UNIVERSITY OF GOTHENBURG

Effects of Confirmation Bias on Web Search Engine Results and a differentiation between Non-assumptive versus Assumptive Search Queries

Bachelor of Science Thesis in the Software Engineering and Management Programme

NEDA ASHRAFI-AMIRI
JOSEF AL-SADER

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
Göteborg, Sweden, June 2016

Effects of Confirmation Bias on Web Search Engine Results and a differentiation between Non-assumptive versus Assumptive Search Queries

Neda Ashrafi-Amiri
Josef Al-sader

Examiner: Agneta Nilsson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden June 2016

Effects of Confirmation Bias on Web Search Engine Results and a differentiation between Non-assumptive versus Assumptive Search Queries

Neda Ashrafi-Amiri
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusashne@student.gu.se

Josef Al-Sader
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusalsajo@student.gu.se

ABSTRACT

This study examines how confirmation bias influences the search results of large scale web search engines for software engineers. We show how the correctness of results can change depending on the formulation of search strings that the user inputs in the search box. The study is conducted as an experiment where some of the most popular software engineering topics are formulated as non-assumptive or assumptive queries and their web search results are then compared.

Keywords

Confirmation Bias, Information Retrieval(IR), Exploratory Search Query, Assumptive Search Query, Non-Assumptive Search Query, Query Suggestion, Web Search Engines(WSE), Search Engine Results Page(SERP)

1. INTRODUCTION

Searching for information has been an inevitable human activity since the beginning of human life. Before, most information was retrieved through synchronous (in-person or telephone conversations) and asynchronous (letters or books) interactions. Nowadays it is taking a more complicated twist whereby information is becoming more accessible through the World Wide Web and it is playing a bigger role in the decision making process for individuals.

Web search engines (WSE) are a type of information retrieval system and thus function in the same way. According to Manning et. al [1], "Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)".

Popular web search engines such as Google and Bing are serving millions of users and billions of search queries daily [2]. Thus, the interaction between the user and the WSE software is both important and sensitive. What the user inputs in the search box (i.e. the search query) directly influences the search engine results page (SERP) and hence the users post search belief.

Today, WSEs are moving towards providing more "effortless" search functions for users so that they are faced with IR systems that require the least amount of actions (clicks). We are faced with machines that will give us information with only a swipe gesture (Smart-phone touch) or a gaze (virtual reality glasses).

As futuristic and promising this all may sound, it is important to understand that information is now being proactively pushed towards us. Functions such as voice recognition have gone even further by basing their search algorithms on the background noise of our devices [2]. Despite the ethical issues involved in this and many other scenarios, we should be asking whether WSEs are responsible enough as the gatekeepers of our society and are they taking extra measures to ensure that the information we retrieve is the "correct" information?

In relation to user interaction with WSE's, cognitive bias can be observed in situations where users search for information to confirm their own hypothesis/beliefs. This specific kind of bias is known to psychologists as confirmation bias [3].

Let us examine the following queries: "study gluten intolerance" and "gluten intolerance cause and symptoms". The first query aims to study the matter whereas the second query already assumes that gluten intolerance is a medical condition. Each query input into Google search box will give different SERPs. The first query which is by our definition (definition discussed in section 2.5 of this paper) a non-assumptive query, will lead to more in depth knowledge about the subject and how researchers are now discovering that the Non-Celiac Gluten Sensitivity is not actually a real medical condition [4].

Query suggestions and related search models assist WSE users to find their inquired information quicker [5]. It is very often that users choose one of the suggested queries as their search query. This means that the IR community should take extra care constructing machine learning algorithms that formulate these suggestions.

Query formulation and confirmation bias are intertwined in more than one way. A user with the intention to confirm their belief about a certain topic has the tendency to formulate their query in a way that would trigger a predefined answer.

Previous research in the area of medical science has put much effort into developing strategies to minimize cognitive errors such as confirmation bias. These efforts range from having to consider alternatives, to gaining more in depth

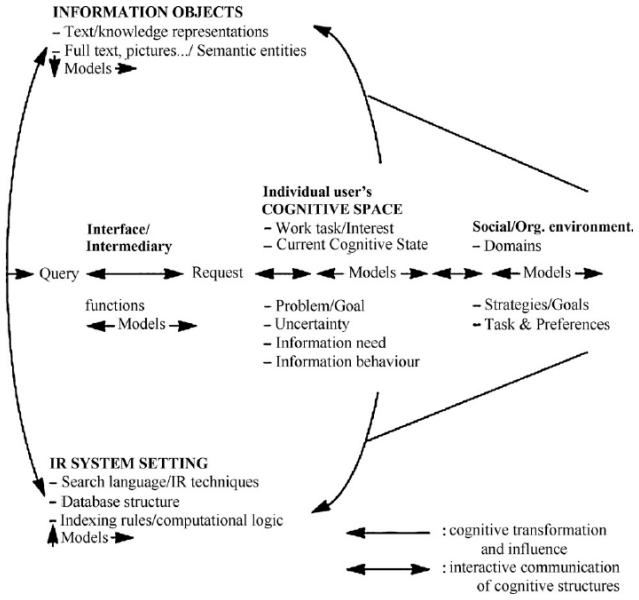

Figure 1: The cognitive model of information transfer (adapted from Ingwersen, 1992)

insight about matters under question [6].

In this paper we continue with some of the works by previous scholars on the effects of confirmation bias on SERPs. We specifically look at a group of individuals, software engineers and their domain specific search queries (DSSQ).

