
# USE OF NEGATION IN SEARCH

THESIS

Kristen M Lancaster, Contractor, USA

AFIT/GCO/ENV/10-J01

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

## AIR FORCE INSTITUTE OF TECHNOLOGY

**Wright-Patterson Air Force Base, Ohio**

This research was supported in part by an appointment to the Postgraduate Research Participation Program at the United States Army Aeromedical Research Laboratory (USAARL) administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the United States Department of Energy (DOE) and United States Army Medical Research and Materiel Command (USAMRMC).

AFIT/GCO/ENV/10-J01

# USE OF NEGATION IN SEARCH

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cyber Operations

Kristen M Lancaster

Contractor, USA

June 2010

AFIT/GCO/ENV/10-J01

# USE OF NEGATION IN SEARCH

Kristen M Lancaster

Contractor, USA

Approved:

| | |
|---|---|
| //signed | 02 June 2010 |
| Lt Col Jason M. Turner, PhD (Chairman) | Date |
| //signed | 07 June 2010 |
| Valeta. Carol Chancey, PhD (Member) | Date |
| //signed | 01 June 2010 |
| Alan R. Heminger, PhD (Member) | Date |

## Abstract

Boolean algebra was developed in the 1840s. Since that time, negation, one of the three basic concepts in Boolean algebra, has influenced the fields of information science and information retrieval, particularly in the modern computer era. In Web search engines, one of the present manifestations of information retrieval, little use is being made of this functionality and so little attention is given to it in the literature. This study aims to bolster the understanding of the use and usefulness of negation. Specifically, an Internet search task was developed for which negation was the most appropriate search strategy. This search task was performed by 30 individuals and followed by an interview designed to elicit more information about the participants' use or non-use of negation during the task. Negation was observed to be used by approximately 17% of users in the study, suggesting that negation may indeed be infrequently used by Internet users. The data obtained during the post-task interview indicate that lack of use of negation stems from users not knowing about negation, having little experience with negation, or simply preferring other methods, even when negation is one of the foremost appropriate methods.

*When you joined the Army and I left AFIT, leaving only one class and my thesis undone, I did not expect that the opportunity to return and finish it would come five months before the beginning of your yearlong deployment. I am beyond grateful for the opportunity you have given me to finish, even though it meant being apart for a total of 17 months.*

*This thesis is written for you, my husband, and my teammate.*

**Acknowledgements**

Thanks are not enough appreciation to give to my adviser, Lt Col Jason Turner, my sponsor, Dr. Carol Chancey, and my committee member, Dr. Alan Heminger. The support and encouragement that you have each shown during the writing of this thesis has been invaluable. Thank you for professional expertise and your confidence in me.

Thanks are also due to my thesis writing partner for pairing with me to stay on track as we both wrote our theses.

Kristen M Lancaster

## Table of Contents

## List of Figures

Figure Page

## List of Tables

Table Page

# I. Introduction

## Prevalence of Search

In November of 2009, comScore (2009), a digital marketing intelligence company, reported that individuals in the United States collectively conducted 22 billion web searches at the websites of search engines and at websites with searching functionality. *The World Factbook* published by the CIA (2010) gives the Internet population of the United States as 231 million in 2008. According to data provided by the U.S. Census Bureau (2010), the resident population of the U.S. was approximately 304 million at that time. Thus, Internet users accounted for approximately 75.9% of the U.S. population in 2008. Assuming that the percent of the population that are Internet users has not changed, the Internet population of the US in November 2009 was 233 million (U.S. Census Bureau, 2010). Thus, during the month of November 2009, Internet users conducted an average of 94.6 searches each, or more than three per day.

Searching requires a significant time investment. It is quite difficult to quantify the portion of search time that may be improved by user skill or search engine improvement. The time spent browsing results can be considered a reflection of the quality of the search query. A user may, for instance, spend quite a bit of time traversing the web site from one of the results, only to find that it does not contain what he is

looking for. This may be for a variety of reasons. The information he is looking for may not exist, or he may have chosen a poor site on which to look for it, or the search query could have been specified poorly and so results are poor, or the search engine algorithm may not return the best results for the particular query. White and Drucker (2007) reported that the mean search trail length is 476.4 seconds, nearly 8 minutes, and consists of an average of 4.3 queries. This yields an average of 110 seconds to execute and evaluate the results of each query. Based on this metric, US Internet users spend approximately 8 billion man-hours per year on searching!

It is easy to see that improvements in searching stand to have a large effect on both personal and corporate productivity. Even a single second of improvement in the average time spent on a query could yield increased productivity of approximately 73 million man-hours per year in the United States alone. Searching can be improved on two fronts. First, the user can improve his search skills. Second, the search engine technologies can be improved in order to serve the user better.

Search engine improvement is happening rapidly. Google, the industry leader in web search, "will introduce 550 or so improvements to its fabled algorithm…" this year alone (Levy 2010). More than two improvements every single working day of 2010 will change and hopefully improve the way the Google interprets a user's query and returns matching results. Google processes two out of every three web searches in the United States, nearly four times as many as the next most popular search engine according to comScore (2009). It is no wonder that a survey conducted by *Nature* (2010) revealed that among Chinese researchers, 83.8% of them stated that Google is so vital to their work

that their research would be somewhat or significantly hampered if they were no longer able to access it, even though other search engines are available.

User search skill improvement may also make a significant difference in obtaining satisfying results and is the overarching consideration here. Even though search technology is improving and becoming more user-friendly, it stands to reason that users should bear some of the responsibility for improving their search prowess. Information retrieval systems play a role not only in improving their own algorithms for interpreting queries and returning results, but also in encouraging the growth of user search skills. Search skills can be encouraged by, for example, suggesting queries or formulating complex queries from entries made on the advanced search page. Research into users' thoughts and behaviors regarding the search process provides a foundation for information retrieval systems to improve and encourage improvement in their users. The vast quantity of literature in Information Retrieval is a testament to the gap that search engines are trying to bridge between a user query and the actual information need of the user.

**Preliminary Definitions**

The details of these definitions and their role in Information Retrieval will be further explored in Chapter 2. For the sake of this analysis, the following definitions will be used. A query is a word or string of words specifying a request to a store of information, such as the indices cataloging the web that are maintained by Google. Many queries may also include operators to more clearly specify the request. In the case of

Google, popular operators include the Boolean operators AND and OR, as well as the minus sign (-) in lieu of the NOT operator. Quotation marks are used to indicate that the enclosed words must appear in sequence. A search including the use of operators is generally referred to as an advanced search.

In order to quantify the effectiveness of a search engine, two metrics are often discussed—precision and recall. Precision measures the fraction of retrieved documents that are relevant to the search. Recall measures the fraction of relevant documents that are retrieved (Morville, 2005, p. 53). Two types of relevance are important to our discussion: user relevance and topical relevance. User relevance refers to the user's judgment of relevance based on his information need, while topical relevance refers to an impartial judgment of relevance based on the specified query (Croft, Metzler, & Strohman, 2010, p. 234). Croft et al. (2010, p. 5) show the difference between user and topic relevance with an example query for *severe weather events*. Specifically, a user may not consider a result relevant if it is about an event that occurred in another part of the world, or happened many years ago, or is in another language that the user cannot read. To an unbiased observer, results such as these are certainly topically relevant. Topically relevant results are much simpler to produce than results relevant to the user because no context about the user or his information need is required.

**User Search Behaviors**

How are users' searches formulated and executed? Suppose that an individual, call him Andy, was at his neighbor's house and that his neighbor had just acquired a

playground for his children from Rainbow Play Systems, Inc. ®. Andy is very interested in getting one for his son Opie, but only after he checks out some reviews and looks for the best price. He quickly goes home and opens his trusty Google search engine and types in the query *Rainbow*. He thinks briefly about adding another word like playground or playset or fort, but there are too many commonly used names for this toy to select one that he believes widely represents the toy. He does implicitly understand that the Google algorithm closely approximates a search for the pages which contain all of the words in the query. So, he does not add any additional terms to his query out of concern that he might miss out on other relevant references to this play system. He also figures that the results might give a clue as to what most people call it and he can add that term to his query. Additionally, he is fairly confident in Google's page rank algorithm, though he doesn't know it by this name, and expects that the top search results will match his desire.

To illustrate, let U be the set of all the pages indexed by Google and let R be the set of results returned by Andy's search for *Rainbow* as seen in Figure 1. Google has many rules that factor in to its algorithm and determine if a page matches a query, but for simplicity, we generalize by saying that the elements of R are pages that are *topically relevant* to the concept R, as determined by an impartial observer. In this case, that observer is the Google algorithm. Based upon Google's algorithm, elements of R will also generally contain the words indicated in the query and so this understanding is interchangeable with topical relevance and will henceforth be described using the term *about*.

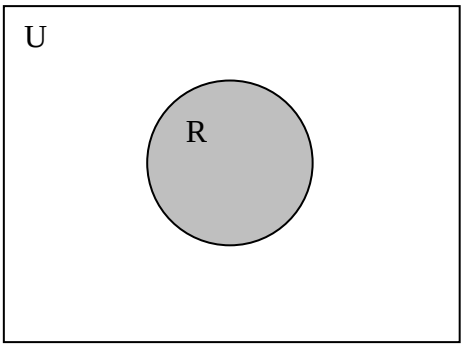


**Figure 1: Results of Search for *Rainbow***

What Andy discovers is that his results are far different than the homogenous response depicted in the previous figure. Let V be the pages about Rainbow vacuum cleaners (something Andy did not think about!), let S be about the type of spectral rainbow seen in the sky, and let P be the pages about the Rainbow playground set in which Andy is actually interested. The following diagram is a simplistic representation of what Andy observes within the results set. For simplicity, no other types of search results exist within the search for *rainbow*, except those described by *P*, *V*, and *S*. Thus $R = P \cup V \cup S$. Figure 2 shows this interpretation of the results.

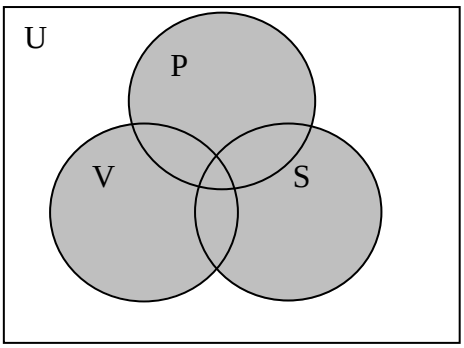


**Figure 2: Results of Search for *Rainbow***

In this case, Andy's need for information is the area indicated with hash marks in the Figure 3. He is interested in as much information as he can find about the playground set he wishes to purchase, but is not interested in vacuum cleaners or spectral rainbows.

Even the pages that would exist in the regions $P \cap V$ and $P \cap S$ are unlikely to be of value to him. Not only are there very few of them to worry about, but if they cover two very dissimilar topics, they are probably of little use to him. Pages in these regions may be the likes of spam pages or Wikipedia disambiguation pages.

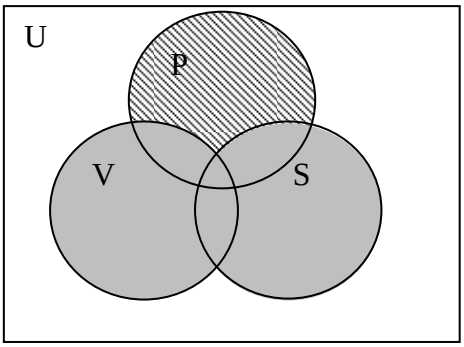


**Figure 3: Results of Search for Rainbow**

So far, this example has been considered across the universe that is Google's entire index, but Andy is an average user and according to iProspect (2006), a search engine marketing research firm, he is very unlikely to click a link beyond the first three pages of results. iProspect also states in their white paper that 88% of users who do not find what they seek in the results of an initial query will either change their search terms or the search engine they are using. Thus, this likely query reformulation is the prime opportunity for Andy to improve his chances of finding the variety of relevant pages for which he is looking.

At this point, Andy may join the majority of average users in his reformulation and add words to his query (iProspect, 2006, p. 3). First, he might search for *rainbow playground*, as represented by the region L in Figure 4 in an attempt to isolate the region described in Figure 3 to be his desired result set.

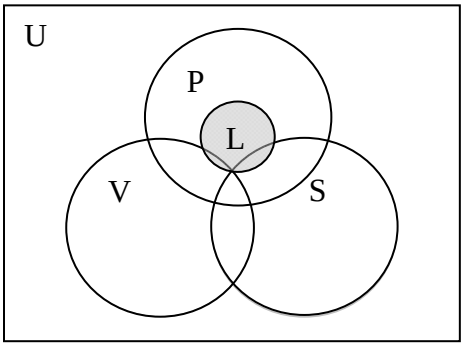


**Figure 4: Results of Search for rainbow playground**

After browsing the results, Andy might proceed with another query such as *rainbow swing sets*, seen as set W in Figure 5.

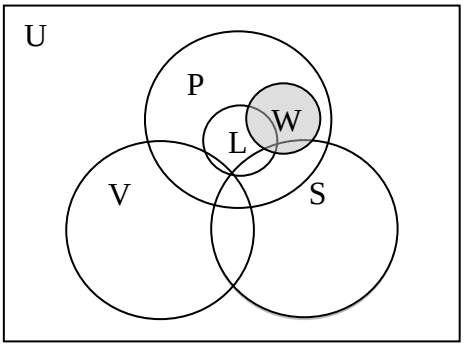


**Figure 5: Results for Search for *rainbow swing sets***

In both of these cases, Andy has sacrificed recall for precision. He has certainly eliminated irrelevant results, thus improving precision, but he has sacrificed many relevant results, thus degrading recall. Consequently, his objective of extensively reading reviews and blogs before buying might be hampered. Specifically, he has given up access to the region $W' \cap L' \cap P$ and has already browsed two results sets to see the limited selection of pages that he has seen.

