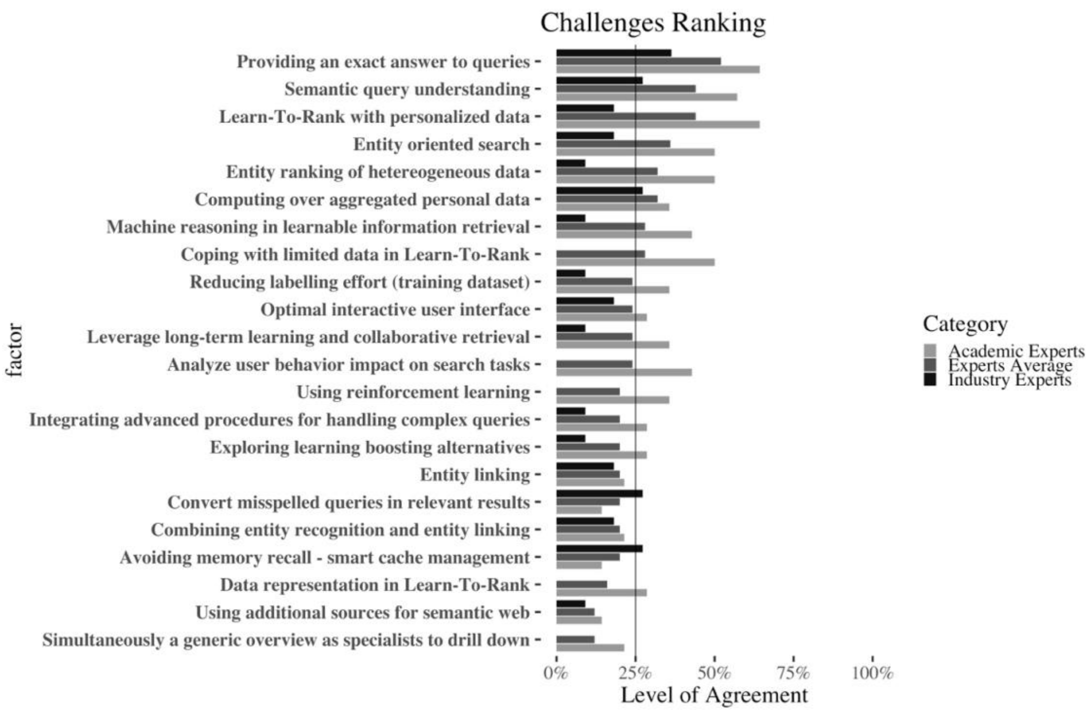
Challenges crawling



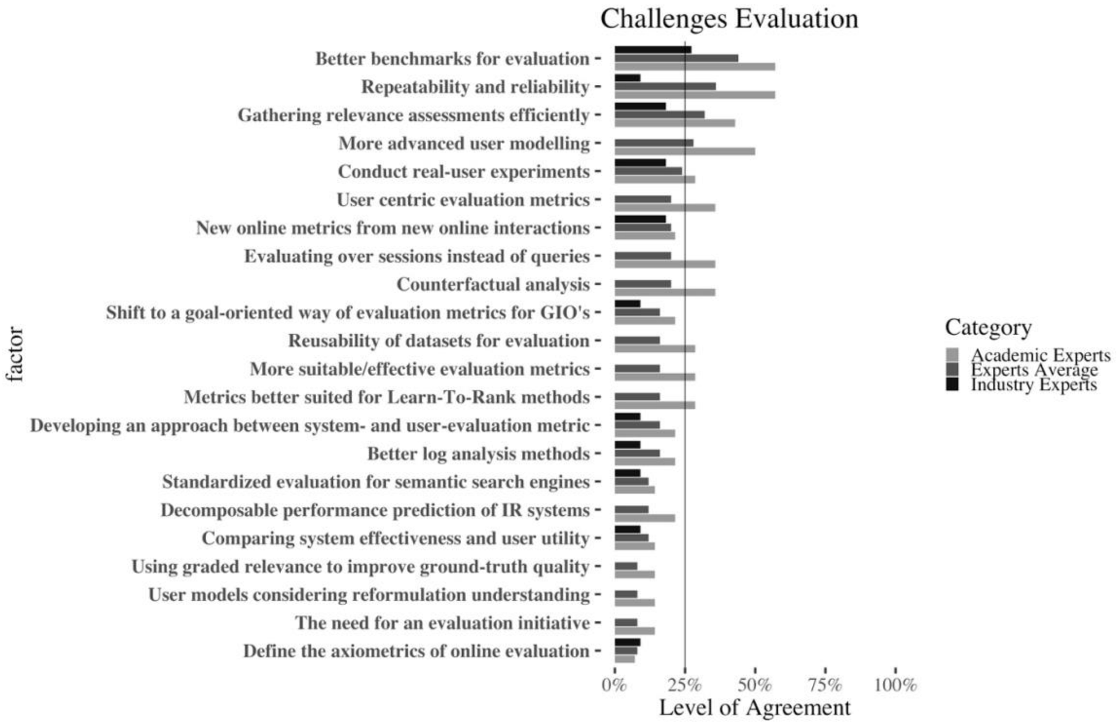
Challenges indexing



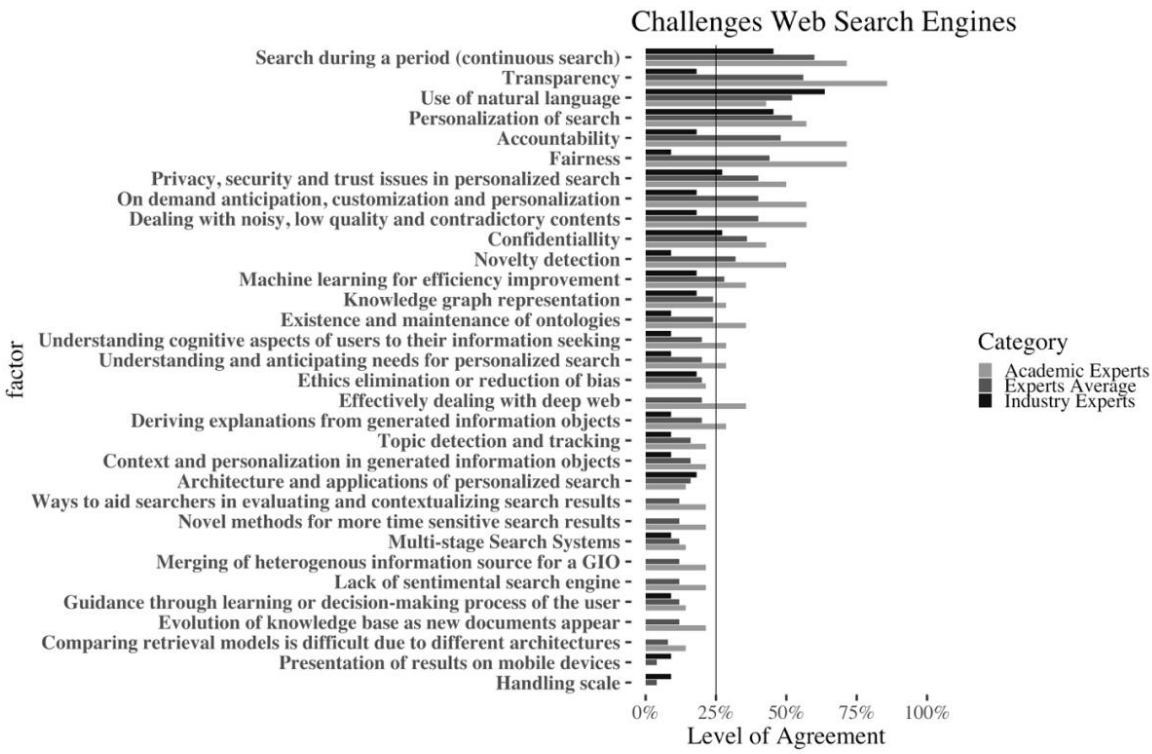
Challenges ranking



Challenges evaluation



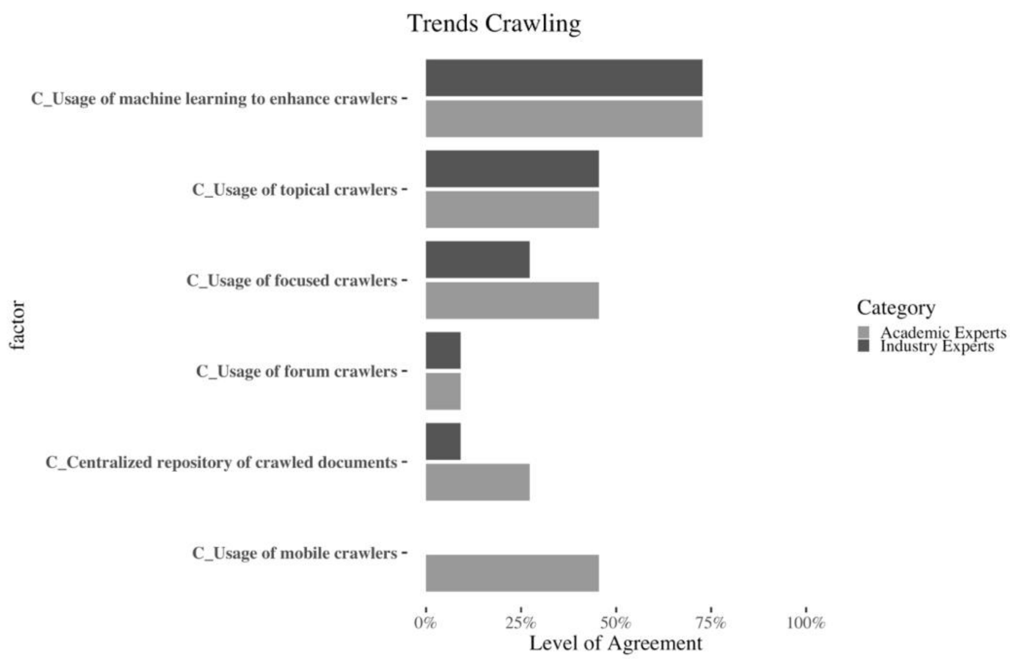
Challenges web search engines



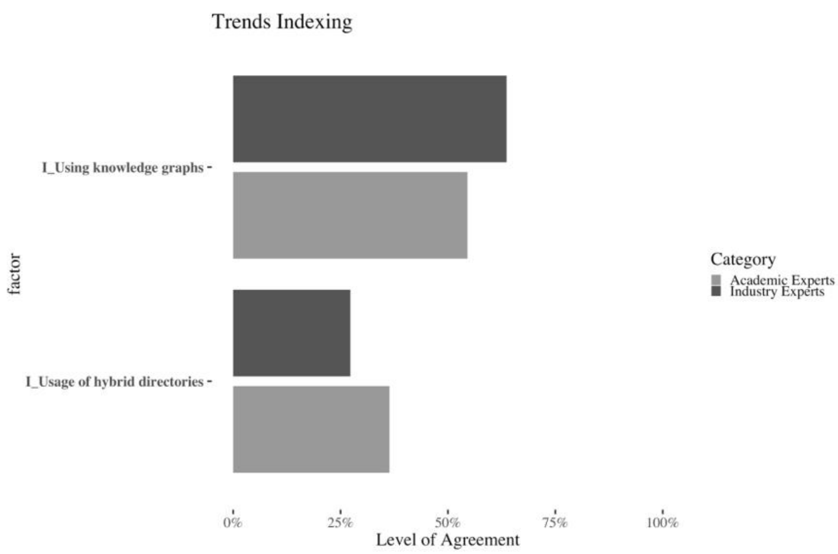
Appendix 8: 2nd Survey results analysis, academic- & industry-experts

The results of the 2nd survey are analyzed in a similar manner as the first survey, however now the factors are selected per panel on the basis of rank (only the top half) per subcategory (trends crawling, trends indexing, etc.).