The reason that software engineers are studied is that we believe this group is very familiar with the use of WSEs on a daily basis and as a majority interact with different IR systems much more than many other groups. We gather data about the participants' prior and posterior beliefs in relation to a real-time software engineering search topic. This way we aim to understand how different types of queries lead to

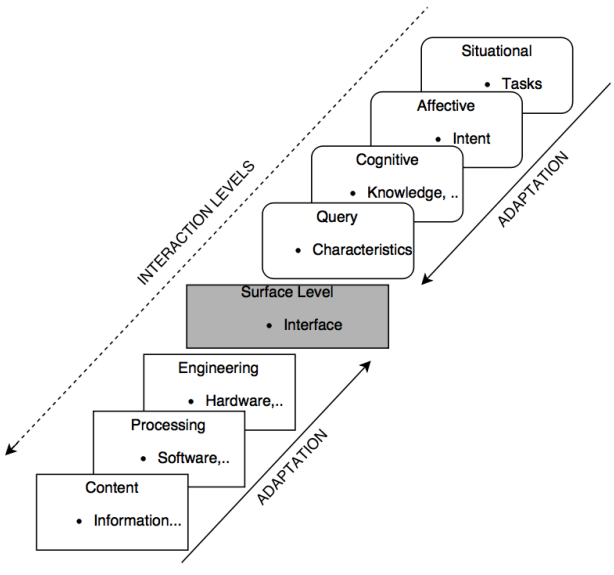

Figure 2: Stratified model of information retrieval interaction (adapted from Saracevic, 1996)

change in beliefs of individuals.

We later conduct a solution-oriented experiment in which a number of popular software engineering topics are formulated as search queries and categorized into two types of DSSQs (non-assumptive and assumptive queries). The answer found on the SERPs for each query is analyzed and compared against industry's (software engineering industry) beliefs of what the correct results are. This paper is aimed to acknowledge the effects of confirmation bias on SERPs and provide initial direction towards more de-biased IR systems.

2. FRAMEWORKS, RELATED WORK AND COGNITIVE BIAS

In this segment, we discuss four relevant areas to our topic: (1) WSEs and human behavior (2) WSE interaction techniques (3) cognitive bias in relation to search engines and (4) bias mitigation strategies. We also further iterate on query definitions and the scope which defines the framework of this study.

2.1 WSE and Human Behavior

There are numerous models and frameworks for information seeking and human interaction with search systems. Cognitive search models such as ones adapted by Ingwersen [7], are built on the idea that search interactions are complex in nature as they are dependent on the psychological functions of searchers (ref: Figure 1). Ingwersen acknowledges that an IR system is an integrated system that includes both system oriented (e.g. authors' texts, retrieval techniques and queries) and cognitive IR research (e.g. searcher's problem space, information need and interface design). The model further illustrates that there exist multiple interactive communication channels of cognitive structure.

In 1996, Saracevic [8] devised an illustration of an IR system by devising a stratified model of information interaction. The model (ref. Figure 2) differentiates between the searcher and the system. It also illustrates how both agents have same level of influence on a successful search. The arrows clearly show the direction of adaptation. For example it illustrates how the intent of the individual can influence other IR interaction strata such as the knowledge and query strata.

Both mentioned frameworks stress on mainly three concerning IR segments; the cognitive aspect, the engineering aspect, and the interactive aspect. In our study we aim to examine the cognitive (confirmation bias) and interactive aspects (search queries) in order to present a solution-oriented theory on how a change in the engineering of the WSEs can minimize the unwanted and inevitable effects of cognitive bias.

Web search engines are consistently pursuing new and improved search algorithms in order to increase the quality of their software and be able to compete in the WSE industry. The next-generation search interactions need to be more complex and handle sophisticated search behaviors such as complex queries and answer-oriented search intents [9]. Information is not only to be retrieved but learnt by future searchers [10].

2.2 Some of the current WSE interaction techniques

Interaction may be too generic of a word for this con-

text. To break it down and understand how WSEs interact with users we can look at a few examples. Auto-complete functions are those which the system offers in a real time or posterior-search (on SERP) to the user on the basis of popularity of the query or other methods such as prefix matching [11]. Shokouhi in 2013 [12], points out how auto-complete functions can be personalized on the basis of user's search history.

A study by Bozdag [13], tells us about large scale search engines such as Google and their attempt to cater user's needs via web-personalization. The user is in effect influencing the searching algorithm every time she commits a search.

These are only a few of the many ways (e.g. query suggestion, advertising clicks, hyperlinks, caption designs, SERP layout, etc.) that WSEs interact with their users. However, there also exists certain risks involved both for the users and the WSEs' reputation in the long run. Epstein and Robertson, 2015 [14], explain how the searchers' beliefs are prone to change as a result of a sequence of search sessions. If those beliefs were to be changed for wrong reasons, or skewed towards inaccurate understanding of topics, then the search engines are failing.

2.3 Cognitive Bias and Search Engines

The underlying cognitive factors that people may carry also have an influence on how interaction is made with search engines. In 2013, Ryan White [15] conducted a study about biases in web search. The study primarily investigates whether people's beliefs in yes-versus-no questions change as a result of searching. He also discussed the extent to which WSE results are biased to favor certain outcomes. The results from this study also showed that people tended to choose search links which were more favorable to their perceived notion of a subject, despite it being wrong.

In 2014, Schweiger et al. [16] conducted a study to investigate the public's confirmation bias while searching online. The topic of search was whether psychotherapy was more effective than pharmacotherapy for treating depression. The results concluded that despite the latest evidence of the two practices being as effective, public tended to believe psychotherapy was more effective and hence projected that belief in their after-search beliefs.

More evidence of confirmation bias playing a role in web search is reflected in a meta-analysis conducted by Hart et al. [17]. In their study, they investigated whether the underlying factors for selective exposure to information is influenced by defensive or accuracy motives. They concluded that those with some prior-knowledge about a given subject tended to lean towards congenial information (supportive of participant's pre-existing beliefs), while the absence of prior-knowledge showed a stronger leniency towards uncongenial information (driven by accuracy reasons in search for correct solutions).