**Negation**

Though not all information needs result in queries that are candidates for being improved by negation, Andy's query provides great opportunity to demonstrate how the successful use of negation can improve recall, a goal commensurate with his information need. Andy's information need requires the broadest possible coverage of *P*, and there does not exist a single word to add to his query that would closely match *P*. In this case, Andy could change his original search of just *rainbow* to *rainbow –vacuum -sky* and could be fairly certain the he has not excluded important search results as he did with the other queries just described. For continuity, consider the following diagram (Figure 6) representing Andy's new search results. Instead of conducting multiple queries to try to view subsets of *P*, by using negation, Andy eliminates the results not relevant to his information need. In the case of Andy's information need, the fact that he has not eliminated valuable results is important.

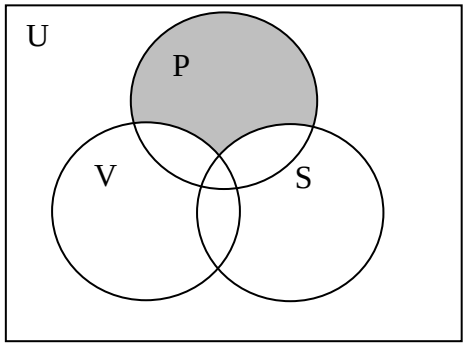


**Figure 6: Results of Search for *Rainbow -vacuum -sky***

In fact, when negation is properly executed within the context of an appropriate information need, there is the potential for increasing recall. Remember that precision is the percentage of retrieved documents that are relevant while recall is the percentage of

relevant documents that are retrieved. Table 1 shows the precision and recall calculations for each of the described search strategies. Remember that $R = P \cup V \cup S$. Also, for simplicity, let $P_1 = P \cap V' \cap S'$.

**Table 1: Precision and Recall**

| | **Precision** | **Recall** |
|---|---|---|
| **Original Query** | $\frac{\lvert P_1 \rvert}{\lvert R \rvert}$ | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert}$ |
| **Addition of non-negated keywords (2 searches)** | $\frac{\lvert P_1 \cap (W \cup L) \rvert}{\lvert W \cup L \rvert}$ | $\frac{\lvert P_1 \cap (W \cup L) \rvert}{\lvert P_1 \rvert}$ |
| **Addition of negated keywords** | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert}$ | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert}$ |

A few sample numbers will make this comparison more readily apparent. Assume $P_1$, $R$, $W \cup L$, and $P_1 \cap (W \cup L)$ contain 8, 24, 1, and $\frac{11}{12}$ million results respectively. Table 2 shows that precision is improved and recall is maintained for the case of the negated keyword. For the case of added non-negated keywords with these values, precision is decreased from the original value, but more importantly, recall is significantly decreased.

**Table 2: Precision and Recall with Sample Values**

| | **Precision** | **Recall** |
|---|---|---|
| **Original Query** | $\frac{\lvert P_1 \rvert}{\lvert R \rvert} = \frac{8}{24} = \frac{1}{3}$ | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert} = \frac{8}{8} = 1$ |
| **Addition of non-negated keywords (2 searches)** | $\frac{\lvert P_1 \cap (W \cup L) \rvert}{\lvert W \cup L \rvert} = \frac{11/12}{1} = \frac{11}{12}$ | $\frac{\lvert P_1 \cap (W \cup L) \rvert}{\lvert P_1 \rvert} = \frac{11/12}{8} = \frac{11}{96}$ |

| | | |
|---|---|---|
| **Addition of negated keywords** | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert} = \frac{8}{8} = 1$ | $\frac{\lvert P_1 \rvert}{\lvert P_1 \rvert} = \frac{8}{8} = 1$ |

It is important to note that negation must be executed by a user carefully and only in the case of appropriate information need in order to be effective. In some cases, a word describing the desired set will easily be found, or a word describing an undesired set may not be readily available. In this example, if a word could be found to describe *P*, negation would not be nearly as valuable. Alternatively, if no good words existed to describe *V* or *S*, then negation would not be the most appropriate search strategy in this case. However, the existence of information needs requiring high recall, such as those performed in the course of research or in a quest for rare collectibles on eBay, is sufficient cause for further investigation into the notion of negation in search.

**Research Question**

Negation is only one tool in the belt of the search engine user. Understanding the searcher's use of negation will further the quest to build search systems that can meet the user at his point of need and his ability to express that need. Though we have demonstrated that negation could be valuable in some situations, users do not use it very frequently (Spink, Wolfram, Jansen & Saracevic, 2001; Markey, 2007). This investigation will begin to explore why that is the case. Research specifically concerning the use of negation in query formulation is limited, so here we begin by asking "When negation would be an appropriate search tactic, do users use negation?"

To answer this question, we must consider situations in which negation may be used successfully to meet a user's information need. Beyond that it may be used successfully, we are specifically interested in the cases where negation is the best or only way of meeting that need. Consider some examples. Negation is valuable in cases when the user requires access to the entirety of a class of results that is not easily characterized by additional search terms and is intermingled with irrelevant results. This was precisely the case seen in the above example. Similarly, a collector might find it difficult to add additional search terms to isolate the items he is looking for due to abnormalities in spelling and description of items for sale. This is a similar scenario to the researcher who is looking for literature on a specific subject. He is looking for all cases where a piece of research matches his need and cannot afford to miss any relevant results. As previously discussed, additional non-negated search terms inherently possess this risk while negated terms, when properly selected, decrease this risk. Consider also the case when a user's results are overwhelmed by irrelevant results. Even though the user does not require an entire class of results in this case, sorting through to find relevant results becomes significantly less difficult if some or all of the irrelevant results can be removed by adding negated search terms. Negation has been considered a valuable tool since the advent of Boolean search schemes and the fact that users are not frequently using negation, despite its value, is the motivation for this investigation.

## II. Literature Review

### What is Information?

To study in the realm of information retrieval, one must first consider the notion of information itself. Is it possible to begin an in-depth study of a topic without first obtaining a baseline definition from the likes of Webster? The American Heritage Dictionary gives a variety of definitions of information.

> 1. Knowledge derived from study, experience, or instruction.
> 2. Knowledge of specific events or situations that has been gathered or received by communication; intelligence or news…
> 3. A collection of facts or data: *statistical information.*
> 4. The act of informing or the condition of being informed; communication of knowledge: *Safety instructions are provided for the information of our passengers.*
> 5. *Computer Science* Processed, stored, or transmitted data.
> 6. A numerical measure of the uncertainty of an experimental outcome.
> 7. *Law* A formal accusation of a crime made by a public officer rather than by grand jury indictment. (Information, 2010)

In these general definitions the words data and knowledge are used in a way that is deeply entwined with the word under consideration—information.

For a more technical definition of information, consider that provided by the *International Encyclopedia of Information and Library Science.* Information is "data that has been processed into a meaningful form…capable of communication and use…; the essence of it is that meaning has been attached to the raw facts" (Information, 2003, p. 244). Knowledge, on the other hand, is defined to be "information evaluated and organized in the human mind so that it can be used purposefully" (Knowledge, 2003, p. 341). Data, as the foundation of both of these definitions is "[n]umerical or other

information represented in a form suitable for processing by computer" (data, n.d.). For example, the number 29,029 is data, but unless the data is given context and units it is rather meaningless and is data rather than information. In this case, stating that Mount Everest is 29,029 feet tall is information. When a person learns and internalizes this information, he now possesses knowledge and may then use the knowledge to plan his ascent of the mountain, or the distance from the mountain he must be in order to photograph it, or whether he must fly around it instead of passing overhead.

### *Theories of Information*

The history and evolution of information theory are integral to understanding the field of information retrieval today. In 1948, Claude Shannon published his *Mathematical Theory of Communication.* Shannon, employed by Bell Labs at the time, was understandably concerned with the transmission of information. His theory was a "measurement for the entropy of a message" (Wersig, 2003, p.310) and was intended to be used to "[optimize] the transport of signals on a communication channel." Though his work did not directly or immediately inform an understanding of information within the scientific community, it was the foundation for his work with Weaver 1949 in which they proposed three levels of problems associated with communicating information. These problems help to refine an understanding of what information is. First, Shannon and Weaver (1949, p. 96) address the technical problem by asking, "How accurately can the symbols of communication be transmitted?" Semantic problems are concerned with "How precisely do the transmitted symbols convey the desired meaning?" (1949, p. 96)

Lastly, the effectiveness problem, stated as a question, is "How effectively does the received meaning affect conduct in the desired way?" (1949, p. 96) These questions defined the context in which the formal study of information blossomed.

Wersig (2003, p. 312) goes on to state that two main issues contributed to the confusion regarding the nature of information. First, technology became closely tied to information such that as the technologies changed, so did the definition of information. Second, the birth of the field of information science around the 1950s, as a trade profession, added to the confusion with its focus on information management and diverted attention from a theoretical definition of information. Nevertheless, a great deal of thought and study has been devoted to the subject of information in its many possible forms.

In 1971, Wersig published a detailed study of the definitions of information in use across a spectrum of fields of study. For instance, the philosophers' definition hinged upon the structure approach in which the "structures of the world – whether perceived by man or not – *are* information" (Wersig, 2003, p. 312). Scholars of decision theory adopted the knowledge approach wherein information is "knowledge developed on the basis of perception" (Wersig, 2003, p. 312). Mathematical communication theory continued the perspective developed by Shannon and Weaver which contended that "information is the message itself" (Wersig, 2003, p. 313). Computer science took the "meaning approach [stating] that the meaning assigned to signs or data is the information, usually connected to the conventions used in the decoding of the sign" (Wersig, 2003, p. 313). The effect approach, touted by behavioral scientists states that "information is a specific effect of a specific process, usually on the part of the recipients in a

communication process" (Wersig, 2003, p. 313). The final approach to defining information described by Wersig was the process approach espoused by information practitioners which "states that information is neither objective nor subjective, but a process" (Wersig, 2003, p. 313).

What had been called the effect approach grew into the cognitive approach and eventually into the subjective approach in the 1980s and 1990s. Two main strains of this theory are evident. The first is that information is the output of knowledge held by a knowledge bearer for the purposes of communication. The later is that information is that which affects a recipient's actions. Thus we see a duality where information is both the product of knowledge and that which affects knowledge. (Wersig, 2003, p. 313).

Ongoing development of information theory seemed to presume that something we would intuitively understand to be information exists without formal definition within a particular situation such that the situation helps to define the concept of information. Wersig (2003, pp. 31-315) describes four such theories that are in vogue today: constructivism, systems theory, action theory, and modernization theory. The constructivist model is one in which information exists only within an individual's perceptions and constructs of reality. Systems theory expresses information by describing it as the catalyst for a change of state of a system. In doing so, systems theory defines "information as a choice for something and thus against everything else that competes" (p. 315). Out of the field of sociology, action theory developed. This theory maintains that in order to make a specific decision, an actor needs knowledge, expressed in a way readily usable by that particular actor in a particular situation. This expression of knowledge is information. Finally, modernization theory is a work in progress that is

contingent upon continued development of a post-modern understanding of science. The basic idea is that information is a guide in various knowledge realities or perhaps even less clearly, in the way of post-modernism, "information as the development of ordering structures within the ambiguous." (p. 316)

Throughout each of these four modern information theories is a common theme of complexity. For example, Constructivism is founded on complexity of the self. Systems theory uses self as a frame of reference and is concerned with external complexity relative to the reference frame of self. Complexity of action is reduced by information in action theory. Finally, the complexity within modernization theory is simply that there is no certainty. This notion of complexity gives rise to the challenges within the information science community which must contend with great complexity surrounding the producers and consumers of information as well as the management of stored information.

**Information Science**

Information science is a very practical discipline. Though still concerned with a theoretical definition of information and the nature of the transmission of information, the discipline in no way "los[es] sight of the practical aspects of collecting, collating and evaluating information and organizing its dissemination through appropriate intellectual apparatus and technology" (Bottle, 2003, p. 295). Restated, information science is "concerned with the gathering, manipulation, classification, storage, and retrieval of recorded knowledge" (Information Science, n.d.).

This practicality is a result of the origins of information science as a discipline. This field began in the early 20th century when scientists began having more difficulty finding desired information (Bottle, 2003, p. 295). Scientists who assisted other scientists in finding information began to be called information scientists in the 1940s; however, the first use of the term information science to denote a field of study was in 1956 (Bottle, 2003, p. 295). In contrast to an information scientist, a librarian is the "curator of collections of books and other information materials, administering conditional user access to these collections" (Sturges, 2003, p. 370). Because information scientists came out of the scientific tradition, information scientists, distinct from librarians, often had enough knowledge or skill to evaluate sources of information being returned to a client.

Information science is a broad field, embracing concerns ranging from how to catalog collections of materials to how long a user might spend looking at each of the ten search results returned from a web search engine. In order to study users' ability to use negation when it is an appropriate tool, we must consider of the topics of information retrieval and web searching behavior. We begin with information retrieval, which is the science of supporting information seeking behavior (Göker, 2003, p. xxi). After reviewing information retrieval, we will address search engines as information retrieval systems. Within that context, web searching behavior readily follows and leads directly into a discussion of users' use and understanding of negation.

## Information Retrieval

Though the genesis of information science was as a profession of assisting researchers in finding information, the science reached a new level during the rise of the computer. This hybridization with the field of computer science produced the field of information retrieval. Dr. Vannevar Bush was one of the foremost American scientific leaders in an era of enormous scientific growth that included the development of the atomic bomb. He was one of the first notable individuals to speak publicly about the future difficulties of finding information that would develop as the production and storage of information continued to dramatically increased (Bush, 1945). This difficulty arises because it is not easy to find, sift, and sort through vast amounts of information in a reasonable amount of time. These are precisely the challenges that the field of information retrieval seeks to conquer today.