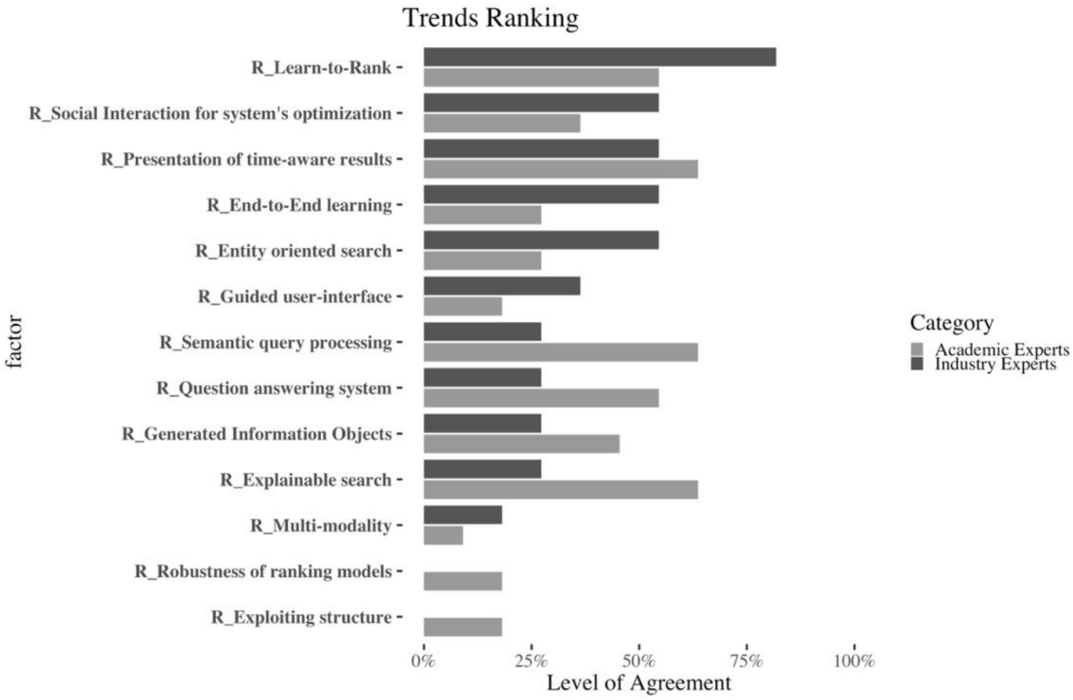
Trends crawling



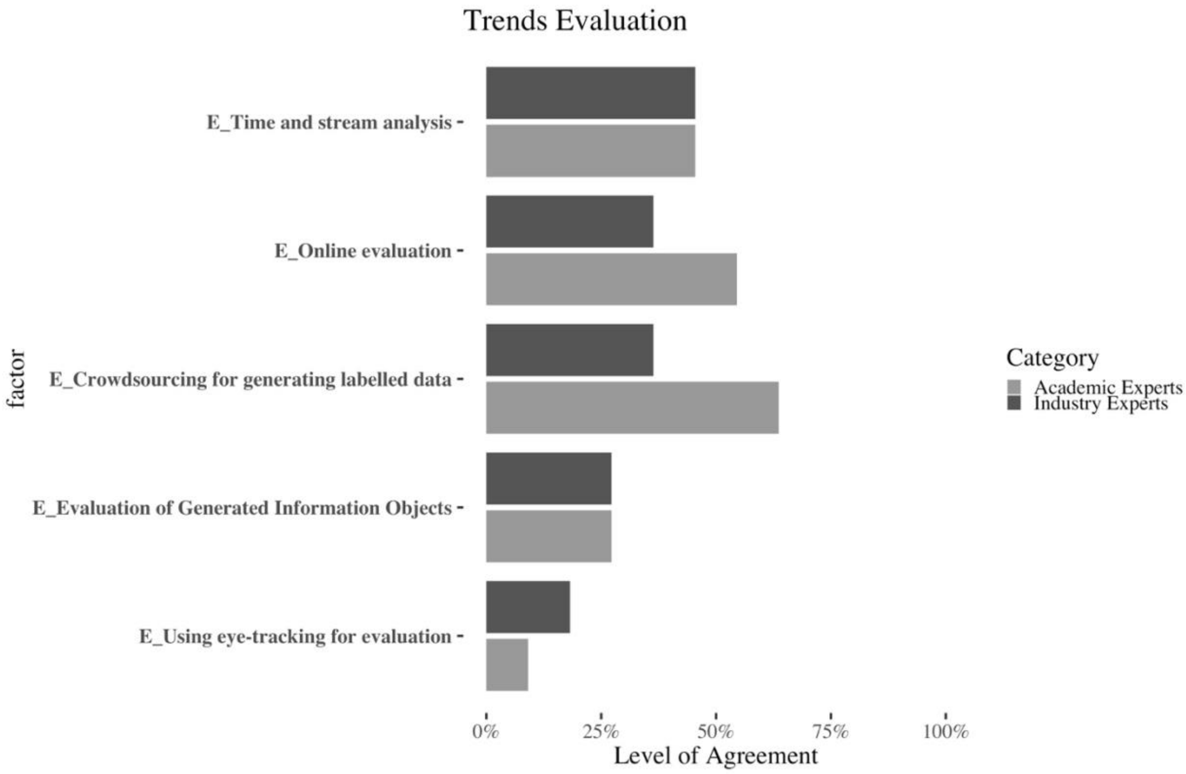
Trends indexing



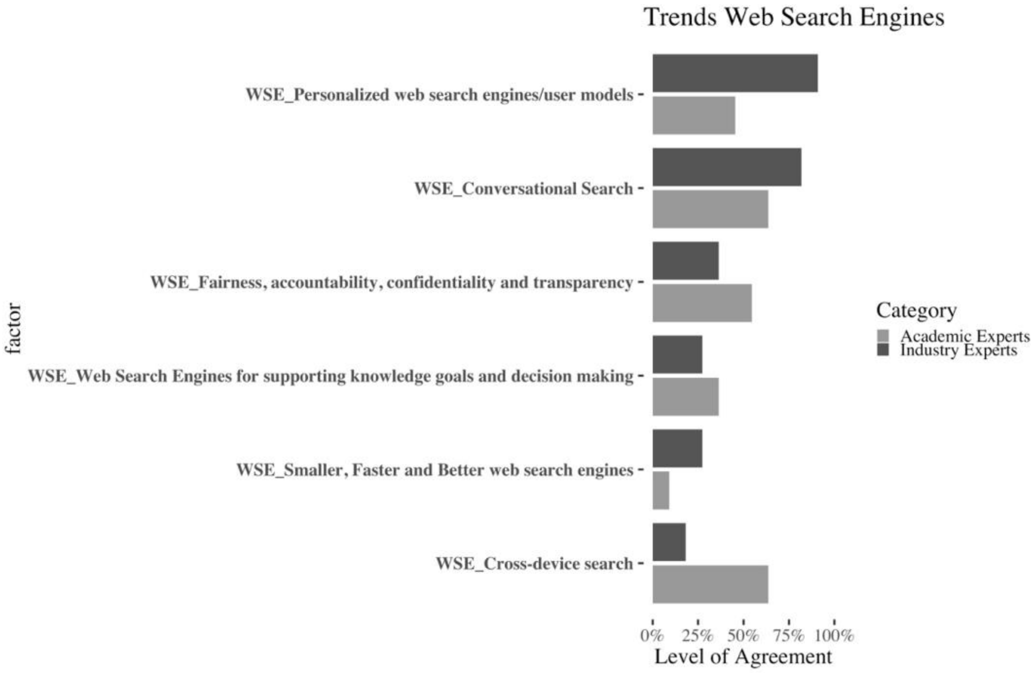
Trends ranking



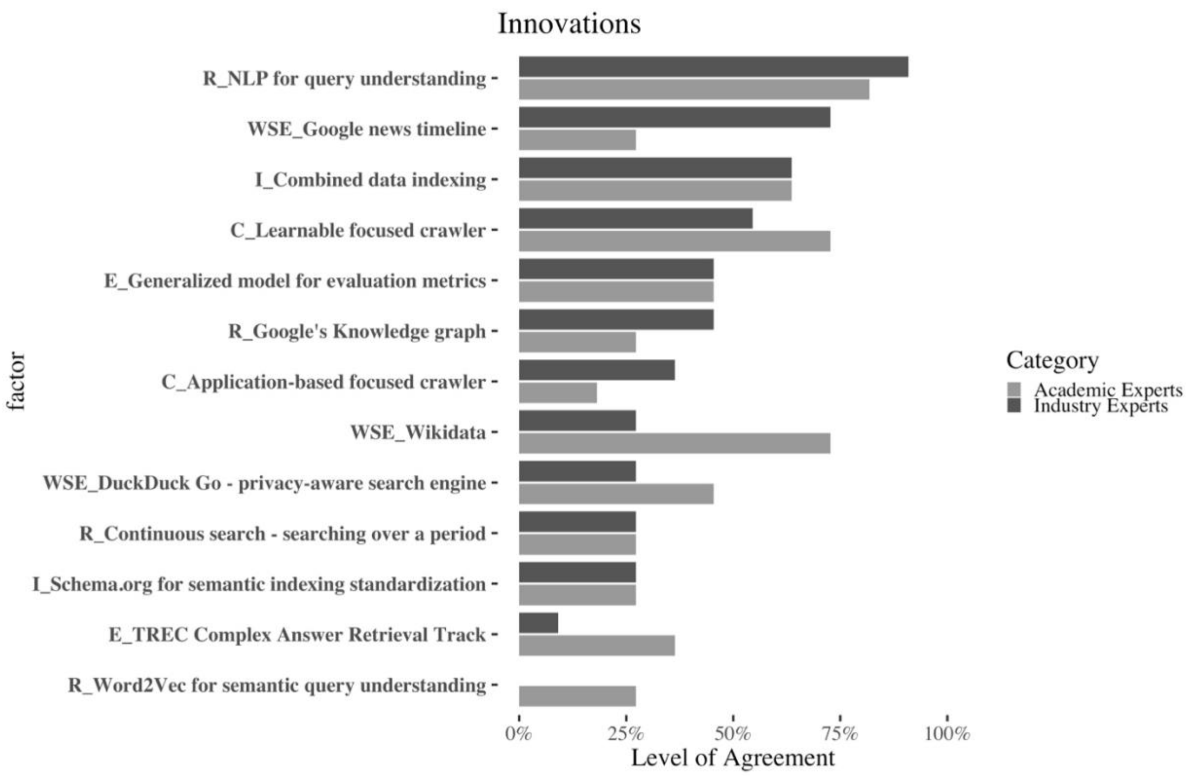
Trends evaluation



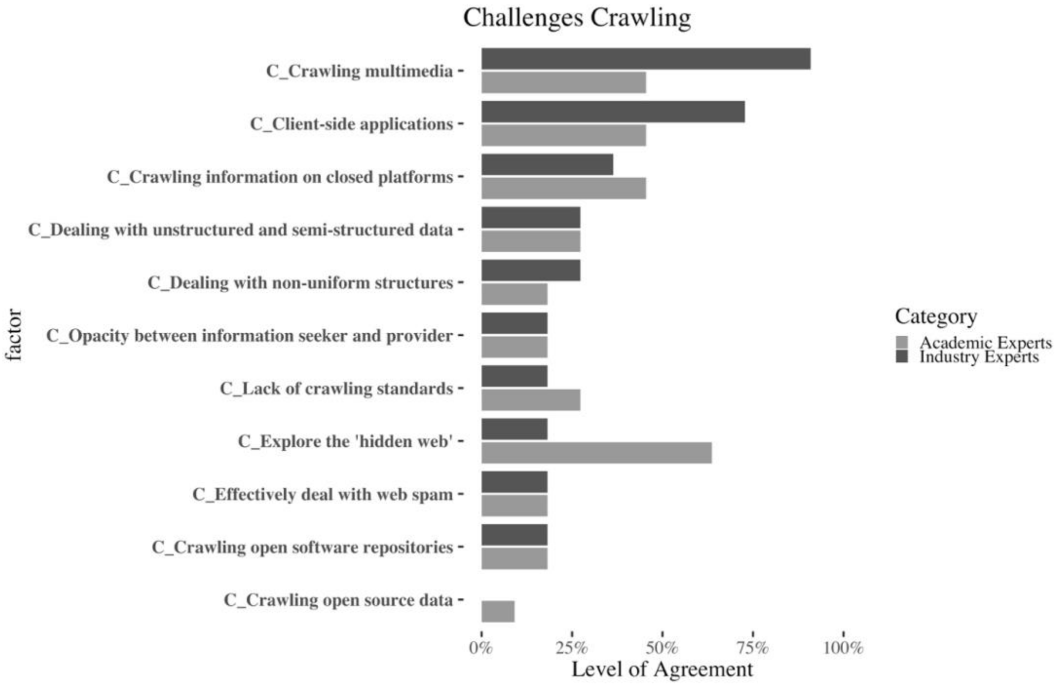
Trends web search engines



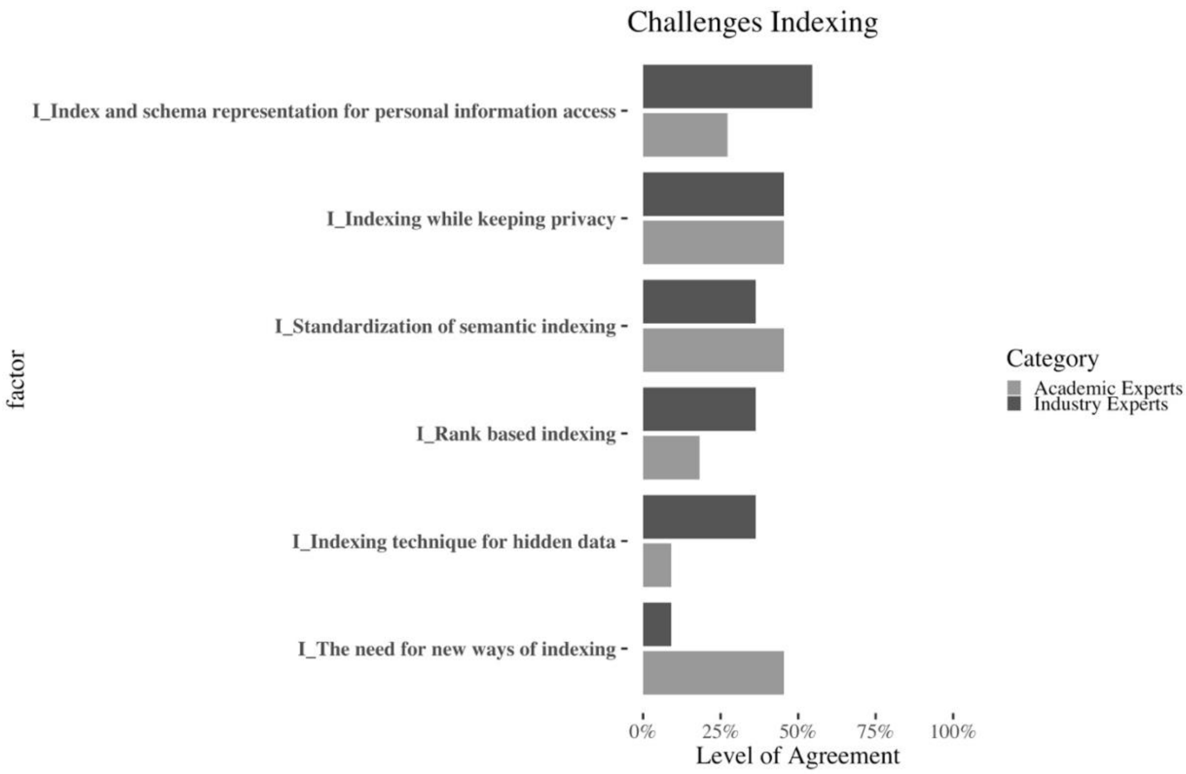
Innovations



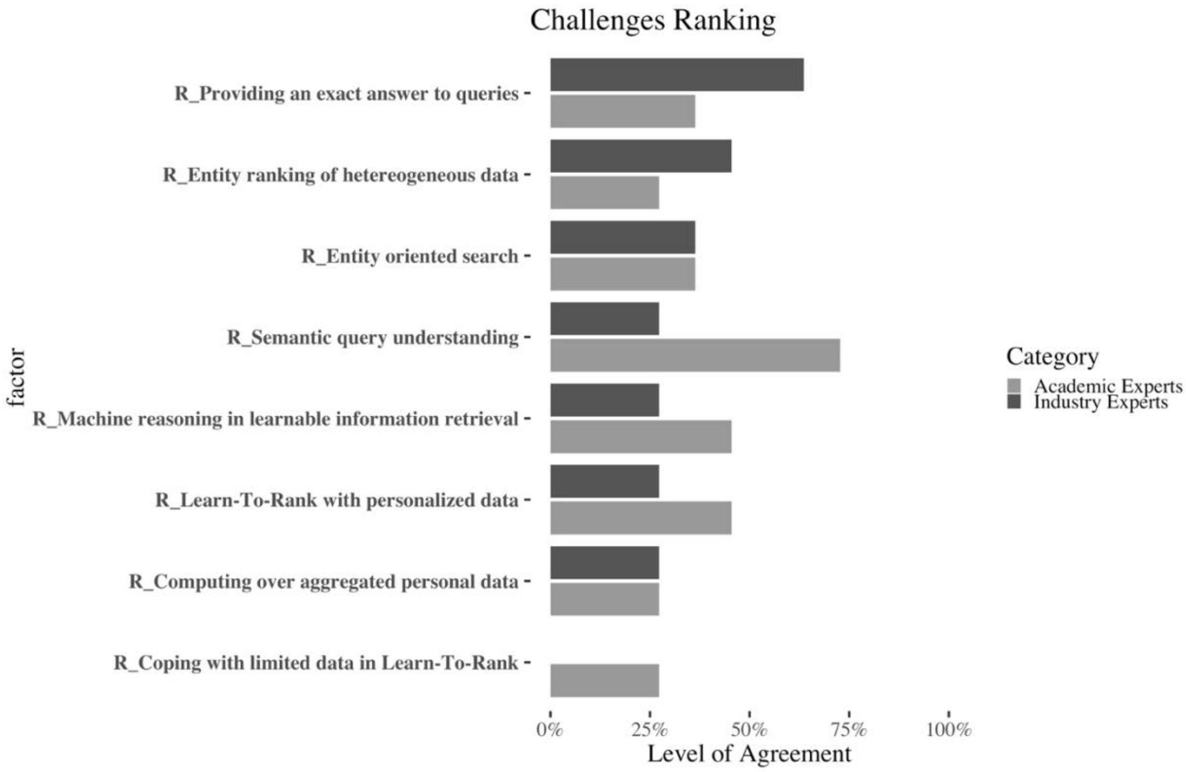