2.4 Bias Mitigation Strategies

According to Schwind et al. [18] the web contains a vast amount of unexploited material because of searcher's inclination for preference-consistent information, that is to say, selectively choosing material on the internet that satisfies their biased view of a given subject. The research team conducted two experiments where they showed that preference-inconsistent query-recommendations could mitigate confirmation bias and as such stimulate divergent thought.

Lewandowsky et al. [19] mention that psychological ways can be used to retract misinformation and even render truth instead of false beliefs for individuals. In their study, the authors devised a set of methods to erase misinformation from individuals by promoting factual information. Some of these methods employed by the researchers involved invoking an alternative account, repeated exposure of that account, emphasizing on facts and pre-exposing an individual with a warning whenever misinformation was to be presented.

Another bias mitigation strategy is not only to change what kind of information is being presented to an individual, but rather how that information is being presented in a way that would reduce bias. Hernandez et al. [20] emphasizes that disrupting the procedure of processing information (disfluency) can reduce confirmation bias by changing how information is presented. The results of their study argues for a more critical response of analysis from recipients of information if they experience difficulty in processing the information of which they are being exposed to.

2.5 Definition of Query Types

Bates [21] identifies four levels of search activities, whereas in this study we focus on the first two. The levels differ in term of the extent of user interaction with the search system. For atomic search behaviors, users tend to execute a one time query that Bates identifies as a "*move*". A set of moves (queries) aimed to complete a search task is described as a search "*tactic*". At this level of search interaction, the user is more goal-oriented and tends to search an iteration of queries. Throughout this paper, this definition is used to easier explain concepts involving these search-patterns.

In 2006, Marcionini [10] pointed out that query-moves may take different forms. He divided them into three categories (as seen in figure 3); look-up, learning and investigating. While look-up moves are good for narrowed down and precise information searches such as fact retrieval, he described the combination of learning and investigative qualities as exploratory moves, where the aim of information retrieval is to attain a higher-level and richer acquisition of knowledge or discovery.

As a result of our exploratory interview, which is described in section 4.1, we have defined two types of queries; assumptive and non-assumptive. The definition of these are as follows:

Non-assumptive queries: Shares both learning and investigatory characteristics. They overlap only within the exploratory search domain shown in figure 3. By this defi-

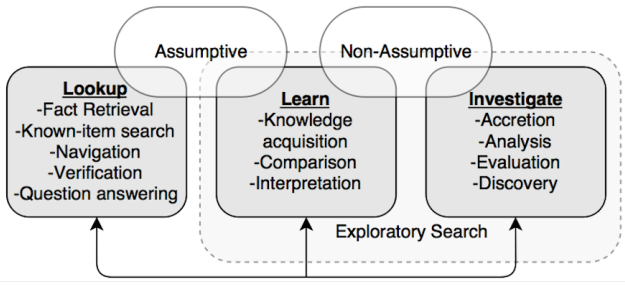

Figure 3: Different categories of query-search (adapted from Marchionini, 2006)

Table 1: Sorting categories for analysis of interview results

QueryTactic	*QueryMove*	*Non − assumptive*	*Assumptive*	*Answer*	*Belief*
1			X	Found	Changed
	Audio-book Android	X			
	import and play audio-book Android		X	X	

nition, an example of a query which is non-assumptive would be *What does "use strict" do in JavaScript, and what is the reasoning behind it?*, because it seeks to learn more about the concept of use strict applied in a JavaScript environment, and analyzing the rationale behind it.

Assumptive queries: The connotation of Marcionini's model in relation to what we call assumptive moves imply that assumptive moves encompass more look-up querying traits. They tend to be of verifying nature, returning more narrowed down information that results in some part of the domain-knowledge being found (albeit not giving a conclusive image of the whole). By this definition, an example of a query which is assumptive is: "What's so bad about Template Haskell?", which will result in a specified acquisition of domain-knowledge, but is also biased in its formulation because it assumes that template Haskell is bad.

In a latter section of this paper (Ref: Section 5.1) we will investigate these definitions further.

2.6 Domain Specific Knowledge and IR

Marcionini, 1988[22], mentions that there are other major factors that play a role in a successful search such as domain knowledge of the searcher. Domain knowledge has a direct influence on the information seeking behavior of the searcher. If the domain knowledge of a group of searchers differ by a great extent then making comparisons between their search queries will be an unjust one.

Previous mentioned studies such as ones conducted by White[15], and Schweiger et al. [16] in relation to bias in WSEs have not sampled participants within the same range of domain knowledge. In our study, we aim to control this variable to some extent, namely the "searchers' domain knowledge" by only investigating software engineers and their domain specific search queries (DSSQs). We understand that the knowledge of domain varies between different software engineers, however, they will all have at least some fundamental knowledge about most common software engineering topics.

3. PURPOSE

Our main goal for this study is to raise awareness to the IR community on how confirmation bias is an inevitable human behavior and needs to be taken care of through algorithmic solutions in the IR system settings. WSEs are playing a bigger role as information leaders of our societies today and need to take steps in order to show accurate information to searchers instead of confirming users wrong beliefs. It may be true that at times WSEs promote confirming users' belief despite the accuracy of the retrieved information in order to keep the users happier.

The purpose of this experiment is to assess the effect of using non-assumptive queries as opposed to assumptive ones on web-search results' correctness. But before we experiment we need to understand how software engineers' beliefs change as a result of searching on WSE's. We want to also look at how they form different queries in natural settings and on the basis of that we will formulate and categorize the experimental queries.

Research Question 1:Does Confirmation bias and the use of non-assumptive and assumptive queries influence web search engine results for software engineers?