Information retrieval can be broken down into four component topics—information need, information search, information retrieval systems, and information access (Göker, 2009, p. xxi). These topics are deeply related. Generally speaking, a user begins with an information need. This need must be expressed by the user as a search statement, which is commonly called a query. The topic of information search encompasses the journey of the user in the entirety of his quest to satisfy his information need. Information retrieval systems are technologies and processes that interface with the user during his information search and try to discern and return to him the information he desires. Information access covers rights to information and includes topics such as

copyright, security, and privacy. Each of the first three of these topics is relevant to the question at hand and will be discussed in turn.

### *Information Need*

Taylor (1962) gives four levels of information need which will clearly influence the discussion of information search. First, there is the *visceral* need, "the actual, but unexpressed, need for information" (p. 392). The *visceral* need gives rise to the *conscious* need, which Taylor describes as the "within-brain description of the need" (p. 392). The *formalized* need is easily understood as the "formal statement of the question" (p. 392). Lastly, information need becomes information search with the expression of the *compromised* need, or query, defined as the "question as presented to the information system" (p. 392).

The unique name for the last level of information need is a clue as to the difficulties encountered at the crossroads between information need and information retrieval. At this level, the user must anticipate how the system to which he submits his information need will interpret his query. Though such systems are improving, the nuances of the information need are generally lost on a sterile information system that does not know the context within which the user operates or his history or background which will influence his perception of the relevance of a result. A system rarely has more than the query upon which to base a judgment of the relevance of results.

Taylor goes on to discuss the "variables… which affect the question and its formation" (p. 393). He presents the factors which affect the process by which the

information need becomes an information search. First, there are the general or practical factors. This includes concerns about whether the system is manual or automatic or how soon the results can be returned to the user. For example, in a system where the user must wait days or pay high costs for the results from a query, the user will put much more time and effort into formulating a precise query. System-input factors include whether the system takes a verbal description or a full sentence or a drawing as the query. The user must know about and adapt to the requirements of the system for input into it. Internal organization is the issue of how information is stored within the system. This includes information like metadata and whether the full text of included materials is indexed. The question input factor is the human user element and deals with the likes of how capable the user is in expressing his need in the form of a query appropriate for the system. Output factors are the interim feedback of the system during the quest for satisfaction of the information need. One query may provide feedback to the user for refinement of his query (Taylor, 1962).

### *Information Search*

The information search process, proposed by Kuhlthau (1991), comes from a user perspective of information seeking. The process has six steps. In the first step, *initiation*, the individual becomes aware of a lack of information. During this phase, the individual will begin to collect background information concerning the vacancy. *Selection* refers to the choosing of a general topic or approach for further investigation. The *exploration* phase is an unstable portion of the information search process as the person begins to

explore the vastness of the general topic under consideration. At this point the individual still does not know exactly what information is required. *Formulation* follows and is the beginning of clarity concerning the desired information. *Collection* is the period when information that meets the user's need is gathered. Lastly, *presentation* is the formalization and synthesis of the found information. These stages of information search formalize the generic process for a user to obtain resolution for an information need.

### *Information Retrieval Systems*

Discussion of information retrieval systems as a topic within the field of information retrieval, and alongside topics like information need, information search, and information access, may be somewhat confusing. Simply, an information retrieval system is a method of indexing holdings, accepting a request from a user, and comparing the request to the index in order to return results to the user. Information retrieval in the topical sense is concerned with the information system and how it processes input from a user.

## Search Engines as Information Retrieval Systems

Though information retrieval systems come in all shapes and sizes, search engines are under consideration here and serve as an example. However, due to the focus on the user side of the equation, only basic properties of search engines will be covered. A web search engine functions by crawling the Internet using automated scripts and indexing the

websites and documents found. Indexes built by crawling the web are catalogues of web pages and the words they contain. Information about each document found, such as metadata and a snippet containing pertinent parts of the text, are also maintained. When a query is presented to the search engine, intricate algorithms, which are trade secrets, are applied to the index and to the query to determine which documents and web sites match the query and to order those results appropriately (Hepworth & Murray, 2003).

In the following sections we will consider three aspects of the task assigned to search engines. First, we will consider the information retrieval models describing how a query may be matched with an index to produce results. Then, we will consider how each query is interpreted so that a match may be made. This will give us insight into how the user expresses information needs as search queries. Lastly, we will look at the measurements used to determine how effectively a search engine meets a user's information need by returning relevant results.

### *Information Retrieval Models*

Though web search algorithms are trade secrets, they are built upon significant research into three information retrieval model classes—exact match, vector space, and probabilistic models. The most rudimentary information retrieval model is the exact match or Boolean model. This model operates on the principle that a query directly defines a set of results by word matching. The value of a model like this is the user's ability to explicitly control results through the use of the Boolean operators AND, OR, and NOT. For example, a query for *tree AND paper* would return all documents using

these two words. A query for *tree OR paper* would return all documents using either of the two words. Lastly, a query for *tree NOT paper* would return only documents containing the word tree that do not contain the word paper. In Boolean logic, these operators can be combined extensively with the use of parenthesis to define order of evaluation. For example, the query *(tree OR chart) AND decision AND NOT (paper AND writing)* is a valid query. As Boolean models are a very strict system of determining relevance, they are straightforward in application, thus allowing the user to be very specific. This ability for specificity allows users to clearly understand why the system returns some results and not others and to control exactly which results are returned. To some degree, all search engines implement the Boolean model by allowing for the use of the operators, though the operators may be treated within the context of a larger set of rules rather than strictly by their definition (Göker, 2009, pp. 3-4).

Unfortunately, Boolean models do not result in the ranking of results, and so sorting through a large volume of exact match results can prove difficult or impossible for a user. The vector space approach to information retrieval addresses the lack of ranking in the Boolean model. Recall that a vector is a geometric construct that is described by its length and direction, but is frequently represented by assuming the origin of the space as the initial point and expressing the coordinate of its terminal point. The Vector Space model describes each document as a vector in an n-dimensional space where each of the n axes is a term in the index. The query is also represented as a vector in this n-dimensional space. For example, if an index contained three terms and a document contained only the first two of these terms, the vector would be (1,1,0). The vector inner product is the score assigned to each potential search result as a means of

comparing them to the query vector. The higher a document scores, the higher it ranks in the results. Various interpretations of the vector space model have been proposed using different scoring metrics and different representations in n-dimensional space, but all of these interpretations are still geometric in their approach. The classic vector space model, though one step more advanced than Boolean models in its ability to rank results, does not implement any sort of interpretation of a query beyond the discrete words and operators of which it is composed. This is where probabilistic models come in to play (Göker, 2009, pp. 5-8).

Probabilistic models are wholly about term weighting. Within this model, probabilistic predictions are made as to whether a particular document is relevant to a given query. Language concerns are highly integrated in probabilistic models. For example, they may take into account use of words that belong together both in a document and in a query. That is, they may be designed to recognize that in a query for *North Carolina* the words belong together and so should also be related in highly relevant documents. The PageRank model used by Google also falls within this class of models. PageRank identifies pages to rank highly based on the number and quality of other pages that link to them. To understand the probabilistic aspect of this model, consider surfing the web, navigating from page to page. The pages that are most likely to be viewed because of significant incoming links from other highly viewed pages will be ranked first. The probability of arrival at any given page is the PageRank. PageRank, unlike many other metrics for ranking pages, is static, is not based on the query, and can be computed at index time (Göker, 2009, pp. 8-15).

In the models that incorporate ranking, a variety of indicators in a web page, besides those already mentioned, are used to indicate relevance. For example, Google takes special interest in the title of a page, how recently a page was updated, the geographic area that might be associated with a page, and whether the page is from a recognized authority (Levy, 2010).

Each of the models introduced here plays a role in the history of information retrieval systems, and remnants of each are left in the systems in use today. Most systems will implement aspects of a variety of models. However, before these models can be directly used to return results, the user's intention must be rendered from the query.

### *Query Interpretation*

According to Croft et al. (2010), web search queries fall into three distinct categories. Informational queries are executed by users looking for topical information. Navigational queries are intended to return a particular web site that the user knows exists. For example, in order to ensure that he goes to the correct website, a user may execute the query *Dick's* or *Dick's Sporting Goods*. His search engine, which is in safe search mode, will correctly return the website for the sporting goods store rather than the undesired destination at the seeming logical address of www.dicks.com. Such queries are valuable because they allow the search engine and its determinations of authority of a website to inform the user on the correct website for the desired entity. Transactional queries are similar to navigational queries; however, the exact destination is unknown.

Instead, the user knows that he wishes to participate in an activity on the web such as purchasing a good or playing a game and is looking for a web site conducive to his goal (Croft et al., 2010, pp. 279-280). Identification of a query as fitting into one of these categories is one facet of many used by a search engine in order to determine the relevance of a result. For example, a query for *www.yahoo.com* may be interpreted by a search engine to be navigational and so the search engine may use a different algorithm than if it was not a navigational query. In this case, the algorithm could call for the query to be first checked against URLs rather than keywords and the search algorithm may also look for results just containing the word Yahoo.

In addition to these vastly different user intentions in web search, language interpretation of queries presents another challenge entirely. There are a variety of fronts on which search engines must fight to understand the meaning of a query. Noted columnist for *Wired* Steven Levy (2010) gives a firsthand look at some of these challenges, some of which we explore here. Winning each of these fronts contributes to a successful search engine. The first front is spelling. Typing errors and lack of knowledge of how to spell a word both can lead a user to enter a search term that is misspelled. Search engines may even seek to identify cases where the misspelling resulted in a different actual word.

Another consideration that distinguishes the best search engines is the complexity and effectiveness of their stemming algorithms. Stemming algorithms are used to broaden the query's reach so that the user does not have to explicitly do so. For example, if a user executes a query that includes the word *clean*, he may also be interested in pages that would be identified with the word cleans or cleaning or cleaner or a host of others, so

the algorithm interprets the query as matching a document with any of these variations of the original word.

Queries are complex representations of information need, and because information itself does not exist as discrete words, interpretation of queries in their real-life context is imperative. An n-gram is a group of words that are used together. If a user issued a query for *New York Times*, successful negotiation of the n-gram would involve the search engine recognizing this triplet as a single entity. In a successful search engine, identifying these three words together should not carry over to the interpretation of the query *New York Times Square*. Names are closely related to the problem of n-grams, but present their own special challenges with optional middle names or middle initials, forms of address, and misspellings. Colloquialisms are also a challenge to search engines. If a user executes a query for *President Lincoln bio*, he means biography, but if he executes a query for *bio warfare*, he means biological. This whirlwind trip through some of the challenges faced in evaluation of a query provides evidence of complex nature of query evaluation and establishes the context in which we will investigate negation.

### *Metrics of Effectiveness*

Relevance, as mentioned before, is the ultimate goal of information retrieval systems. Both user relevance and topical relevance were defined in Chapter 1. In light of relevance, two primary metrics for measuring the performance of a search engine in returning relevant results stand out—precision and recall. As mentioned in Chapter 1, precision is measured as the proportion of pages that are retrieved that are relevant and so

shows how well a search engine does at identifying and excluding irrelevant documents in the search results. In research, precision is generally measured with respect to topical relevance since the users' information need is frequently unknown beyond the details included in the query. In contrast with precision, recall is the percent of pages that are relevant that are retrieved and thus measures how well a search engine returns all of the relevant pages.

The precision metric can also be limited to evaluation of only the first $n$ search results and is written as $P@n$ and read "precision at $n$" where $n$ is the number of results that have been included in the metric. So $P@3$ is the precision only taking into account the first three results so that if the first result was relevant, but the next two were not, $P@3$ would be $\frac{1}{3}$. This is a valuable metric in accordance with the findings (iProspect.com, 2006) that users will generally browse no farther than the third page of results. Unfortunately, the P@n metric still does not differentiate between two cases where the number of relevant documents is the same, but they are returned in a different order. For example, P@10 does not distinguish between the case where the first two results are relevant and the case where the last two results are relevant. P@10 would be 0.2 in both cases. To distinguish between the two different results sets most recently described, the metric of average precision may be used (Croft et al., 2010, p. 310-311).

To calculate the average precision over a set of results, calculate the P@n values at every position where a relevant document is listed and then average those values. For example, in a set of 10 results where only results 1, 2, 3, and 10 are relevant, first calculate P@1, P@2, P@3, and P@10 and then average these values. To see how this metric compensates for the failing of the P@n metric just discussed, consider the

following example. To calculate average precision of a set of 5 results where only results 1 and 3 are relevant, the following calculations would be appropriate: $\frac{P@1+P@3}{2} = \frac{1+{}^{2}/_{3}}{2} = \frac{5}{6} \cong 0.83$. For a set of 5 results where only results 4 and 5 are relevant, the average precision is $\frac{P@4+P@5}{2} = \frac{{}^{1}/_{4}+{}^{2}/_{5}}{2} = \frac{13}{40} \cong 0.325$. Thus, the metric of average precision gives a better measure when relevant results are ranked higher even if an equal percentage of relevant results are returned (Croft et al., 2010, p. 311-313).

Each set of results may be measured by the above metrics, but in order to measure the overall effectiveness of a system, the most common metric is the mean average precision. This metric is calculated by obtaining the average precision for all of the queries on a system. These average precision metrics are then averaged together. Generally, this 'complete' list of queries is finite and based upon a test set of queries and documents which has been assigned relevance judgments (Croft et al., 2010, pp. 313-314).

These and many more metrics may be used to describe the effectiveness of a search engine from a mathematical perspective. However, the user is only incorporated in these metrics to the extent that he is the judge of relevance. In order to design and understand the performance of search engines and other information retrieval systems in the wild, the ways that a user interacts with the system must be explored.