Challenges crawling



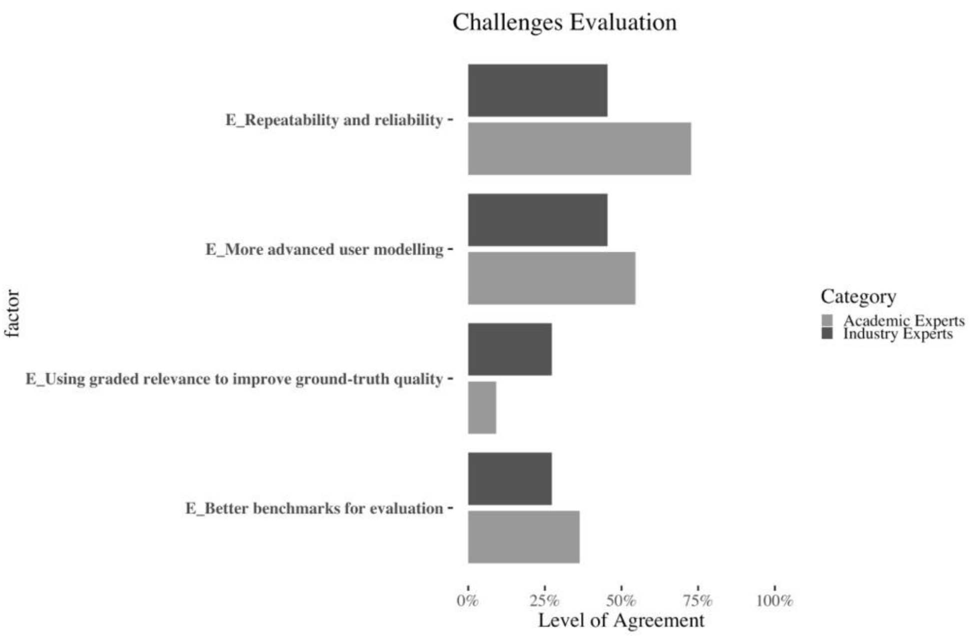
Challenges indexing



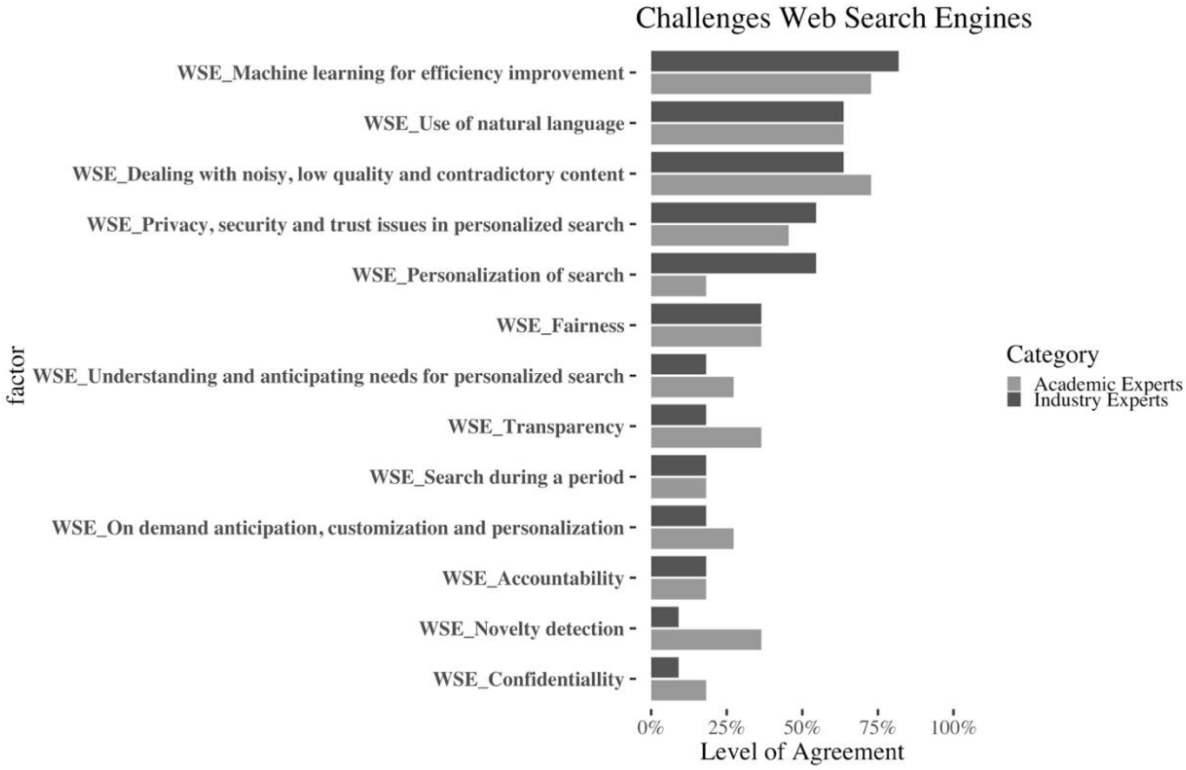
Challenges ranking



Challenges evaluation



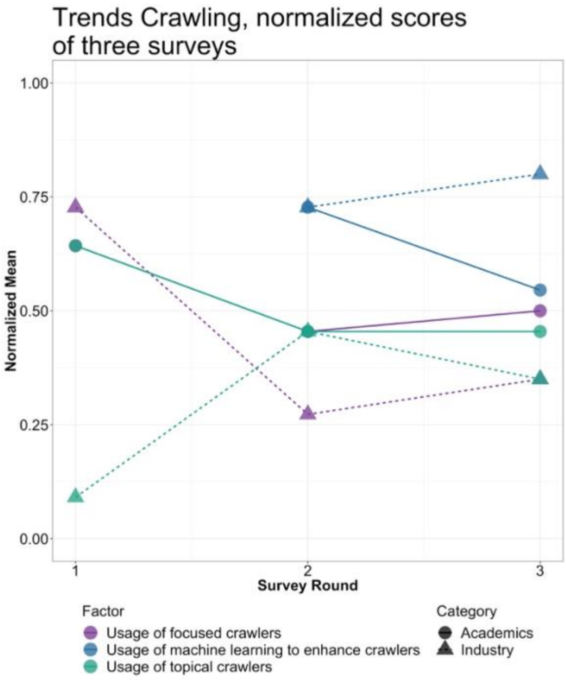
Challenges web search engines


Challenges Web Search Engines
WSE_Machine learning for efficiency improvement
WSE_Use of natural language
WSE_Dealing with noisy, low quality and contradictory content