Research Question 2: Is there a difference of accuracy of results when using non-assumptive vs. assumptive queries on web search engines?

Null Hypothesis: The use of non-assumptive queries leads to the same web search results as the use of assumptive queries.

$$H_0 : \mu 1 = \mu 2$$

Alternative Hypothesis: There is a difference of accuracy of results when using non-assumptive vs. assumptive queries on web search engines.

$$H_1 : \mu 1 \neq \mu 2$$

4. RESEARCH STRATEGY

Our research strategy is built on two consecutive parts. An exploratory interview will be devised to get a better insight on the behavioral aspects of how software engineers use web search engines. This will then lay the foundation upon which the actual experiment is conducted on. In this section, we discuss the methodology behind this approach.

4.1 Exploratory Interview

The first phase is a qualitative exploration of how software engineers format their queries on Google and a recording of their prior and posterior beliefs about their searched-for topics. We execute this phase by interviewing 10 software engineers. Findings from this qualitative phase are then used to answer the first research question on how confirmation bias and the use of non-assumptive queries influence web search results and serve as a guideline for the second phase of this study.

The interview is conducted in a semi-structured manner, meaning that the questions are constructed in a way to cover certain areas of interest and also tend to be open-ended. This allows the participants to open-up and be more unique. There is a scheduled time and a specific goal for the interview. The requirements for this interview is that each participant brings their personal laptop or any other device that they use to search for information on.

There is an observational part in the interview that aims to explore more about real-life search experiences of the participants. This part is of course carried out upon the participants' full consent. The interviewees present their past week's (or more) search history on Google. They are given

time to clear any browsing data that they do not wish to share with us. If the participants do not use Google at all then we omit this part of the interview and focus on their given examples rather than observing proof.

4.1.1 Step-by-Step guide to the interview

Here are the steps we have taken for the interview:

1. Prepare questions for the interview.
2. Select the participants on the basis of our selection criteria
3. Set a time for the interview.
4. Send an email to the participants informing about the confidentiality of their data (following the Code of Practice for Research Ethics Concerning Human Participants (Non-NHS) [23]). The email also includes the requirement for them to bring in their search devices (e.g. laptop) and the time that they need to be present at their specified location.
5. On the day of the interview we will once again inform the participants of how we will use their data and record their voice once they consent for it.
6. The interview takes place. There are four main areas we wish to get further insight in: (1) participants' domain specific query habits (the domain being software engineering) (2) change of belief in after-search and (3) closeness of answers found in relation to their prior perception. To acquire relevant data that covers these three areas, the following questions are going to be used as a part of a semi-structured framework:
   (a) Do you recall a recent software engineering search topic that you used a web search engine to find the answer to?
   (b) Do you have the search query or queries you have used to find the answer for this question saved on your browser's history? If yes, can you tell us what the query was?
   (c) What was your prior belief on the answer to that question?
   (d) What do you think the answer is now?
   (e) Would you say that what you have found out was close to what you had in mind prior to your search?
7. The data will be transcribed and analyzed.
8. The participants will receive an email mentioning our gratitude for their time. This email will include an attachment of the case study.

4.1.2 Sampling and Data methodology

Ten participants were sampled for the exploratory interview. The primary aim was to sample software engineers from both different levels of academia as well as the industry. However, mostly software engineering students showed interest, which resulted in all ten test subjects having a software engineering student background.

The sampling method applied can be described as convenience sampling taken from a pool of individuals studying at the department of computer science and engineering, both from Gothenburg University and Chalmers. The average time for each interview was between 20 - 30 minutes, depending on how fast a test subject could recall/find relevant software engineering-related search history. These interviews were then recorded and transcribed into digital documentation. From these documents, each subject's query set of query moves was organized into a table where, based on the information provided from the interviewees, prior and posterior belief was identified as well as the query-type (being either assumptive or non-assumptive).

4.1.3 Data Analysis

The data extracted from the exploratory interviews is qualitative, offering an in-depth insight into how software engineers format their queries on Google and comparing their prior and posterior beliefs after they have searched for different software engineering topics. Consequently, since using this approach will encompass the use of open-ended questions, the resulting data from this exercise is anticipated to be qualitative and is thus transcribed and interpreted within a deductive framework [24].

The extracted data will be analyzed to see if there are any changes in posterior-belief of a given question about a software engineering topic. The observational part of the interview which involves examining the Google search history of our interview subjects is intended to strengthen the answer to this question by giving us an insight on the search behavior of interviewees.

By transcribing the collected data and sorting it into categories as seen in table 1, we will get a better comprehension on the existence of bias within their search queries and how this bias then effects the interviewee's perceived accuracy of results. Each interviewee will have their data organized in a separate table. Queries are put into groups as some were only iterations aiming to answer a single question or explore the same topic. If the answer is marked as found by the participant, then there will be an **X** sign next to the specific query that led to the found answer.

4.1.4 Validity Threats

One of the risks posed to validity in regards to the exploratory interview in our study is posing questions which are not open-ended, and as such driving the interview results towards a certain level of bias [24]. Kitchenham et al.[25] describes how data collection procedures and analysis can be prone to open interpretation and research bias when conducting such an interview. This is reflected not in the least to how we are trying to prove a hypothesis in this part of the study which can drive the interview towards a certain level of bias, creating a confirmatory set of data.

The use of different web search engines was seen at first as a validity threat, however none of the participants used search engines other than Google. This helped us to be able to compare the results easier as this independent variable remained unchanged throughout the course of the interviews making it a superficial controlled variable.

Choice of language for a given interview may also prove to have an affect on the quality of the interview. If for instance English is not the interviewee's mother tongue, this might deter the quality of information collected.