## Web Search User Behaviors

Evaluation of a search engine is not limited to statistical metrics, but also includes the user element. As described previously, search providers will continue to improve the relevance of the results returned to their users by improving the underlying technologies and developing new algorithms to more successfully interpret queries to get to the heart of each user's information need. But between the information need that vaguely exists within a user's mind and the production of results to meet that need is the user with his query formulation skills, however limited they may be. The user has the power and responsibility to formulate his information need into a query based upon his understanding of the operation of the information system. Here, then, we survey the literature describing the behavior of the average searcher so that the context for the potential usage of negation is established.

In the late 1990s, a significant portion of information regarding the average searcher came from query logs. Two major search engines, Excite and AltaVista, each released a number of logs for public study. The most recent release of logs came from AOL in 2006.

The Excite search engine released query logs in 1997, 1999, and 2001. Spink, Jansen, Wolfram and Saracevic (2002) summarized research on all three of these logs, including the research of Jansen, Spink, and Saracevic (2000) and Spink et al., (2001). Spink et al. (2002) produced the metrics in Table 3. Of note are the lack of significant change in average query length and average number of pages viewed and the increase in Boolean operator usage from 1997 to 2001. Another noticeable change is that the

number of users going to later pages decreased significantly. This may reflect a significant improvement in the desired results being returned in earlier results pages and improved rankings within pages, or it may indicate that users were giving up much more quickly. It is valuable to recognize that in June 1998, Excite was the second most popular search engine (Sullivan, 1998), but by October 2001 had become the seventh most popular search engine (Sullivan, 2001). This information adds another confounding factor to the analysis because of the possibility that the most popular search engines are the best and so draw the most sophisticated users who might reasonably be expected to be better information seekers from the start. For example, analysis of a log from a search engine used by less sophisticated users may reveal a lack of use of advanced search techniques, while the general population may use them regularly.

**Table 3: Excite User Statistics from Spink et al. (2002)**

| Comparative statistics for Excite Web query data sets-- one million queries per study | | | |
|---|---|---|---|
| **Variables** | **1997** | **1999** | **2001** |
| **Mean terms per query** | 2.4 | 2.4 | 2.6 |
| **Terms per query** | | | |
| **1 term** | 26.3% | 29.8% | 26.9% |
| **2 terms** | 31.5% | 33.8% | 30.5% |
| **3+ terms** | 43.1% | 36.4% | 42.6% |
| **Mean queries per user** | 2.5 | 1.9 | 2.3 |
| **Mean pages viewed per query** | 1.7 | 1.6 | 1.7 |
| **Pages viewed per query** | | | |
| **1 page** | 28.6% | 42.7% | 50.5% |
| **2 pages** | 19.5% | 21.2% | 20.3% |
| **3+ pages** | 51.9% | 36.1% | 29.2% |
| **Users modifying queries** | 52.0% | 39.6% | 44.6% |
| **Session size** | | | |
| **1 query** | 48.4% | 20.8% | 30.8% |
| **2 queries** | 60.4% | 19.8% | 19.8% |
| **3+ queries** | 55.4% | 19.3% | 25.3% |
| **Boolean queries** | 5.0% | 5.0% | 10.0% |
| **Terms not repeated in the data set** | 57.1% | 61.6% | 61.7% |
| **Use of 100 most frequently occurring query terms** | 17.9% | 19.3% | 22.0% |

AltaVista released portions of its logs in 1998 and 2002. Jansen, Spink and Pedersen (2005) reported metrics on these query logs. Interestingly, AltaVista was the fifth most popular search engine in 1998 (Sullivan, 1998) and had been downgraded to the ninth most popular by October of 2001 (Sullivan, 2001). Table 4 shows the results of Jansen et al. as a basis for comparison to the Excite search engine results. The later query logs for both engines show an uptick in the average number of query terms. Logs from both search engines show that users were increasingly stopping after viewing only one page. Unfortunately, usage of Boolean operators is difficult to compare between the two search engines as they are grouped with other operators, (e.g. *site:*) in the AltaVista report. However, the AltaVista logs show evidence of approximately 20% of queries containing Boolean or other advanced operators. The Excite logs show 5-10% usage of Boolean operators alone. Some of the disparity in these numbers may come from the AltaVista study more liberally counting terms like *And*, *Not*, and *or* as attempts to use Boolean operators, though only fully capitalized versions are actually accepted by the engine.

**Table 4: AltaVista comparison from (Jansen et al., 2005, p. 563)**

| Variables | AltaVista 1998 | AltaVista 2002 |
|---|---|---|
| **Mean terms per query** | 2.35 (sd = 1.74) | 2.92 (sd = 1.91) |
| **Terms per query** | | |
| **1 term** | 25.8% | 20.4% |
| **2 terms** | 26.0% | 30.8% |
| **3+ terms** | 27.6% | 48.5% |
| **Mean queries per user** | 2.02 (sd = 123.4) | 2.91 (sd = 4.77) |
| **Users modifying queries** | 20.4% | 52.4% |
| **Session length** | | |
| **1 query** | 77.6% | 47.6% |
| **2 queries** | 13.5% | 20.4% |
| **3+ queries** | 6.9% | 32.0% |
| **Results pages viewed** | | |
| **1 page** | 85.2% | 72.8% |
| **2 pages** | 7.5% | 13.0% |
| **3+ pages** | 7.3% | 14.1% |
| **Boolean queries and other operators** | 20.4% | 20.0% |
| **Terms not repeated in data set** | | 5.6% |
| **Use of 100 most frequently occurring query terms** | | 18.9% |

Jansen and Spink (2006), though they did not include the 2006 AOL query logs, analyzed the results from nine query log studies, including those mentioned here. Their conclusions are instructive on the trend in information search. They conclude that "users are viewing fewer results pages" (p. 248), have not increased the number of terms in their queries, and the use of query operators, including Boolean operators, is holding steady.

In 2006, AOL's release of some of its query logs caused an outcry in the media about privacy concerns; AOL retracted their logs in response and no query logs have been released since. Nearly 4 years have elapsed, and much has changed in the search landscape, both on the user side and on the search engine side. Recall that Google has been reported to be making over 500 improvements per year to its algorithms. Even by the time a study of a query log makes it into the literature, things have already changed, so all of the reported statistics and findings must be considered in this context. Additionally, AOL was ranked as the fifth most popular search engine in the US in late 2006 (comScore, 2006), recording only approximately 13% of the number of searches submitted to the leading search engine and only holding 5.9% of the market share.

Brenes and Gayo-Avello (2009) analyzed the 2006 AOL query log but chose to report quite differently from previous research conducted on the AltaVista and Excite query logs. Their study collected queries into groups based upon their frequency of use in the query log. Each of these groups was then analyzed for its characteristics. Unfortunately, because of the grouping of queries and lack of generalized metrics, little comparison is possible to previously analyzed query logs outside of the result that the metrics produced are on the same order of magnitude as analysis of other query logs.

**Usage of Negation**

Negation is the specific web search tool under consideration in this study. Klein (2009), one of the more contemporary sources for a full definition of negation in Boolean queries, defines negation of a word in a query as "prohibiting the occurrence of the negated terms in the retrieved documents" (p. 299). Negation is implemented in all five of the top search engines as ranked by usage by comScore (2010). However, definition and broad implementation of this operator do not necessarily lead to usage of it as seen in the following studies. Unfortunately, beyond these few studies, little current research exists on the usage of negation due to lack of release of query logs for research and a lack of dedicated study for what is perceived to be a little used functionality.

The following studies present a view that negation is a complex, but valuable functionality for a search engine. Clark (1976) gives us a linguistic perspective on negation. In his introduction to his review of literature and studies, he pronounces that "[n]egation is undoubtedly one of the most fundamental conceptual devices of language" (p. 18). He goes on to conclude that "negation is more difficult to understand than affirmation" (p. 19). Even though negation is rarely used by users of Internet search engines, Klein (2009) argues for the importance of research into advanced query syntax. He cites the community of users of information retrieval systems who focus on very specific types of information and require full answers (high recall) to their queries. He includes lawyers and medical professionals in this community. Eastman and Jansen (2003) make a similar case that the average search engine user does not need negation. Scheffler and March (1972) investigated a procedure for proactive distribution of

abstracts to researchers in a research laboratory in which each researcher received abstracts commensurate with a profile (query) prepared by an information specialist based upon written statements of research interest or keyword lists. They documented that the careful use of the NOT operator to reduce the number of irrelevant abstracts was highly effective.

Only one of the studies mentioned here contains more than anecdotal evidence that negation is useful. A number of studies call into question the value of advanced operators. Lucas and Topi (2002) concluded that "search term selection and usage are much more significant predictors of query performance than [operators]" (p. 105). While Eastman and Jansen (2003) indicated that advanced operators probably have a purpose for certain types of users, their study, which did not study negation, shows that other advanced search operators have no significant positive effect on traditional effectiveness measures.

A number of additional studies have indicated the minimal use of negation in search. Spink et al. (2001) provide a detailed analysis of a log from the Excite search engine, which was popular at the time. One of few studies to distinguish between correctly and incorrectly used operators, they reported a high incidence of incorrectly used negation operators. Only 1,963 queries out of 1,025,910, approximately 1 in every 500 queries, contained a correctly used negation operator (- or NOT). Spink et al. contrast their results with previous studies of professional searchers and digital librarians and suggest that the Internet population is not nearly as sophisticated. Similarly, Herskovic, Tanaka, Hersh, and Bernstam (2007), reporting on an analysis of a PubMed query log, indicate that 0.2% of queries used the Boolean NOT operator or its cousin, the

minus sign (-). While percentage of use of this operator is small in the scope of queries, it is certainly existent in all studies of sufficient size. Similarly, Chau, Fang, and Liu Sheng (2005) report a 0.09% usage of the negation operators in queries to a Utah government web site. Within an analysis of a university OPAC, Lau and Goh (2006) found that only approximately 0.02% of queries utilized the NOT operator. Finally, Markey (2007) summarizes a large number of studies by stating that negation is used in less than 2% of searches. This figure is somewhat higher than that reported in the forgoing studies due to the inclusion of some studies where the search engine being studied encouraged the use of Boolean operators. Some library catalog search systems have this feature.

Finally, a number of researchers have implied by their methodology a correlation between use of Boolean operators, including negation, and search expertise. White and Morris (2007) conclude that advanced searchers, defined as those who used an advanced operator (plus, minus, double quotes, and 'site:') and submitted at least 50 queries, are more successful at searching. Lucas and Topi (2002) contrast novice and expert searchers' use of Boolean operators for a series of eight query formulations. For the most complex problem, both advanced and novice searchers used significantly more Boolean operators, on average, indicating the value of such operators. This effect was even more pronounced among the advanced searchers. White and Morris (2007) concluded that use of advanced search syntax is correlated with users being online for more time, "spend[ing] less time querying and traversing search trails" (p. 262), propensity to explore search results, clicking on results more quickly, and success is search.

If users using advanced search operators are indeed more successful at information search, this is certainly cause to explore search operator use further, especially having noted that use of negation is extremely uncommon. However, due to the scarcity of related research, it is unclear if this situation is due to a lack of opportunities to use negation or due to other user issues. In the following chapter, the current investigation into the use, or disuse, of negation will be described. The objective, as identified in the opening chapter of this report, is to investigate if a user will even use negation in a situation when it is advantageous or appropriate to do so.

# III. Methodology

## Participants

The study sample consisted of 30 graduate students assigned to a small, government-run engineering and management university. Two participants were female while 28 were male. The participants ranged in age from 23 to 42, with an average age of $(32 \pm 6)$. Two participants had only obtained a Bachelor's degree. Seventeen had completed a Master's degree and the remaining 11 were partially finished with a Master's degree.

Participants were recruited primarily through the social network of the principal investigator via paper flyers and emails. A study invitation was also posted on a sign at a public location within the university near where the study was conducted. A snowball sampling and referral technique was also used to recruit additional participants as part of a participation incentives program. Specifically, a $100 Amazon gift card was to be given away to one randomly selected participant. Each study participant would initially receive one entry into the drawing. Participants were also provided with three referral coupons; recruiting another participant who returned the coupon would be worth one additional entry into the drawing for both the referrer and the referred individual.

## Materials

Participants were invited to an indoor public café located on the grounds of the government university; where a table was set up for the purposes of this study. The search task was completed on a Dell Latitude E6400 laptop using the Firefox web browser version 3.6.3 with the 'Hide Google Options' Add-On installed. The laptop had the Windows XP Professional Service Pack 3 operating system installed. Participants were provided with an external mouse in order to allow the participant to use whichever pointing device was most comfortable. Screen capture software, Jing version 2.3.100089, which was freely available from TechSmith, was also installed on the laptop and used to record the participants' search task sessions. A horizontal piece of paper was taped across the bottom inch of the laptop screen to obscure the 5 minute countdown timer employed by the Jing software.

A standardized script and set of study materials was used to ensure that the search task and follow-on interview was administered to each participant in the same way. The script and informed consent form are included in Appendix A. A hard copy of an excerpt from Google's Advanced Search Tips page explaining the use of negation was also provided to the study participants. Participants were also provided with a hard copy of the image shown in Figure 7 as part of the narrative describing the search task given to each participant in this study. For the purposes of the search task, the copyright and site information was removed from the picture in order to prevent the participant from using that additional information to find the image during the search task.


**Figure 7: Beetle Image for Search Task (Krásenský, n.d.)**

## Procedures

In order to examine the use of negation in the context of a real-world Internet search, participants were given a manufactured search task in which negation was determined a priori to be the most efficient search option. Upon completion of the search task, participants were then interviewed to explore the factors that may have contributed to their use or non-use of negation. The entirety of the interaction with each participant was designed to last about 20 minutes. Before the formal conduct of the study, both the search task and the follow up interview questions were administered to approximately ten willing colleagues in order to identify and mitigate potential pitfalls.