WSE_Privacy, security and trust issues in personalized search
WSE_Personalization of search
WSE_Fairness
WSE_Understanding and anticipating needs for personalized search
WSE_Transparency
WSE_Search during a period
WSE_On demand anticipation, customization and personalization
WSE_Accountability
WSE_Novelty detection
WSE_Confidentiallity
factor
0%
25%
50%
75%
100%
Level of Agreement
Category
Academic Experts
Industry Experts

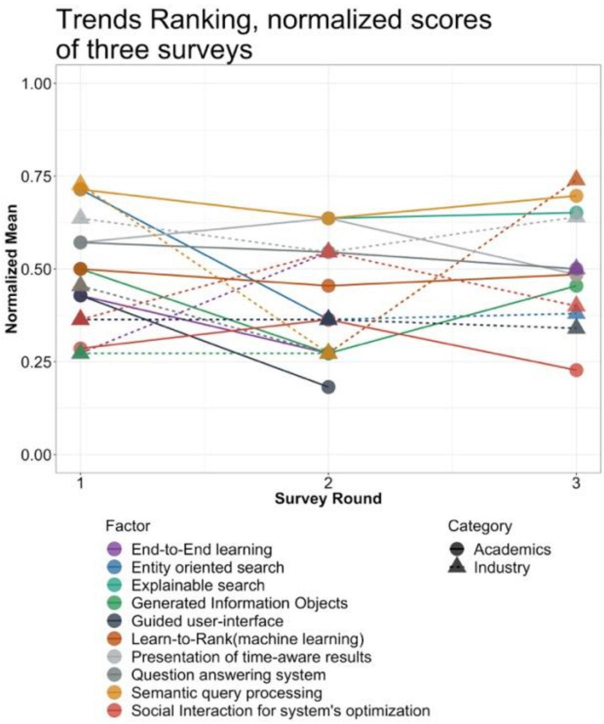
Appendix 9: Analysis of all three survey rounds compared to each other

The following plots describe the progress of the factors over the three consecutive survey rounds. The values are normalized and made compatible to construct an egal manner of reading the plots. A higher normalized mean means a higher support/importance of the factor.