The subjective nature of classifying whether a query-move is non-assumptive or assumptive can hold a threat towards

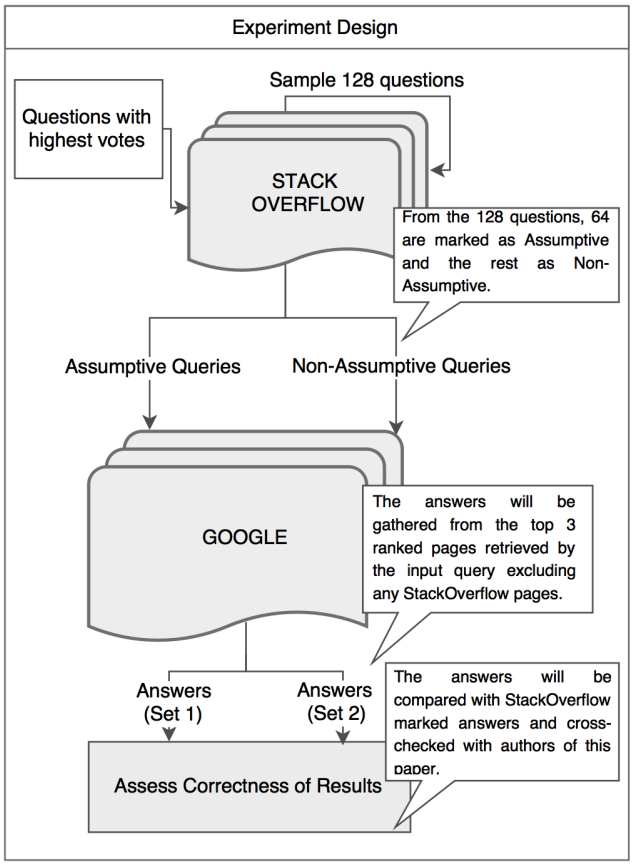

Figure 4: The design strategy for the experiment phase

the validity of the experiment. It especially runs a high risk if only one individual is to distinguish what is one or the other. To lower the risk, the query-classification is done independently by each research-member, upon which the classification is done once more together with all research-members. The aim of this step is to come to an agreement of the classification during discussion of the matter. If disagreement upon certain queries ensues, an independent third-party researcher (who has comprehension of the study material) will have the final say.

Furthermore, human behavior can be influenced by peer-pressure when doing an interview, therefore the aim of the data collection part of this study is to interview people individually rather than in groups to ensure a higher level of quality of the data extracted [26].

4.2 Experiment

The second phase, (i.e the experiment) is of quantitative nature and it is aimed to test the findability when using WSEs, in this case Google in relation to the the type of query the user inputs (non-assumptive versus assumptive). By findability we are referring to the extent that the top-ranked pages include correct answers to the questions presented as queries.

The strategy for this phase involves firstly finding 128 software engineering questions asked by the members of this community. We will derive to the 128 questions by referring to the famous website namely *Stack Overflow* which is a question and answer site for professional and enthusiast programmers. The sampling of these questions will be further analyzed in section 4.2.2 of this report.

The 128 questions comprise of 64 non-assumptive and 64 assumptive questions. The process of addressing whether a question is non-assumptive or assumptive is of subjective nature, however a filtration process on the basis of the definitions of the two forms of queries will be carried out before the categorization is final.

Each question on the Stack Overflow website has an accepted answer which we will consider as the *correct* answer. Each of the selected questions will be posed as a query and input into Google's search box individually. The top three results for each question will be analyzed simultaneously excluding the Stack Overflow link. If the answer is present in the first page, then we do not visit the next page and if no answer is found on any of the three top pages, then the results will display "no answer found". Figure 4 is a simplified overview of the explained strategy.

The independent variables for this experiment are the 128 formatted queries and the dependent variables are the indexed pages by Google search engine.

A Mann-Whitney U-test examination will be conducted to statistically examine the two sets of queries. The test examines whether two samples are different or not. The below symbols represent our hypothesis and the expected outcomes:

$$H_0 : \mu 1 = \mu 2$$

$$H_1 : \mu 1 \neq \mu 2$$

4.2.1 Sampling Method

The population for this experiment is comprised of software engineering questions. Due to the unlimited nature of this population, we have used the calculations below to derive to a sample size number:

- Statistical power: 0.8
- Anticipated effect size: 0.3
- Significance level: 0.05
- Minimum total sample size (two-tailed hypothesis): 128
- Minimum sample size per group (two-tailed hypothesis): 64

The above formulas are involved in the calculation of a priori sample size values for a Mann-Whitney U-test. In this experiment as mentioned before we are dealing with two sets (groups) of queries. The above calculations infer that for this experiment, 64 non-assumptive and 64 assumptive queries will be formulated to be input into Google Search Engine. This size was chosen to increase the statistical power of our experiment. By choosing a statistical power of 0.8 we ensure that there is at least an 80 percent chance of rejecting a false null hypothesis and as such avoid committing a type II error [27], which is the failure to reject a false null hypothesis. The sampling method used is a simple random sampling technique, however, the choice of using Stack-overflow as a source of population proposes certain validity threats discussed further in the validity threats section.

Table 2: An Assumptive Query Tactic for Subject 9

QueryTactic	*QueryMove*	*Non − assumptive*	*Assumptive*	*Answer*	*Belief*
1			X	Found	Not Changed
	finite-state transducer closure intersection		X		
	finite-state transducer no insertions deletions closure intersection		X	X	
	regular relations closure properties	X			
	regular relations intersection	X			

Table 3: A non-assumptive Query Tactic for Subject 10

QueryTactic	*QueryMove*	*Non − assumptive*	*Assumptive*	*Answer*	*Belief*
1		X		Found	Changed
	GCM android tutorials	X			
	GCM android java	X			
	GCM android studio	X			

4.2.2 Data Analysis

This segment of the research deals with quantitative data which consists of accuracy measurement of search results. The sampling stage produces data which will be sorted into two categories; non-assumptive or assumptive. The criteria for categorizing them is in accordance with segment 2.2.