### *Search Task Design*

The search task was designed to create a situation in which negation would be the most appropriate and effective method for finding the desired information. Specifically, a scenario was described to each participant in which the participant's child had found an image of a beetle (Figure 7) using the Google Image search website. The scenario indicated that the child had forgotten to write down proper citation information for the image to include in his school report; participants were therefore instructed to find the exact image the child had used. The scenario indicated that the first search conducted at the Google Images website was for the term *beetle*, a search that, as of May 2010, produced a large number of images of Volkswagen Beetles as shown in Figure 8. After indicating they were ready to proceed, participants were instructed to press a key on the laptop which would signal the start of their search task session and initiated the screen capture software to record all of their ensuing search activities. Prior to each participant's use of the computer, the recording software was started and then paused in order to take advantage of a hot key to restart the recording. Participants were then allowed to search the Google Images website in whatever way they saw fit. The search task ended when participants successfully found the image or they gave up of their own accord. Alternatively, if the participants continued to search for the image for 5 minutes without success, the study administrator terminated the search task. This amount of time was born both out of software convenience and the pre-testing conducted. During pre-tests, individuals were found to use negation within the 5 minute time frame if they were going to use it at all.

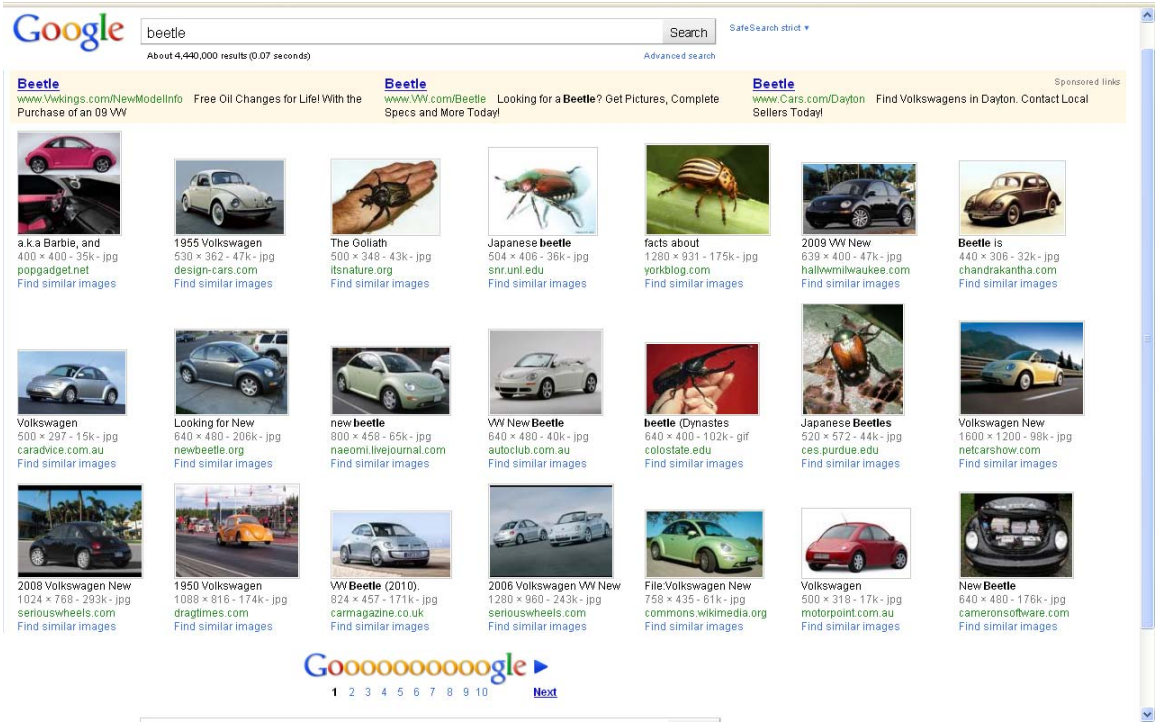


**Figure 8: Search Results for *beetle***

The selection of the test image was based upon the variety of ways that participants searched for the image during the pre-testing. Those ways included paging through the results, further describing the content of the image, describing the image itself (e.g. describing its orientation), and negation. The image was selected to be difficult to find by means other than negation. Stabilization of these factors was impossible due to the ever changing nature of search engine technologies, but this image was selected to meet the needs of the study as closely as possible over time. Specifically, this image never appeared earlier than the sixth page of the results for the *beetle* query during all pre-testing of the study protocol. It was assumed that many participants would attempt to further describe the image, as in the example search described in Chapter 1, by adding words like ‘black’ or ‘plant’ to the query. However, this image was very infrequently returned by the inclusion of such additional search terms. Lastly, the image

was provided without copyright information or digital clues that could be used to locate the image. Some features, such as the image being a full color image, could not be avoided. Thus, the risk of failing to find this image was high without the use of negation to eliminate the multitude of unrelated Beetle automobiles.

The rationale behind of the construction of the scenario was as follows. First, because the scenario describes how the child originally found the image using Google Images, the participant could easily assume that the image itself was, in fact, findable with the Google Images search engine. Additionally, it was reasonable to assume that the search results for the term *beetle* should contain the image. However, choosing a search term not used in the Google index to identify the image caused the image to not be returned. As such, the use of negation proved to be a fairly safe way of narrowing the search results without risking the absence of the desired image in the results. However, negation was only useful if the term negated was unlikely to be associated with the desired image. In this case, it was reasonable to assume that words such as *VW* and *Car* would not be associated with the image and could therefore be safely negated.

During each of the three days of the study, a variety of queries were pre-tested to determine the approximate location of the desired image in the pages of search results. Queries using negation returned the test image within the first page of results. The original query for the single search term *beetle* returned the test image on page 6 of the results on all days of the study. Table 5 describes the pre-tested queries. The locations at which the selected image appeared within the search results were advantageous based on the previously reported findings (Spink et al., 2002; Jansen et al., 2005; iProspect, 2006) that few users will venture beyond the first page of their results and practically none will

venture past the third page. The query for *black beetle* was also included as this was the most common query among pre-test participants.

**Table 5: Position of Desired Image in Results**

| Query | Results Page |
|---|---|
| **beetle** | 6 |
| **beetle -VW** | 1 |
| **beetle -car** | 1 |
| **beetle -Volkswagen** | 1 |
| **beetle -VW -car** | 1 |
| **beetle –car -Volkswagen** | 1 |
| **beetle -VW -Volkswagen** | 1 |
| **beetle -VW -car -Volkswagen** | 1 |
| **black beetle** | Not within first 30 pages |

### *Search Task Procedural Considerations*

Two specific contrivances in this study design must be noted. Less than 2 weeks before this study was conducted, Google modified their search page to include a persistent sidebar for searching the results in a faceted manner as seen in the metadata selection options shown in Figure 9. To maintain and ostensibly isolate user focus on Boolean search, specifically negation, an Add-On was used in the Firefox Web browser to mask this new functionality as shown previously in Figure 8. Additionally, participants were considered to have located the image if it appeared on their screen, even if they did not identify it as a match. This was done to isolate the value of negation rather than the participants' matching abilities.

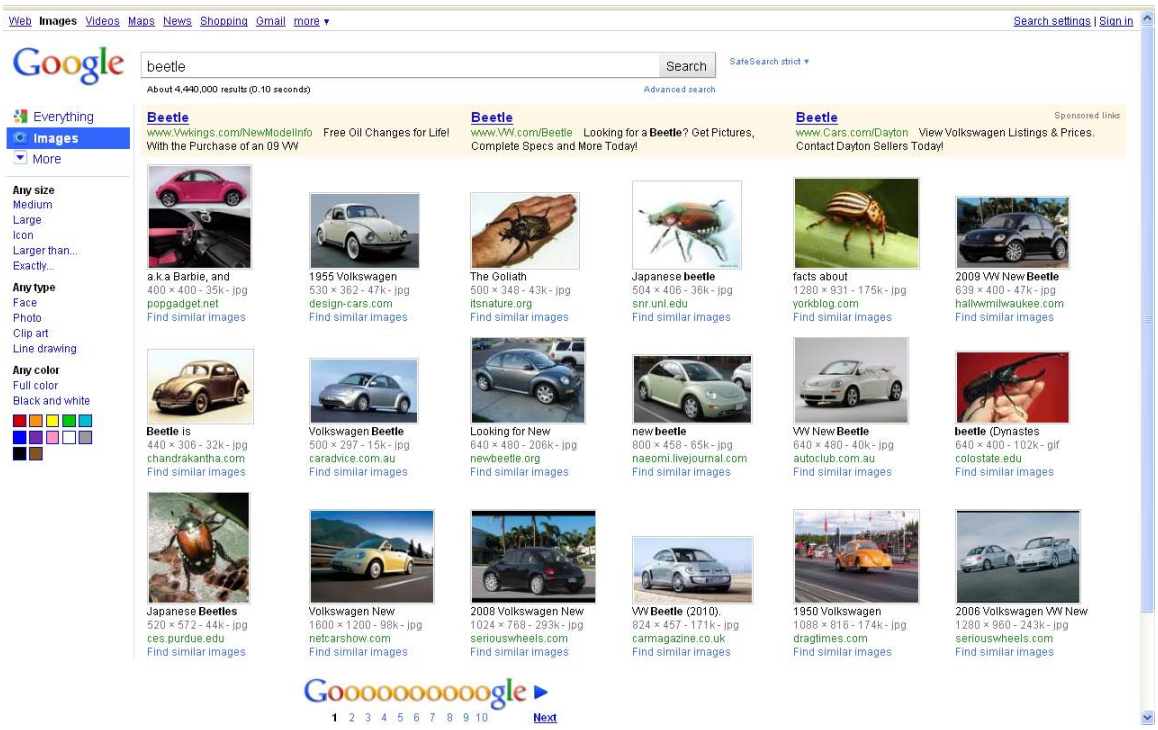


**Figure 9: Google Search Sidebar**

### *Post-Search Interview Protocol*

Using a semi-structured interview protocol, a series of questions was posed to the participant following the completion, successful or not, of the search task, in an effort to explore some of the underlying issues that may have contributed to the lack of use of negation in this experimental setting. The questions were designed around the notion of a spectrum on which search users could be arrayed with regard to their search sophistication and use of negation. This spectrum was assumed to range from users not knowing about negation, through users not finding it useful, to users not having complex enough information needs to merit its use, to users who regularly used negation.

During the follow-up interview, participants were given the opportunity to discuss their various search strategies and tactics and reflect on their attitudes and perceptions about the functionality of the Google Images website. They were also provided an introduction to the concept and use of negation as described in the Google website help files and were given an opportunity to express their knowledge, attitudes, and perceptions regarding negation prior to their participant in the study, during participation in the study, and in their future searches. Participants were also asked to provide non-identifying demographic information as well as self-reported assessments of their own search skills given the linkage suggested by Lucas and Topi (2002) and White and Morris (2007) between advanced search engine users and the correct use of Boolean operators. The interview script is included in Appendix A.

Two primary considerations were followed in the design of the interview. First, each participant was asked the same series of questions while skipping those that did not apply based upon their search task performance or previous answers. Secondly, each question was also carefully designed to guard against instilling a sense of task failure, particularly for participants who did not find the image. Though the search task was designed for the use of negation and failure to find the image was highly likely in the absence of the negation operator, each protocol question was framed to fault the search engine for not interpreting the participant's information need correctly. The following chapter describes the task-based performance measures and follow-up interview results that were obtained during this study.

# IV. Results

## Search Task Results

The aim of this study was to determine if participants would use negation when the presented search task was most readily accomplished through the use of negation. The data collected in this study showed that 5 of 30 participants (16.7%) did use negation. Three of the 5 participants (60%) who used negation located the image, while 10 of the 25 participants (40%) who did not use negation located the image. The two participants who used negation but did not locate the image included terms in their queries that were not associated with the image while at the same time including negated terms. Of the 5 participants who used negation, 3 had completed a Master's degree and the other two had completed only some graduate coursework. Due to the small number of participants, the value of negation in finding the image cannot be said to be statistically significant, however, the numbers presented here should in no way cause rejection of the hypothesis that negation is beneficial in some situations.

When negation was successfully employed to locate the image, one of the participants queried for *beetle -car* while the other two queried for *beetle -vw*. Of the participants that located the image without negation, three queried for *beetle* and paged to the sixth page of results. Another queried for *beetle* but found the image on the second page of results by applying advanced search criteria (e.g. the image is a full color picture in portrait orientation). Among the others who did not use negation but found the image, one queried for *beetle insect* with additional advanced search criteria, finding the image

on the fifth page of results; three queried for some permutation of *beetle insect wheat*, each finding the image on the first page of results; and another queried for *beetle wheat* and was rewarded on the first page of results. A final participant submitted the ill-formed query for *beetle,black, wheat* and discovered the image on the eighth page of results. None of the participants who used negation submitted a negation query immediately following the query for *beetle*, but first submitted at least one descriptive query such as *black beetle.*

Five major search strategies were self-reported during follow-up interviews or observed following an inspection of the individual search queries. Each of the 196 queries captured was identified according to one or more of these search tactics. The five are defined here:

1. Brute Force – Viewing more than three results pages for a single query.
2. Beetle Identification – Conducting a search specifically to identify the beetle. (e.g. *types of beetles*)
3. Image Content Description – Describing the content of the image. (e.g. *black beetle* or *beetle eating plant*)
4. Image Description – Describing how the image was captured or stored, either within the query or using the advanced search tool.
5. Negation – Negating one or more query terms.

None of these strategies were mutually exclusive, and some queries evidenced none of these strategies (e.g. *beetle* when not explored beyond the third page). Table 6 shows the number of queries that evidenced each described strategy. By far the most common strategy was to include additional content descriptors, with the brute force method a

distant second. Negation was the fourth most used strategy, identified in only 6.6% of the 196 queries.