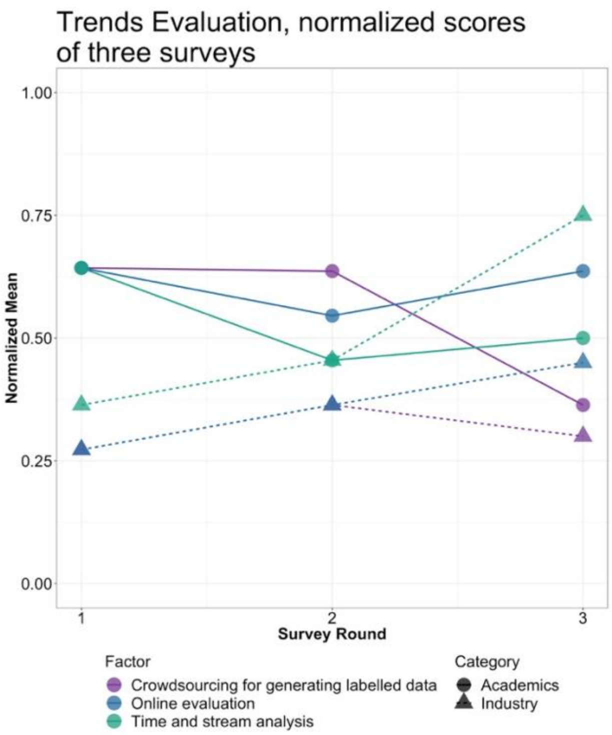
Trends crawling



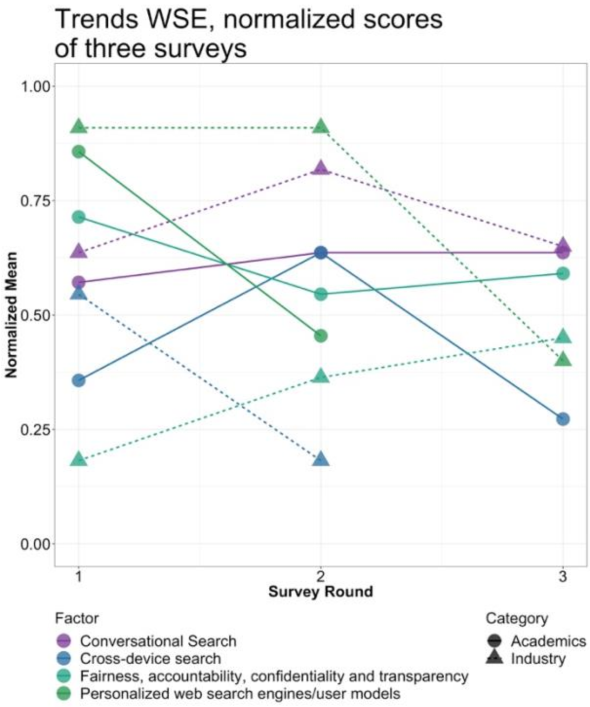
Trends ranking



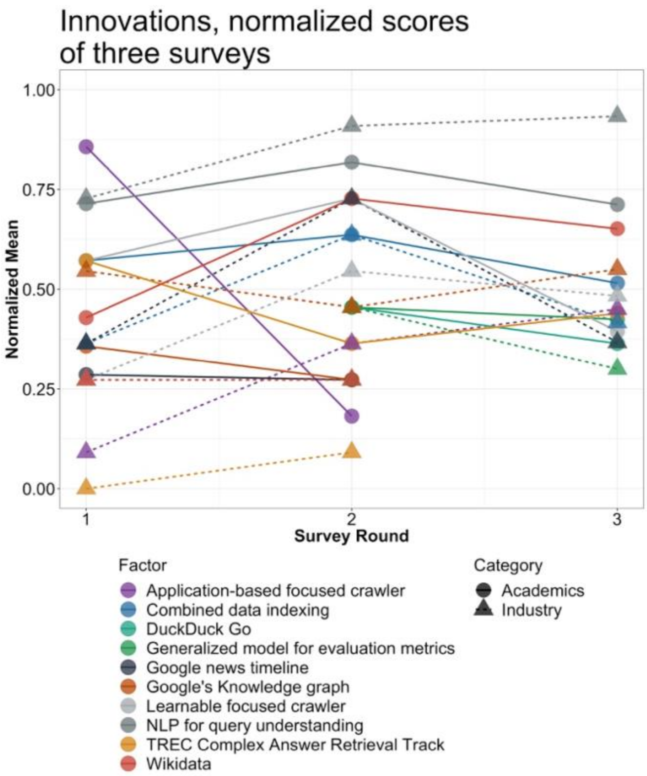
Trends evaluation



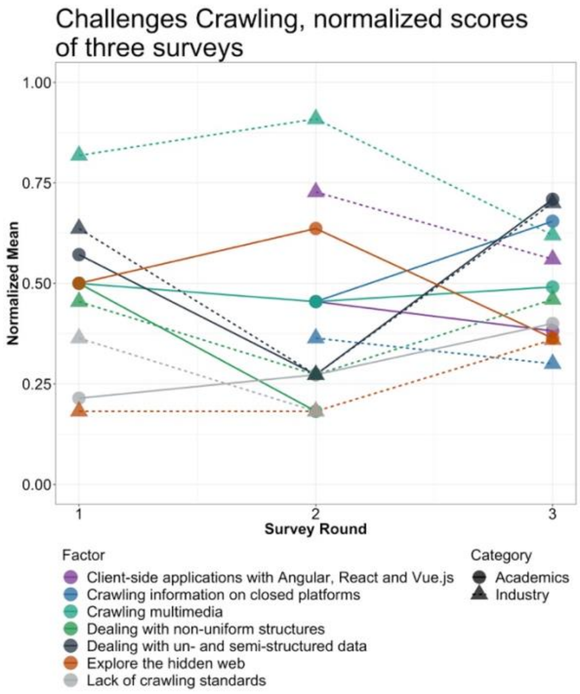
Trends web search engines



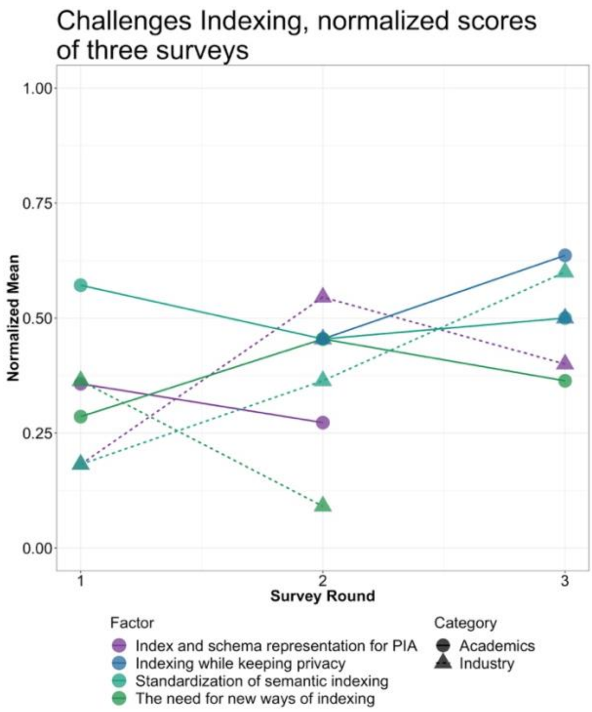
Innovations



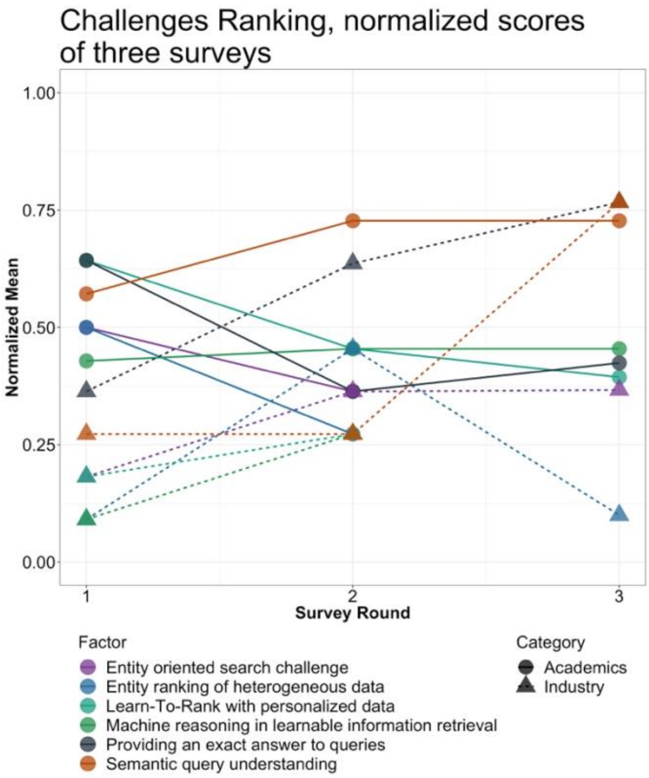
Challenges crawling



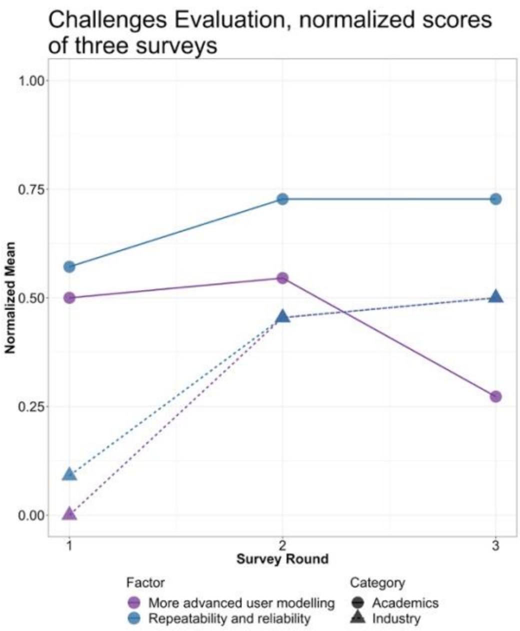
Challenges indexing



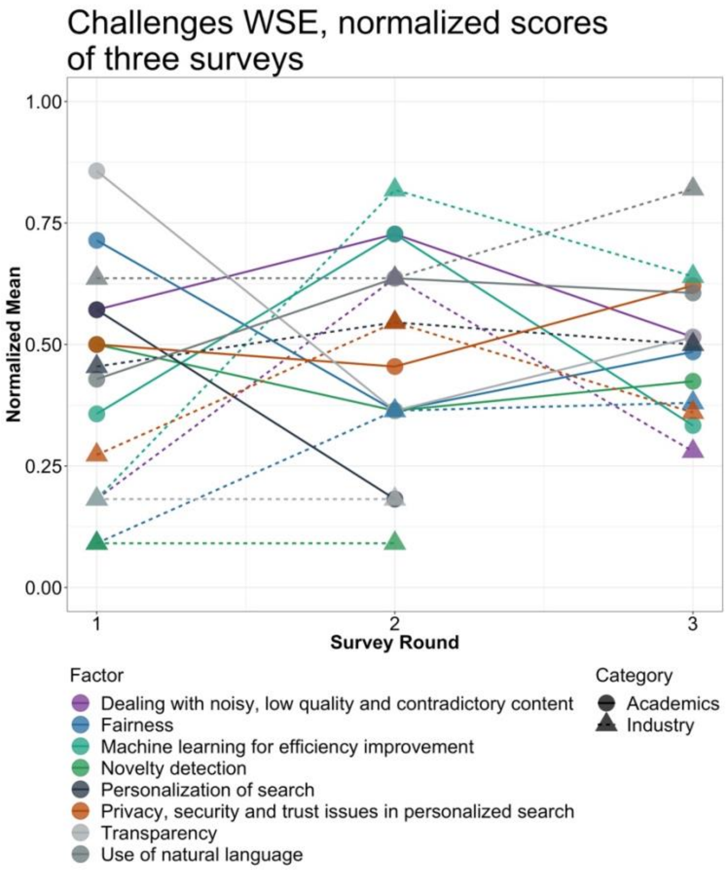
Challenges ranking



Challenges evaluation



Challenges web search engines


Challenges WSE, normalized scores of three surveys
1.00
0.75
0.50
0.25
0.00
Normalized Mean
1
2
3
Survey Round
Factor
Dealing with noisy, low quality and contradictory content
Fairness
Machine learning for efficiency improvement
Novelty detection
Personalization of search
Privacy, security and trust issues in personalized search
Transparency
Use of natural language
Category
Academics
Industry