After this, the data will be further refined when the manual Google search comparison takes fold, leaving a measure of accuracy on the data for both categories of queries. An assessment of accuracy is made to each sampled move by querying it on Google. The important variables to be considered in this regard is which SERP-link the answer is found on and if the source of this answer is independent from stack-overflow. The ordinal scale which is used is of a ranking nature, where answers found on the first SERP-link is given a '1', '2' for second and '3' for third. If the answer can not be found within the first three SERP-links, then the accuracy value will be set to 0 (meaning no answer found).

This means that the dependent variable consists of ordinal data, given that we are dealing with two independent groups which is not anticipated to be normally distributed (because of the small variance of the ordinal-scale being 0-3) [28]. This is also why we have chosen to instigate a non-parametric equivalent of the t-test, i.e. Mann-Whitney U test. It is the accumulated accuracy level of all non-assumptive, respectively, assumptive queries which will be compared against each other through a Mann-Whitney U-test in order to examine the difference of distribution (i.e. the difference of the mean ranks between two independent groups) within the data-set.

The procedure of a Mann-Whitney U test is to compare two sample groups by calculating the rank of each data-value among all data-values from both groups, and then calculating the rank-sum value for each group, which we denote as U1 and U2 respectively. The computed statistic is the U-value, which is the minimum of U1 and U2 (U = min(U1, U2)), and is the basis of which measurements such as significance level and effect size is derived from. This can be done by implementing the mathematical theory of the Mann-Whitney U-test [29] by hand, however for simplicity, the test will be done using the statistical analysis tool R which has inbuilt algorithms for the purpose of running a Mann-Whitney U-test. The effect size will also be calculated using these algorithms. In theory, the effect size measure is found by dividing the Z-score (standard deviation from the mean in our data) with the square-root of the total number of samples involved in the experiment, and then taking the absolute value of this result to acquire the effect size (which we denote as r).

4.2.3 Validity Threats

Here, like the analysis of the exploratory interview, the subjective nature of classifying whether a query-move is non-assumptive or assumptive can hold a threat towards the validity of the experiment. The same approach of identifying query-types for the exploratory interview section will be applied here, whereas if disagreements ensue after discussion of a given query, a qualified independent third-party researcher will have the final say.

Choosing stack-overflow as the pivotal factor for the sampling methodology was done to reduce bias by a great extent. A large and reliable community of software engineers is what drove this decision. However, stack-overflow is still a forum regulated by the people who make up that forum, which can result in some bias exposure. Also, misinterpreting the answers given on stack-overflow when undergoing the accuracy assessment phase can undermine the validity of the experiment. Therefore, measures are taken here to minimize the risk by having both research members discussing and comparing the answer material.

5. RESEARCH FINDINGS

In this section, we present the findings from our exploratory interview and query-experiment. This includes excerpts from the transcribed interview results, depicting practical means of identifying non-assumptive and assumptive queries. We

also present the experiment results. These results include the comparison of percentages in relation to answers found, both for non-assumptive and assumptive queries, as well as a Mann-Whitney U test conducted on the experimental data.

5.1 Exploratory Interview Results

While conducting our exploratory interview, we noticed a pattern. Test subjects tended to often search in an iterative manner, querying a variety of search strings in order to find their answer. Another pattern that was noticeable throughout our analysis of the interview was that two different kind of moves were prevalent; those which sought to dig deeper within the subject-domain, and those which tried to narrow down the results to more specific information. Examples of non-assumptive query-moves were characterized by searching for more general, large topics such as 'Whatsapp architecture' but also when comparing entity A with entity B as for example 'Magento vs. WooCommerce'.

On the other hand, assumptive queries were expressed when subjects wanted domain-specific information. Such queries could involve look-up tasks like specific fact retrieval ('Android image view stretch height to key-ratio'), or error messages such as 'Error 43554', but also queries of more conspicuous confirmatory nature such as 'Why use strong instead of bold?'.

Table 2 depicts an example of how one subject's query tactic was classified during our analysis of the interview. In this particular case, our analysis found that the second query move of the query tactic was assumptive while also being the query move that gave the subject's desired answer. The subject had some preconceived belief about the properties of finite-state transducer, which was confirmed by the answer the subject found, therefore the core-belief was not changed in this particular case.

Table 3 illustrates a case where a query tactic did not have any preconceived assumption, and was thus classified as non-assumptive. In this particular instance, the interview subject in question had no pre-conceived notion about how GCM (Google Cloud Messaging) functions within an android environment. After having found the solution to the problem, the subject's notion went from no belief to having a belief, prompting a belief change.

A noteworthy pattern from the interview data that was collected is that query sets which are non-assumptive tend to change the preconceived notion about the answer that is found. In contrast to this, when an answer is found through an assumptive query move, the preconceived belief has a tendency to not change. Figure 5 depicts a summary of the differences between non-assumptive query tactics and assumptive query tactics. It was found that roughly 76.5 percent of assumptive querying led to an answer being found, while 23.5 percent of the total results of assumptive querying (including cases where answers were not found), sparked a change of belief. In contrast to this, subjects always found their answers when adopting a non-assumptive tactical approach to querying, whereas 91.7 percent of the results stimulated change of belief.

5.2 Experiment Results

The pretext for conducting this experiment was to: (1) analyze the accuracy of non-assumptive query-moves (2) analyze the accuracy of assumptive query-moves and (3) comparing the difference of accuracy of results between non-assumptive and assumptive query-moves.