**Table 6: Search Strategies**

| Search Strategy | Count |
|---|---|
| **Image Content Description** | 155 |
| **Brute Force** | 69 |
| **Image Description** | 23 |
| **Negation** | 13 |
| **Identification** | 5 |

The query log captured in this study shows a mean of 3.3 terms per query with a standard deviation of 1.6. On average, participants in this study viewed 4.2 pages for each query conducted. Another consideration in some query log research is the incorrect use of advanced search operators. In this study, none of the 13 queries containing a minus (-) sign showed evidence of incorrect implementation of negation.

**Follow Up Interview Results**

Table 7 shows the spectrum of reasons for lack of use of negation by the 25 participants did not use negation. While only 10 of those 25 participants did not know how to use negation, an additional 5 indicated no experience with negation, thus indicating that 15 of the 25 individuals who did not use negation were entirely unlikely to use negation no matter the search task. Many participants who indicated that they prefer

other search tactics over negation appeared to have a limited grasp of how a search engine works, often indicating a lack of understanding of the value of negation which was precipitated by a lack of understanding of search functionality. Particularly, in this case, they seemed to be unaware of how tagging of images occurs and the resulting limitations.

**Table 7: Spectrum of Reasons for Lack of Use of Negation**

| Group Characteristics | | | Count |
|---|---|---|---|
| **No previous knowledge of negation** | | | 10 |
| **Previous knowledge of negation** | **No previous use of negation** | **Prefer another method over negation** | 3 |
| | | **Did not think to use negation** | 2 |
| | **Previous use of negation** | **Did not think to use negation** | 2 |
| | | **Prefer another method over negation** | 6 |
| | | **Would have used negation if time permitted** | 2 |

Table 8 and Table 9 show how likely participants reported they were to use negation in the future on a five point scale, broken down by previous use and previous knowledge of negation. A two-tailed t distribution was used to measure the significance of the results with confidence levels indicated in parentheses throughout the text. Participants' previous use of negation was a significant predictor of how likely they reported using negation in the future ($p = 0.03$), with experienced participants rating likelihood an entire point higher than inexperienced participants. However, there was no

significant difference between groups with and without previous knowledge of negation ($p = 0.92$).

**Table 8: Likelihood of Using Negation in the Future by Previous Use**

| Previous Use | Count | Average Rating | Standard Deviation |
|---|---|---|---|
| **Yes** | 15 | 4.3 | 1.2 |
| **No** | 15 | 3.3 | 1.2 |
| **Overall** | 30 | 3.8 | 1.3 |

**Table 9: Likelihood of Using Negation in the Future by Previous Knowledge**

| Previous Knowledge | Count | Average Rating | Standard Deviation |
|---|---|---|---|
| **Yes** | 20 | 3.75 | 1.4 |
| **No** | 10 | 3.8 | 1.0 |
| **Overall** | 30 | 3.8 | 1.3 |

Table 10 shows no significant difference in self-reported Internet search skills between users who used negation during the search task and those who did not ($p = 0.93$). Furthermore, Table 11 shows no significant difference in self-reported Internet search skills and success in locating the image ($p = 0.49$).

**Table 10: Self-reported Internet Search Skills by Use of Negation**

| Negation Used | Count | Average | Standard Deviation |
|---|---|---|---|
| **Yes** | 5 | 3.6 | 0.55 |
| **No** | 25 | 3.6 | 1.0 |
| **Overall** | 30 | 3.6 | 0.93 |

**Table 11: Self-reported Internet Search Skills by Image Location Success**

| Image Found | Count | Average | Standard Deviation |
|---|---|---|---|
| **Yes** | 13 | 3.8 | .83 |
| **No** | 17 | 3.5 | 1.0 |
| **Overall** | 30 | 3.6 | 0.93 |

This study also provided a collection of information needs for which negation is a valuable tool. During the pre-test phase of this study, a number of individuals indicated that negation was valuable for them in their shopping on Craigslist and eBay, particularly for collectible items. In these cases, the individuals were interested in finding as many matches as possible without missing any potential matches. During the follow-on interviews for the actual study search tasks, the vast majority of study participants described situations like the search task in which a query term was ambiguous and caused an undesired class of results within the result set as prime opportunities for the use of negation. In addition, four participants indicated that negation would be valuable in their research or class work. The task of locating a vast array of information about oneself or another person on the Internet was mentioned by two participants as a prime opportunity for the use of negation. Only one study participant directly expressed hesitation about the usefulness of negation in any instance, but this participant also reported that negation was valuable to him/her for image searches and searches within Google Scholar.

# V. Conclusions and Recommendations

## Discussion

The search tasks metrics presented in the previous chapter tell a different story than those usually presented in analysis of query logs. In most studies reviewed previously (Spink et al., 2001; Herskovic et al., 2007; and others), negation was generally reported to be used in only about 0.1% of queries, but in this study, it was identified in 6.6% of the queries conducted likely due to the design of a search task that was benefited by the use of negation.

The increased number of page views reported in the study, compared to that reported in Spink et al. (2001), Jansen et al. (2005), and iProspect (2006) might be attributed to the participants' intrinsic understanding that any result returned could be the desired image and that there is little guarantee that it appear on the first page of results. Additionally, iProspect (2006) reports that a small portion of individuals would change search engines when they were unsuccessful in their searching. This method was forbidden during this study, potentially contributing to higher pages viewed per query.

In this case, the number of terms per query was higher than reported by Spink et al. (2002), Jansen et al. (2005), and others. The discrepancy between the metrics in this study and previous research may be accounted for by the single informational query, as defined by Croft et al. (2010), compared with the average query. Brenes (2009) reported an overwhelming majority of the queries submitted to AOL and recorded in the logs released in 2006 were single term queries. Thus, it is understandable that a single

complex information need might generate longer than average queries. Additionally, there was no evidence in this study corroborating the evidence from Spink et al. (2001) of incorrectly used negation operators.

Based on the post-task interview, the foremost reason for lack of use of negation was lack of experiential knowledge of negation. Lack of experiential knowledge also was shown to influence the participant's self-reported propensity to use negation in a future search. Previous use of negation was the most frequent predictor of self-reported Internet search skills, paralleling the methodological decisions by White and Morris (2007) and Lucas and Topi (2002) to classify users of advanced search operators as advanced users. The metrics concerning self-reported Internet search skills are different than anticipated. Potentially, other ways of measuring search skill would be more appropriate. For example, search skill might better be measured through successful completion of a series of tasks.

## Conclusions

Three significant conclusions arise from the results presented in the previous chapter and discussed in this chapter. First, in answer to the research question posed in this study, search engine users did, in fact, use negation at a rate of 16.7% when confronted with a search task or information need that warranted its use. Furthermore, most of the users who did not use negation in such a scenario simply did not know about it or had no personal experience using it. At the beginning of this report, a thought experiment was conducted to show that Andy's information need was best served with

negation. This study showed a 60% success rate for participants using negation compared to the 40% success rate of participants not using negation. Thus, information needs benefited by negation do in fact exist.

## Research Contributions

The most significant contribution of this research is in providing a foundation for future exploration of negation in search. While query logs have provided a beginning for negation research, existing public logs are relatively ancient with respect to the quickly changing landscape of Internet search. Additionally, query logs provide no user feedback, in contrast with this study in which search engine user's thoughts and motivations with regard to a series of queries are explored. Negation functionality, though often ignored in the literature because of its infrequent use, garners sizable usage when merited and is widely accepted as useful among those who are aware of its existence. On the basis of the use and usefulness of negation shown in this study, technologies being developed or improved must certainly include negation functionality as well as provide for improved user awareness of the functionality.

## Limitations

The primary limitations of the methodology employed by this study and restricting its widespread application are threefold. First, the study population was not necessarily representative of the general population. The educational level, age, and

gender of the study population selected from a military-sponsored graduate institution are not representative of the general Internet population. In 2003, only 27.2% of the US population over 25 had attained a Bachelor's Degree (Stoops, 2004). A study of a population of graduate students, who may possess more complex information needs and more established information seeking skills, does not necessarily translate to the entire Internet population. Thus, while the population studied for this research does not mimic the Internet population, research with a potentially more sophisticated group of searchers likely provided this study a higher density of individuals who had some prior knowledge of negation. Nevertheless, the obtained results might still be representative of other Internet users or user groups with complex information needs, to whom many advanced search systems and capabilities are targeted.

This study was, in at least a small way, limited by the selection of the image used in the search task. Specifically, the image was findable by methods other than negation. Ideally, an image would have been selected that could only have been found using negation so that no participant would complete the task before exhausting their patience, the time allotted, or their search skills. If that were the case, the 10 participants who found the image without negation would have continued searching, possibly using negation in the process. Thus, the incidence of negation use in this study may under represent the percentage of users who would ultimately or eventually use negation in the event of a search task demanding its use. Another confounding factor surrounding the image is that each participant knew of its existence within the search results. This may have caused unnatural search behavior such as immediately paging through the results beyond the number of pages that would normally be perused.

As with all studies, subjectivity and misinterpretations on the part of the participants and the study administrator may also have impacted the results. For example, it became clear during the study that the question, "How likely are you to use negation in a future search?" was misinterpreted by a number of participants. Specifically, the intent of the question was to elicit the likelihood that a participant would ever use negation in the future. Some participants interpreted the question as asking about the likelihood that they would use negation on any given search in the future. As negation is not valuable for every search, responses were likely lower than they would have been had the question been more carefully worded. Future studies should take careful note of the potential for these issues and mitigate through increased pre-testing of interview questions. In addition to the problem of question interpretation issues, interview responses may also be subject to misinterpretations of the answer. For instance, the researcher's knowledge of and personal investment in the subject area might have contributed to faulty assumptions about a participant's answers that were not entirely warranted or appropriate. To mitigate this concern, the interviewer carefully documented each participant response keeping in mind other answers and their performance on the search task as a context within which to interpret it.

**Recommendations for Future Research**

In addition to developing studies to combat the limitations mentioned above, the area of negation is ripe for exploration. Commensurate with the pilot nature of this study, broader research should be conducted, assessing search engine users in their

natural environments and without the use of contrived search tasks. Many methods for collecting such data exist including in the form of browser add-ons and proxy server software; however, tactics will be required to divine which queries represent information needs for which negation would have been useful. Perhaps a hybrid approach of collecting data naturally, but then following up directly with users would be an appropriate starting point. This natural data collection method may require a better understanding of what percentage of queries would benefit from negation and how they might be identified. During this study, participants possessed an abnormally high level of education. Future studies should investigate any correlation that may exist between education and use of negation or other advanced search operators. Future studies should collect data in sufficient quantity to allow all results to reach the level of statistical significance.

Another area for further research is in the area of negation education methods. Topi and Lucas (2005) have already found that training in Boolean logic affected information retrieval performance. Future research should also consider Venn diagrams as a method of education. Venn diagrams were presented in Chapter 1 as a means of communicating the value of negation. One of the study participants who chose to use a descriptive method, rather than negation, mentioned during the follow up interview that he perceives each search term as a set in a Venn diagram such that the addition of terms causes the results set to be the intersection of all the sets. This explanation was intended to be justification of his descriptive method. He did not, however, recognize that his argument did not support his method; specifically, that a desired result is easily left out of the result set by using a greater number of inclusion criteria for information search. After

the study concluded, this participant wished to discuss negation further. Upon additional explanations of the failings of the description method using the Venn diagram approach, the participant seemed to readily understand the value of negation. Further study into the usefulness of Venn diagrams as a teaching tool or as a search interface tool where a Venn diagram serves to represent query terms as in Jones (2003) would be appropriate. If negation is explored from this educational standpoint, follow up studies should be conducted to determine the lasting value of the education. In addition to Venn diagrams, any future education studies should further investigate and exploit the relationship between experiential knowledge of negation and use of negation.

Along the negation spectrum developed to represent the data gathered in this study, the most interesting group of participants for further study are those who simply chose to use search methods other than negation despite prior knowledge of the concept of negation. Comprising a large subset of the study group, with 7 out of the 25 participants who did not use negation, these individuals readily volunteered without any specific prompt that they prefer to use alternative methods to negation. Further investigation into this group of users who knew about negation, but chose not to use it, could revolutionize the way the community perceives the usefulness of negation or the education of users. Research into this lack of use of negation might best be conducted again in an interview format, but with a more open ended format allowing the expert researcher to dialogue with the participant. Participants could be asked to describe their understanding of how search engines work in order to uncover weaknesses in structural understanding which prohibit comprehension of negation. Studies could also be developed along the lines of the previous suggestion for negation education to determine

if there are issues regarding users' comfort with the use of negation. As reasons for lack of use are identified, further studies will mushroom from this proposed study.

The final recommendation for future research is a study investigating negation in faceted search and browsing. Faceted search is a method of isolating desired results based upon selection of listed metadata rather. In this study, free text search was explored; however, the advent of the Google options sidebar and the prevalence of faceted browsing in digital libraries and on shopping websites call for a similar investigation of negation an a faceted searching context. Studies in this area should address users' ability to comprehend an interface that allows for negation as negation does not currently exist in most faceted interfaces. Such a study should also investigate how users understand the faceted browsing functionality compared to the degree to which they understand how free text searching functions.

**Conclusion**

This study has answered the question of whether users will use negation when it is the most appropriate search strategy. Any of these areas for future research is a viable means of advancing the art of information retrieval and catering to the information seeking behavior of users. This study stands as a call for further research into the use and usefulness of negation.

## Appendix A: Materials for Study Administration

### Study Script

#### *Instructions for Study Administration*

Read only the words italicized and enclosed in square brackets to the subject. Words not italicized are intended as instructions for the researcher only.

#### *Informed Consent*

Present the volunteer with the informed consent documentation. *[Please take a few minutes to read this document.]* (Pause) *[Do you have any questions?]* (Pause) *[If you give your consent to participate in the study, please sign at the bottom.]* After a signature is obtained, provide the participant with a second copy of the informed consent documentation for his personal records. Sign the informed consent. Place the signed informed consent document in the folder marked 'Informed Consent'. Whether the individual chooses to continue with the study or not, ensure that Step 7 is completed to enter the individual in the drawing.