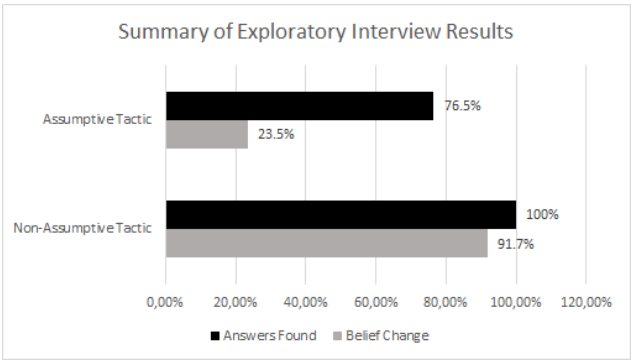

Figure 5: Interview Analysis: Change of belief in relation to answers found

Non-assumptive query-moves showed an approximate of 93 percent level of accuracy, where 60 of 64 sampled queries gave an answer independent of stack-overflow. 55 of these answers were found on the first SERP-link, while only two were found on the second SERP-link. The queries often gave many independent but complex answers that involved spending more time when analyzing them, to acquire a conclusive assessment. Figure 6 depicts the difference of answers found between the 64 non-assumptive queries in relation to SERP-links.

Assumptive query-moves had an accuracy level of approximately 66 percent, with 42 out of 64 answers being found. Out of these 42, most answers were found on the first SERP-link (39 answers to be precise), while three answers were found on the second SERP-link. Answers which were not found were often due to no other source than stack-overflow (or forums pointing to stack overflow) having the answer to a given query. So in essence, the nature of how assumptive queries are formulated meant mostly that if no other forum independent from stack-overflow had answered the question, an answer would not exist on the first three SERP-links. When an answer was found, it was often because the queried question had a popular trend of being asked. Query-moves which sought to extract perceived fact-retrieval such as "what is reflection and why is it useful?" tended to result in an answer with shared consensus in rela-

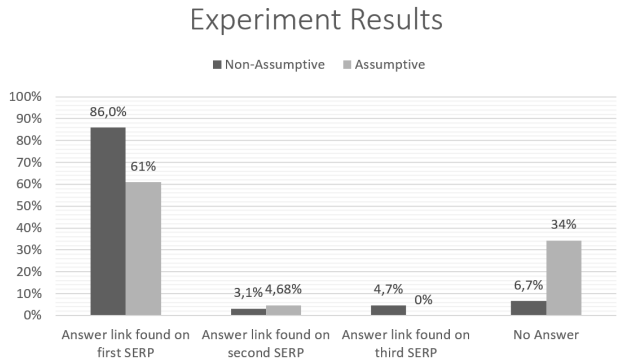

Figure 6: Experiment Results: Assumptive and Non-assumptive answers distributed over SERPs

Table 4: Snippet from the experiment results analysis and sorting table

$QueryNr.$	$StackOverflowQuery$	$Non-assumptive$	$Assumptive$	SERP's $nth\ link for\ the Answer Found$
1	What does "use strict" do in JavaScript, and what is the reasoning behind it?	X		1
2	Why is printing B dramatically slower than printing #?		X	no answer
3	How do JavaScript closures work	X		1
4	Why is SELECT * considered harmful?		X	no answer
5	Read/convert an InputStream to a String		X	1

tion to the stack-overflow software community, thus generating an accurate answer.

Figure 6 shows the difference of answers found between the 64 assumptive queries. Interestingly enough, no given answers were found on any third SERP when measuring the accuracy of assumptive query-moves.

Comparing the difference of accuracy between the answers provided from the non-assumptive and assumptive query moves from our specific sample of data show that the answers of non-assumptive moves had a higher overall percentage of being accurate, being approximately 27 percent more accurate in comparison to the results of assumptive queries. After devising a Mann-Whitney U test to evaluate the difference of accuracy between the two sample groups, a significant effect of group was found (U = 1452.5, p = 0.0002385, Z = -3.677, r (effect size) = 0.33).

Based on a probability level of 0.05, there is a statistically significant difference between the non-assumptive and assumptive sample groups ($p < 0.05$), implying that the chances of the effect observed in our sample-data being supportive of the null-hypothesis is less than five percent. Furthermore, an effect size of 0.33 show that there exists a medium level (in accordance with cohen's definition [27]) of correlation between the two samples in relation to the strength of the observed phenomena [30]. Table 4 illustrates a snippet from the SERP data of which the Mann-Whitney U test was applied on.

A query-move which gave an independent answer is categorized as found, with the accompanied SERP-link number (between 1-3) in a column to the right of it. In contrast to this, a query-move which was not found was given the SERP-link number 0 and was as such categorized as not found.

The overall data suggests that we can reject the null-hypothesis stating equality between the assumptive and non-assumptive sample groups, and that a difference of accuracy exists in favor of non-assumptive queries.

5.3 Discussion

Evaluating our research findings further in relation to RQ1, our exploratory interview suggests that confirmation bias has an influence on software engineers search habits. This influence is for the most part focused on assumptive queries, prompting a decreased level of belief change in comparison to non-assumptive queries. A similar pattern was found during our experiment, where a difference of result accuracy existed between non-assumptive and assumptive queries when conducting search on Google. After conducting a Mann-Whitney U test, a significance value of less than 0.05 was found, indicating that we can indeed reject the hypothesis stating that there is no difference between the two sample groups of assumptive and non-assumptive queries, consequently confirming RQ2 stating inequality in accuracy of results between assumptive and non-assumptive queries.

Although research has been conducted on the effects of confirmation bias during web-search, our research shows that confirmation bias play a role when querying assumptive and non-assumptive moves. As far as our knowledge goes,we are the first researchers who have explored this specific scope of investigation.

We think that one of the pivotal reasons for why non-assumptive queries lead to a better accuracy of results lay in known previous knowledge of the topic to be searched. Schwind et al. identified this instance in their study of subject's not having a preconceived notion about a search-topic, showing that this often led to a change of belief. Contrary to this, having a preconceived notion about a given subject often left the belief unaltered.

A closer examination of our exploratory interview-results showed often that non-assumptive query-moves made test subjects change belief 91.7 percent of the times, in comparison to assumptive moves which left beliefs altered at only 35.5 percent. However in total there were 12 non-assumptive queries, as opposed to 34 assumptive ones, which suggests that these results might have looked different if there were an equal amount of both types being compared. What acts as an interesting basis for why change of belief was more prevalent with non-assumptive moves is that interviewees often answered that they had no previous domain knowledge about the topic of which they were searching for when conducting this kind of search.

White 2013 [15], has used respondents' recollection of their recent committed search queries as basis for research. In his study, the *ground truth* answers were objectively judged by two physicians. We did not rely on respondents recollection but on what was found on their personal search history. This way the variable of memory has been removed and with it the risk of *faulty* memory. Moreover, instead of relying on the judgment of two experts on what the ground truth was for the experiment results, we relied on a community of experts (i.e. The stack-overflow community). Furthermore, disagreement over identifying queries as either assumptive or non-assumptive throughout this study was always resolved

through discussion within our research-team, and as such we never had to call in an independent third-party researcher to solve an issue of this kind.

5.4 Conclusion

In this study, we have come to the conclusion that non-assumptive queries have a better chance of increasing the findability (accuracy) of search results in comparison to assumptive queries for software engineers. Our means of investigation has been learning more about software engineers' search habits through an exploratory interview and then conducting an experiment where we compared the accuracy of assumptive and non-assumptive query-moves in relation to answers found.

Based on our overall findings, we suggest further studies involve devising a practical solution which promotes the use of non-assumptive queries. This could be done through employing a query-suggestion model [5], assisting users conducting web-search by detecting assumptive queries and suggesting a non-assumptive alternative to this. We believe that this will make WSE more flexible, still maintaining a quick way for simple-level search such as fact retrieval, but also more advanced higher-level search which involves learning and investigative characteristics.

We also suggest to expand the sphere of DSSQ (domain specific search queries), to get a further insight on the differences between assumptive and non-assumptive queries in other areas and by people with different domain specific knowledge.

6. ACKNOWLEDGMENTS

We would like to thank Michal Palka and Inari Listenmaa for mentoring us throughout the course of this research. We would also like to thank all of those students who have given us their time by participating in our interviews.

7. REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. *Cambridge: Cambridge university press*, 1, 2008.

[2] R. W. White. *Interactions with Search Systems*. Cambridge University Press, NY, USA, 2016.

[3] S. Plous. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company, 1993.

[4] J. R. Biesiekierski, S. L. Peters, E. D. Newnham, O. Rosella, J. G. Muir, and P. R. Gibson. No effects of gluten in patients with self-reported non-celiac gluten sensitivity after dietary reduction of fermentable, poorly absorbed, short-chain carbohydrates. *Gastroenterology*, 145(2), 2013.

[5] O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to web search results. *Computer Networks*, 31(11), 1999.

[6] P. Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8), 2003.

[7] P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of documentation*, 52(1), 1996.

[8] T. Saracevic. Relevance reconsidered. in information science: Integration in perspectives. *In Proceedings of the Second Conference on Conceptions of Library and Information Science*, 1(2), October 1996.

[9] M. L. Wilson, B. Kules, and B. Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1), 2010.

[10] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4), 2006.

[11] S. Chaudhuri and R. Kaushik. Extending autocompletion to tolerate errors. *In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, June 2009.

[12] M. Shokouhi. Learning to personalize query auto-completion. *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, July 2013.

[13] E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3), 2013.

[14] R. Epstein and R. E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), 2015.

[15] R. W. White. Beliefs and biases in web search. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 12(2):3–12, 2013.

[16] S. Schweiger, A. Oeberst, and U. Cress. Confirmation bias in web-based search: a randomized online study on the effects of expert information and social tags on information search and evaluation. *Journal of medical Internet research.*, 16(1), 2014.

[17] Alice H. Eagly Inge Brechan Matthew J. Lindeberg William Hart, Dolores Albarracín and Lisa Merrill. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555–588, 2009.

[18] C. Schwind, J. Buder, U. Cress, and F. W. Hesse. Preference-inconsistent recommendations: An effective approach for reducing confirmation bias and stimulating divergent thinking? *Computers and Education*, 58(15):787–796, 2012.

[19] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 2012.

[20] Jesse Lee Preston Ivan Hernandez. Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology*, 49(1):178–182, 2013.

[21] Marcia J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[22] G. Marchionini and B. Shneiderman. Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1), 1988.

[23] University of Leicester. Code of practice for research ethics @ONLINE, May 2015.

[24] P. Runeson and M. Höst. Guidelines for conducting

and reporting case study research in software engineering. *Empirical software engineering*, 14(2), 2009.

[25] B. A. Kitchenham, L. Pickard, and S. L. Pfleeger. Case studies for method and tool evaluation. *IEEE software*, 12(4), 1995.

[26] N. Juristo and A. M. Moreno. Basics of software engineering experimentation. *Springer Science and Business Media*, 2013.

[27] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[28] Evie McCrum-Gardner. Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, 46(1):38–41, 2008.

[29] H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[30] Gail M. Sullivan and Richard Feinn. Using effect size–or why the p value is not enough. *Journal of Graduate Medical Education. 2012*, 4(3):279–282, 2012.