#### *Step 1: Introduce the Study*

*[In this study, we will be evaluating how Google's image search engine responds to user input.]*

***Step 2: Present Participants with the Scenario***

Provide the participant with the image of the beetle (Figure 7). *[Your child is writing a research paper. He indicated to you that he found this image of a beetle using Google's Image search, but forgot to write down citation information. He needs your help in locating exactly this image.]*

***Step 3: Instruct the Participant on Boundaries***

Present participant with the computer. *[In this scenario, you have begun with this initial search for the term 'beetle' as shown on the computer screen. When you begin, you may browse your search results or modify your search terms within the Google images website in the way you normally would. However, please avoid clicking on any of the images as you browse the results as this will take you away from the Google images website itself. If Google images successfully returns the desired image for you, click on it and notify me that you are finished. Please take your time as you try to find the beetle. If at any point you determine that you have tried all possible methods to get Google Images to return the image of the beetle, you may stop and let me know that you are done searching.]*

***Step 4: Instruct the Participant on How to Begin***

*[Your search history will be recorded automatically to enable further analysis of this search session. When it is time to begin, please press the F8 button to start the study. Do*

*you understand all of the directions?]* (Pause) *[Do you have any questions?]* (Pause) *[You may begin by pressing F8.]* Observe time on a watch or clock with a second hand so that you can anticipate when the participant's 5 minutes are concluded. Prior to beginning this waiting period you should relocate to the alternative seat in the room.

***Step 5: Move to the Interview***

*[Thank you for your effort. Please give me a minute to review how well Google Images responded to your input.]* Look at the history for the browser and determine if the participant used negation or not. Answer questions 1, 3, and 7 in the interview document. Note that you as the interviewer should note any commentary the participant provides that would color the given answer. For example, if for the question "Have you ever used negation in search?" the participant answers "Yes, but I didn't realize that you couldn't have a space between the minus sign and the word," you should note this important caveat.

***Step 6: Interview***

Complete a copy of the interview document. Words to be read to the participant are italicized and enclosed in square brackets as in these instructions.

***Step 6: Debriefing***

*[The aim of this study was to explore how users interact with Internet search engines. Search engines are designed to meet the needs of their users, so understanding users and their needs are of supreme importance. In this particular instance, we were interested in a particular facet of Internet search known as negation and how negation might be used in Internet search engines to provide users with better search results that more closely meet their needs.]*

***Step 7: Participation in Amazon Gift Card Drawing***

*[Are you willing to stay for a few more minutes to enter into the drawing for a $100 Amazon gift card?]* If yes, provide the participant with a single copy (or two copies, if the participant was referred) of study flyer to write their name and contact information on. Provide the participant with a few flyers to refer friends and inform the participant that they must indicate their name on these fliers and that their friends must bring the flyer and participate in the study in order for the referrer to receive an additional entry. Collect a flyer from the participant if they were referred by another individual.

**Informed Consent**

Participant # _____

INFORMATION PROTECTED BY THE PRIVACY ACT OF 1974

**Informed Consent Document**

**For**

**Search Engine Response to User Search Methods**

Air Force Institute of Technology (AFIT)/ENV, AFIT, Wright-Patterson Air Force Base, OH

Principal Investigator: Lt Col Jason M Turner, DSN Rank/Name, DSN
785-3636x7407, AFIT/ENV
jason.turner@afit.edu

Associate Investigator: Mrs. Kristen M Lancaster, 937-608-1385, AFIT/ENG
kristen.lancaster.ctr@afit.edu

1. **Nature and purpose:** You have been offered the opportunity to participate in the "Search Engine Response to User Search Methods" research study. Your participation will occur at Einstein Bros Bagel store at the Air Force Institute of Technology (AFIT).

   The purpose of this research is to evaluate search engine response to tactics employed by users of web search engines.

The time requirement for each volunteer subject is anticipated to be a total of 1 visit of approximately twenty minutes. A total of approximately 30 subjects will be enrolled in this study. Only volunteers 18 years of age or older will be allowed to participate in the study.

2. **Experimental procedures:** If you decide to participate you will work on one search task and then be asked a series of questions. The entirety of the experiment will last approximately 20 minutes. First, the task will be explained to you. During the task you will use a laptop computer and mouse to perform a search task. The search engine behavior and results returned in response to your user inputs will be recorded for future review. During the interview, the researcher will ask you a series of 18 or less questions, many of which can be answered with a yes or a no. During the task and interview, you will not be judged at any time on correctness or completion. Because of the short duration of the task, no breaks or food or drink will be provided or allowed.

3. **Discomfort and risks:** Discomforts may consist of the possibility of slight physical discomfort due to the use of a computer and mouse and remaining in a seated position for 20 minutes. If you have any concerns now or later regarding the conduct of this study, you may, without penalty, forgo participation. No known risks exist.

4. **Precautions for female subjects or subjects who are or may become pregnant during the course of this study:** There are no special cautions required for female subjects.

5. **Benefits:** You are not expected to benefit directly from participation in this research study.

6. **Compensation:** Participation in this study is voluntary and no compensation will necessarily be awarded, including costs related to transportation to and from the research site. Even if you decline to give consent to participate in this study, you have at this point qualified for entry into a random drawing for a $100 Amazon gift card. If you do choose to participate in the study, referral cards will be provided to you by the study administrator. These cards may be given to other individuals who you believe might be interested in participating in the study; if they return the referral card and participate, you and they will receive an additional entry opportunity for the drawing. Chances of winning are dependent upon the number of total participants and referrals. Information that you release as a part of this drawing for the purposes of contacting you in the event that you are selected for the gift card award will not be connected to the data collected from your participation in the study, and all records of such contact information will be destroyed after a winner has been chosen.

7. **Alternatives**: Choosing not to participate is an alternative to volunteering for this study.

8. **Entitlements and confidentiality:**
    a. Records of your participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C. 552a, and its

implementing regulations and the Health Insurance Portability and Accountability Act (HIPAA), and its implementing regulations, when applicable, and the Freedom of Information Act, 5 U.S.C. Sec 522, and its implementing regulations when applicable. Your personal information will be stored in a locked cabinet in an office that is locked when not occupied. Electronic files containing your personal information will be password protected and stored only on a secure server. It is intended that the only people having access to your information will be the researchers named above the Air Force Surgeon General's Research Compliance office, the Director of Defense Research and Engineering office or any other IRB involved in the review and approval of this protocol. When no longer needed for research purposes your information will be destroyed in a secure manner (shredding). Complete confidentiality cannot be promised, in particular for military personnel, whose health or fitness for duty information may be required to be reported to appropriate medical or command authorities. If such information is to be reported, you will be informed of what is being reported and the reason for the report.

b. Your entitlements to medical and dental care and/or compensation in the event of injury are governed by federal laws and regulations, and that if you desire further information you may contact the base legal office (ASC/JA, 257-6142 for Wright-Patterson AFB).

c. If an unanticipated event (medical misadventure) occurs during your participation in this study, you will be informed. If you are not competent at the time to

understand the nature of the event, such information will be brought to the attention of your next of kin or other listed emergency contact.

Next of kin or emergency contact information:

Name_____ Phone#_____

The decision to participate in this research is completely voluntary on your part. No one may coerce or intimidate you into participating in this program. You are participating because you want to. Lt Col Jason M Turner or Mrs. Kristen M Lancaster, or an associate, has adequately answered any and all questions you have about this study, your participation, and the procedures involved. Mrs. Kristen M Lancaster can be reached at (937) 608-1385. Mrs. Kristen M Lancaster or an associate will be available to answer any questions concerning procedures throughout this study. If significant new findings develop during the course of this research, which may relate to your decision to continue participation, you will be informed. You may withdraw this consent at any time and discontinue further participation in this study without prejudice to your entitlements. The investigator or medical monitor of this study may terminate your participation in this study if she or he feels this to be in your best interest. If you have any questions or concerns about your participation in this study or your rights as a research subject, please contact Lt Col Michael Richards at (937) 904-8100 or michael.richards@wpafb.af.mil.

"Your participation in this study will not be photographed, filmed or audio/videotaped."

YOU ARE MAKING A DECISION WHETHER OR NOT TO PARTICIPATE. YOUR SIGNATURE INDICATES THAT YOU HAVE DECIDED TO PARTICIPATE HAVING READ THE INFORMATION PROVIDED ABOVE.

**Volunteer Signature_____Date_____**

**Volunteer Name (printed)_____**

**Advising Investigator Signature _____ Date _____**

**Investigator Name (printed)_____**

**Witness Signature _____Date _____**

**Witness Name (printed)_____**

**Privacy Act Statement**

**Authority**: We are requesting disclosure of personal information. Researchers are authorized to collect personal information on research subjects under The Privacy Act-5 USC 552a, 10 USC 55, 10 USC 8013, 32 CFR 219, 45 CFR Part 46, and EO 9397, November 1943.

**Purpose**: It is possible that latent risks or injuries inherent in this experiment will not be discovered until some time in the future. The purpose of collecting this information is to aid researchers in locating you at a future date if further disclosures are appropriate.

**Routine Uses**: Information may be furnished to Federal, State and local agencies for any uses published by the Air Force in the Federal Register, 52 FR 16431, to include, furtherance of the research involved with this study and to provide medical care.

**Disclosure**: Disclosure of the requested information is voluntary. No adverse action whatsoever will be taken against you, and no privilege will be denied you based on the fact you do not disclose this information. However, your participation in this study may be impacted by a refusal to provide this information.

**Negation Explanation from Google's Advanced Search Tips Page**

The following is an excerpt from the Google advanced search tips page on a technique often called **negation**.

"Attaching a minus sign immediately before a word indicates that you do not want pages that contain this word to appear in your results. The minus sign should appear immediately before the word and should be preceded with a space. For example, in the query `anti-virus software`, the minus sign is used as a hyphen and will not be interpreted as an exclusion symbol; whereas the query `anti-virus -software` will search for the words 'anti-virus' but exclude references to software. You can exclude as many words as you want by using the **-** sign in front of all of them, for example `jaguar -cars -football -os` " (Google, 2010) This query will return pages with the word jaguar, but without the words car, football, and OS so that most pages about the Jaguar car brand, the Jaguars football team and the Jaguar operating system are excluded from the search results.

**Interview Script**

Participant # _____

Follow each of these number steps. Read only the italicized words to the participant. Embolden or write the subject's response. Follow any instructions associated with a response about skipping to another question.

1. *[Please describe your search strategy or strategies.]*
2. Did the participant work until the end of the 5 minutes?
    a. Yes.
    b. No. Go to Step 5.
3. *[Are there any search tactics that you did not have a chance to pursue? If so, could you briefly describe them?]*
4. *["If you were in charge," is there any specific search engine functionality or feature that you would propose to Google that you think would help to improve the results or your experience using this search engine?]*
5. Provide the participant with Supplement 2. *[Please read this excerpt from Google's advanced search tips page. Let me know when you are finished.]* (pause) *[Is this explanation of negation clear to you? Yes or No.]*
    a. Yes.
    b. No. Go to Step 13.
6. Did the participant use negation? This requires a judgment call on the part of the interviewer and should not be based upon whether or not a minus sign was used, but

on the overall context. For example, a query for *beetle -black* would not be considered a use of negation. Circle one.

a. Yes. Go to Step 11.

b. No.

7. *[Some types of search engines allow the use of the word NOT, typed in all capital letters, in the same way that Google uses the minus sign. Did you know how to use negation before the explanation I just provided to you? Yes or No.]*

a. Yes.

b. No. Go to Step 11.

8. *[When confronted with the original search for 'beetle' that had many Volkswagen Beetles in the results, why did you choose not to use negation?]*

9. *[Do you think that negation would have been useful in finding the beetle? Please use a scale of 1 to 5 where 1 is not helpful at all and 5 is very helpful.]* _____

10. *[Have you ever used negation in a search? Yes or no.]*

a. Yes.

b. No.

11. *[How likely are you to use negation in a future search? Please use a scale of 1 to 5 where 1 is very unlikely and 5 is very likely.]* _____

12. *[In what types of searches do you find negation to be a useful tool or do you think it might be a useful tool?]*

13. *[How would you rate your Internet search skills? Please use a scale of 1 to 5 where 1 is poor and 5 is outstanding.]* _____

14. *[I now have a few demographic questions for you. As specified in your informed consent, this information will only be reported in a way that does not identify you personally. You may decline to answer any of the following questions without affecting your participation in the study or your eligibility for the gift card drawing. What is your age? You may choose to decline to provide your age.]*

    a. Prefer not to say.

    b. Age: ____

15. *[What is your gender? Again, you may choose to decline to provide this information.]*

    a. Prefer not to say.

    b. Male.

    c. Female.

16. *[What is the highest level of education that you have attained?]* You may prompt the participant by reading the answer choices if necessary.

    a. No high school diploma or GED.

    b. High school diploma/GED.

    c. Some college.

    d. Associate's degree.

    e. Bachelor's degree.

    f. Some graduate coursework.

    g. Master's degree.

    h. Ph.D.

**17.** *[Thank you for your participation in this study.]*

## Appendix B: Human Subjects Exemption Approvals

May 17, 2010

Lt Col Jason M Turner,

I have reviewed your study entitled " Use of Negation in Web Search" and found that your study qualifies for an IRB exemption.

Per 32 CFR 219.101 (b)(2), Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation is exempt.

Your study qualifies for this exemption because the demographic data you are collecting cannot realistically be expected to map a given response to a specific subject, and the questions you are asking could not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation. Finally, while you are collecting names, this is a required and natural consequence of your selected data collection methodology. These names will be protected at all times, only be known to the researchers, and managed according to the AFIT interview protocol.

This determination pertains only to the Federal, DoD, and Air Force regulations that govern the use of human subjects in research. It does not constitute final approval to conduct the study which should be granted by you research advisor. Further, if a subject's future response reasonably places them at risk of criminal or civil liability or is damaging to their financial standing, employability, or reputation, you are required to file an adverse event report with this office immediately.

WILLIAM A. CUNNINGHAM, PhD
AFIT IRB Research Reviewer

DEPARTMENT OF THE ARMY
US ARMY MEDICAL RESEARCH AND MATERIEL COMMAND
504 SCOTT STREET
FORT DETRICK, MD 21702-5012

REPLY TO
ATTENTION OF

MCMR-RP

19 May 2010

MEMORANDUM FOR THE RECORD

SUBJECT: Determination for U.S. Army Aeromedical Research Laboratory (USAARL) Personnel Participation in the Protocol, "Use of Negation in Web Search," Principal Investigator: Lt Col Jason M. Turner, Air Force Institute of Technology (AFIT), Wright-Patterson Air Force Base, Ohio; USAARL Associate Investigator: Kristen M. Lancaster, USAARL, Fort Rucker, Alabama, USAARL Study Number 2010-013, Log Number A-16185

1. The subject project documentation received 10 May 2010 by the U.S. Army Medical Research and Materiel Command's Office of Research Protections, Institutional Review Board Office (ORP IRBO) has been reviewed for applicability of human subjects protection regulations.

2. The research involves asking up to 30 adult volunteers to conduct a computer search for an image via a search engine. The investigators will then ask the volunteers pre-determined questions about the strategy used to conduct the search. Volunteers' identifiable information will not be collected on survey instruments.

3. In accordance with 32 CFR 219.101(b)(2), the AFIT Institutional Review Board (IRB) determined that this protocol is exempt as research involving the use of educational tests and survey procedures with adult subjects. The IRBO concurs with the determination made by the AFIT IRB. The protocol may proceed with no further requirement for review by the ORP IRBO, pending concurrence of the USAARL Commander. The ORP IRBO protocol file will be closed.

4. In the event that there is a change to the protocol that may affect its exemption status, a description of the change must be sent to the ORP IRBO at IRBOFFICE@amedd.army.mil referencing the Protocol Log Number listed in the "Subject" line above. The ORP IRBO will re-open the file if necessary.

5. The point of contact for this action is Dr. Sarah L. Donahue, at 301-619-1118 or Sarah.L.Donahue@us.army.mil.

ANDREA J. KLINE, MS, CIP
Director, Institutional Review Board Office
Office of Research Protections

DEPARTMENT OF THE ARMY
US ARMY AEROMEDICAL RESEARCH LABORATORY
6901 FARRELL ROAD
FORT RUCKER, AL 36362-5577

REPLY TO
ATTENTION OF:

MCMR-UAC

18 May 2010

MEMORANDUM FOR DIRECTOR, Warfighter Protection Division

SUBJECT: Approval of USAARL Study #2010-013, HRPO Log Number A-15185 "Use of Negation in Web Search " Principal Investigator: Lt Col Jason M. Turner, Air Force Institute of Technology (AFIT), Wright-Patterson Air Force Base, Ohio; USAARL Associate Investigator Kristen M. Lancaster, USAARL, Fort Rucker Alabama. POC for this study is Dr. Carol Chancey.

Subject is approved. You may proceed with data collection.

JOSEPH F. MCKEON
COL, MC
Commanding

CF:
SPD
Chair, SRC
Regulatory Compliance Officer

## Bibliography


Aula, A. (2003). *Query formulation in web information search.* In Isaias, P. & Karmakar, N. (Eds.) Proc. IADIS International Conference WWW/Internet 2003, Volume I, 403--410. IADIS Press.

Belkin, N. J., Kelly, D., Kim, G., Kim, J., Lee, H., Muresan, G., et al. (2003). *Query length in interactive information retrieval*. Toronto, Canada: ACM.

Bottle, R. T. (2003). Information Science. In *International encyclopedia of information and library science* (pp. 295-297). 2nd ed. New York: Routledge.

Brenes D.J., & Gayo-Avello D. (2009). Stratified analysis of AOL query log. *Information Sciences.* 179 (12), 1844-1858.

Bush, V. (1945, July). As we may think. *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/.

Central Intelligence Agency. (2010, April 22). *CIA - The world factbook -- United States.* Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/us.html.

Chau, M., Fang, X., & Liu Sheng, O. R. (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology : JASIST.* 56 (13), 1363.

Clark, H. H. (1976). *Semantics and comprehension*. Janua linguarum, 187. The Hague: Mouton.

comScore, Inc. (2006). *comScore Releases September U.S. Search Engine Rankings.* Retrieved from http://ir.comscore.com/releasedetail.cfm?ReleaseID=245968

comScore, Inc. (2009). *comScore Releases November 2009 U.S. Search Engine Rankings.* Retrieved from http://www.comscore.com/Press_Events/Press_Releases/2009/12/comScore_Releases_November_2009_U.S._Search_Engine_Rankings.

comScore, Inc. (2010). *comScore Releases February 2010 U.S. Search Engine Rankings.* Retrieved from http://www.comscore.com/Press_Events/Press_Releases/2010/3/comScore_Releases_February_2010_U.S._Search_Engine_Rankings.

Cooper, W. (1988). Getting beyond Boole. *Information Processing & Management,* 24(3), 243-248.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Boston: Addison-Wesley.

data. (n.d.). *The American Heritage® Dictionary of the English Language, Fourth Edition*. Retrieved April 21, 2010, from Dictionary.com website: http://dictionary.reference.com/browse/data

Eastman, O. M., & Jansen, B. J. (2003). Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Results. *ACM Transactions on Information Systems : a Publication of the Association for Computing Machinery.* 21 (4), 383.

Göker, A. & Davies, J. (2009). *Information retrieval: Searching in the 21st century.* Chichester UK: Wiley.

Google. (2010). *More Search Help : Google search basics.* Retrieved from http://www.google.com/support/websearch/bin/answer.py?answer=136861.

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's web use skills. *Journal of the American Society for Information Science and Technology.* 53 (14), 1239.

Hepworth, M. & Murray, I. (2003). Search Engines. In *International encyclopedia of information and library science* (pp. 569-572). 2nd ed. New York: Routledge.

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *Journal of the American Medical Informatics Association.* 14 (2), 212.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks.* 33 (1), 337. Information (2003). In *International encyclopedia of information and library science* (p. 244). 2nd ed. New York: Routledge.

information. (n.d.). *The American Heritage® Dictionary of the English Language, Fourth Edition*. Retrieved February 17, 2010, from Dictionary.com website: http://dictionary.reference.com/browse/information

Information (2003). In *International encyclopedia of information and library science* (p. 244). 2nd ed. New York: Routledge.

Information Science. (n.d.). *The American Heritage® Dictionary of the English Language, Fourth Edition*. Retrieved April 08, 2010, from Dictionary.com website: http://dictionary.reference.com/browse/information science

iProspect.com Inc. (April 2006). *iProspect search engine user behavior study*. Retrieved from http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management.* 42 (1), 248.

Jansen, B. J., Spink, A., & Koshman, S. (2007). Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology.* 58 (5), 744.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology.* 56 (6), 559.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management.* 36 (2), 207-27.

Järvelin, K. (2003). Information Retrieval. In *International encyclopedia of information and library science* (pp. 293-295). 2nd ed. New York: Routledge.

Jones, S. (2003). VQuery: A graphical user interface for boolean query specification and dynamic result preview. *International Journal on Digital Libraries*, *2*, 207-223.

Klein, S. (2009). On the use of negation in Boolean IR queries. *Information processing and management.* 45 (2), 298-311.

Knowledge (2003). In *International encyclopedia of information and library science* (p. 341). 2nd ed. New York: Routledge.

Krásenský, P. (n.d.). *Cereal Beetle*. Retrieved from http://www.naturfoto-cz.de/photos/krasensky/cereal-beetle-1800.jpg.

Kuhlthau, C. C. (1991). Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science.* 42 (5), 361-71.

Lau, E. P., & Goh, D. H. L. (2006). In search of query patterns: A case study of a university OPAC. *Information Processing & Management.* 42 (5), 1316.

Levy, Steven. (February 2010), *Exclusive: how google's algorithm rules the web*. Retrieved from http://www.wired.com/magazine/2010/02/ff_google_algorithm/all/1.

Lucas, W., & Topi, H. (2002). Form and Function: The Impact of Query Term and Operator Usage on Web Search Results. *Journal of the American Society for Information Science and Technology.* 53, 95-108.

Markey, K. (2007). Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology.* 58 (8), 1071.

Mischo, W. H., & Lee, J. (1987). End-user searching of bibliographic databases. *Annual Review of Information Science and Technology, 22,* 227-263.

Morville, P. (2005) *Ambient Findability*. Sebastopol, CA: O'Reilly Media, Inc.

Scheffler, F. & March, J. (1972). An Experiment to Study the Use of Boolean Not Logic to Improve the Precision of Selective Dissemination of Information. *Journal of the American Society for Information Science.* 23 (1), 58-65.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-Sex to E-Commerce: Web Search Changes. *Web Technologies.* 35, 107-109.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology.* 52, 226-234.

Stoops, N. (2004). *Educational Attainment in the United States: 2003.* Retrieved from http://www.census.gov/prod/2004pubs/p20-550.pdf.

Sturges, P. (2003). Librarian. In *International encyclopedia of information and library science* (pp. 370-371). 2nd ed. New York: Routledge.

Sullivan, D. (1998). *Ratings of most visited search engines*. Retrieved from http://web.archive.org/web/20020404013943/http://searchengineguide.org/classi2.htm.

Sullivan, D. (2001, Dec 10). *Nielsen//NetRatings Search Engine Ratings*. Retrieved from http://web.archive.org/web/20011214200843/http://searchenginewatch.com/reports/netratings.html.

Taylor, R. S. (1962). The Process of Asking Questions. *American Documentation, 13*(4), 391-396.

Topi, H., & Lucas, W. (2005). Mix and match: combining terms and operators for successful Web searches. *Information Processing & Management.* 41 (4), 801.

U.S. Census Bureau (February 2010). *Monthly national population estimates*. Retrieved from http://www.census.gov/popest/national/NA-EST2009-01.html.

Wersig, G. (1971). *Information, kommunikation, dokumentation: ein beitrag zur orientierung der informations- und dokumentationswissenschaften*. München-Pullach: Verlag Dokumentation.

Wersig, G. (2003). Information Theory. In *International encyclopedia of information and library science* (p. 310). 2nd ed. New York: Routledge.

White, R. W., & Drucker, S. M. (2007). *Investigating behavioral variability in web search*. Banff, Alberta, Canada: ACM.

White, R. W., & Morris, D. (2007). *Investigating the querying and browsing behavior of advanced search engine users*. Amsterdam, The Netherlands: ACM.

## Vita

Mrs. Kristen M. Lancaster graduated from Enloe High School in Raleigh, North Carolina in 2002. She then went on to obtain a Bachelor of Science degree in Mathematics with minors in Physics and Computer Science from Cedarville University in Cedarville, Ohio in 2006. Upon graduation, she married her fellow firefighter, Aaron D. Lancaster. Throughout this period, she completed internships at IBM, Sandia and Oak Ridge National Laboratories, and The Reynolds and Reynolds Company. In the fall of 2006, she began studying for a Master of Science degree in Cyber Operations in the Graduate School of Engineering and Management at the Air Force Institute of Technology (AFIT). In the spring of 2007, she put her thesis and final class on hold to follow her husband, a newly minted Warrant Officer in the United States Army, to flight school at Fort Rucker, Alabama. While there she worked for a government contractor preparing software specifications for a project at the Combat Readiness/Safety Center. In 2009, she began providing research computer systems support as a contractor for the United States Army Aeromedical Research Lab (USAARL), also at Fort Rucker, Alabama. On sponsorship from the USAARL, she returned to AFIT to complete this thesis in order to further the information management interests of the Biodynamics Data Resource. Upon graduation, she will continue her work with the USAARL.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From – To)* |
|---|---|---|
| 17-06-2010 | Master's Thesis | Oct 2009 – May 2010 |

**4. TITLE AND SUBTITLE**
Use of Negation in Search

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Lancaster, Kristen M

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)**

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

**8. PERFORMING ORGANIZATION REPORT NUMBER**
AFIT/GCO/ENV/10-J01

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Dr. V Carol Chancey
US Army Aeromedical Research Laboratory
6901 Farrel Rd.
Fort Rucker, AL 36362
334-255-6952

**10. SPONSOR/MONITOR'S ACRONYM(S)**
USAARL

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Boolean algebra was developed in the 1840s. Since that time, negation, one of the three basic concepts in Boolean algebra, has influenced the fields of information science and information retrieval, particularly in the modern computer era. In Web search engines, one of the present manifestations of information retrieval, little use is being made of this functionality and so little attention is given to it in the literature. This study aims to bolster the understanding of the use and usefulness of negation. Specifically, an Internet search task was developed for which negation was the most appropriate search strategy. This search task was performed by 30 individuals and followed by an interview designed to elicit more information about the participants' use or non-use of negation during the task. Negation was observed to be used by approximately 17% of users in the study, suggesting that negation may indeed be infrequently used by Internet users. The data obtained during the post-task interview indicate that lack of use of negation stems from users not knowing about negation, having little experience with negation, or simply preferring other methods, even when negation is one of the foremost appropriate methods.

**15. SUBJECT TERMS**
Negation, Information Retrieval, Search Engine, Query Reformulation, Exclusion Criteria, Information-Seeking Behavior

| **16. SECURITY CLASSIFICATION OF:** Unclassified | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | **19b. TELEPHONE NUMBER** *(Include area code)* |
| UU | UU | UU | UU | 101 | Lt Col Jason M Turner, AFIT/ENV<br>937-255-3636x7407 |

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std. Z39-18

*Form Approved*
*OMB No. 074-0188*