

Review of  
*Amanda Spink*  
and her work in  
*Web Searching and Retrieval,*  
*1997 - 2004*

Larry Reeve  
for Dr. McCain  
INFO861, Winter 2004  
Term Project

## Table of Contents

Background of Spink.....	2
Web Search and Retrieval.....	3
Studies Conducted.....	4
Relevance feedback.....	6
Advanced search techniques.....	8
Term Characteristics.....	9
Comparison of American and European Web Search and Retrieval.....	11
Conclusion and Future Work.....	12
References.....	15

### Background of Spink

Amanda Spink received her Ph.D. in 1993 from the School of Communication, Information and Library Studies at Rutgers University, New Jersey. Her dissertation title was *Feedback in Information Retrieval*. Prior to receiving her Ph.D., Spink completed an undergraduate degree in History and Political Science from Australian National University, followed by a graduate degree in Information Systems Management and Marketing from Fordham University. Spink is currently an associate professor at the University of Pittsburgh in the School of Information Sciences. Prior to her appointment there, she held faculty positions at Pennsylvania State University (School of Information Science and Technology, Web Research Group), and the University of North Texas (School of Library and Information Sciences) (Spink, 2004).

Spink worked under Tefko Saracevic while at Rutgers, and continues to publish with him. Her earliest work with Tefko precedes her Ph.D. dissertation, and additional work in related areas continues until well after her dissertation. She published with Saracevic as recently as 2002, in an article about web search changes over time. Saracevic's work seems primarily concerned with Information Seeking and Retrieving, as well as with issues of Relevance (Saracevic, 2004). Saracevic's work in information retrieval and its influence on Spink can be seen by reviewing the publication areas that Spink tends to write in. According to her publication list, Amanda categorizes her main research areas as Information Retrieval (IR), Web Retrieval, Human Information Behavior/Information Seeking, and Medical Informatics. The focus of this paper is primarily her work in Web Retrieval, although elements of IR and Information Seeking also (and obviously) cross over into web retrieval.

Spink is rated 5<sup>th</sup> in journal article production and 18<sup>th</sup> in citation production among U.S. Information Science faculty (Budd, 2000). Spink has also been listed as among the most highly-cited authors in information retrieval research (Ding et al., 1998). According to the Institute for Information Science, her joint paper on web user query analysis (Jansen et al., 2000) is the most highly-cited article in web searching behavior (Spink, 2004). The paper is included here as a stream of her research in web searching behavior and query analysis.

### Web Search and Retrieval

Analyzing user queries as well as studying how users search is important for building future information retrieval systems. One area user query failure analysis looks at is user query construction. Studying how users develop queries can tell designers where their systems may be improved. For example, if users typically do not use advanced query techniques that are made available, then some type of automated system may be provided to help the users construct queries that are more likely to satisfy an information need. A corresponding aspect of query construction is determining how users initially develop a query and then refine it in the course of their information seeking process. Users may add search terms, remove them, change terms while keeping the count the same, or may become frustrated and give up their search. Since entering search terms into a form in order to produce a document set is the dominant interface in general purpose search engines, it is useful for system designers to understand what users are actually doing with real systems, and possibly determine what features of systems users find helpful and what features are hindering search effectiveness. The features Spink is looking to enhance relate to search terms, rather than other sorts of interface enhancements, such as visualization displays or personalization. An example of a search term enhancement is an automated agent that could suggest search terms that may help the user find relevant documents.

Spink's work in web retrieval examines user query construction. In particular, she looks at search term effectiveness, relevance feedback in query construction, search terms as topic indicators, and search term characteristics, such as lengths and co-occurrences. Search terms are important for Spink's work because they "are a central determinant in IR, and selection of search terminology is a driving force and variable in IR processes" (Spink, 1997). The web retrieval Spink discusses is primarily involved with users-at-large, as contrasted to mediated searchers. Users-at-large are people in the general population using general-purpose search engines, primarily on the Web. Mediated searchers are those using traditional IR mediated search techniques on specific databases. The distinction between the two types of information seeking groups is important. Systems designed for users-at-large do not expect users to have a well-developed set of skills for performing searches. In contrast, mediated searchers take advantage of specialists with skills in searching particular sets of databases. These professional mediators are more likely to have the exploitative control that

Patrick Wilson writes about. The mediators are able to suggest search terms that are more likely to return relevant documents, because they have experience with and understand the organization of the databases they are searching. Spink's work shows that the query construction of each type of information seeking group is different. Systems designed for each group must consider the searching skills capability of members of each of the groups.

Another aspect of web searching and retrieval related to search term selection is the use of relevance feedback. Relevance feedback is a beneficial feature because it allows users to discover potentially relevant documents without requiring users to modify their existing search terms. Relevance feedback is well-known and widely used within traditional IR systems (Spink, 1999). Relevance feedback is available on most general-purpose search engines, usually by clicking on a hyperlink that is near one of the retrieved documents. For example, the Google search engine implements relevance feedback with a hyperlink entry that reads "Similar pages." However, the use of relevance feedback during Web search sessions is not well understood. Although the feature exists on most Web search engines, its use by users needs further examination to understand how users utilize the feature. Spink also conducts research in relevance feedback, both in mediated searching and in Web searching. Her focus is primarily feedback relevance in Web searching, since a substantial amount of research in feedback relevance has already been done using traditional IR systems.

#### Studies Conducted

Spink begins her research into web searching by first examining selection of search terms when using mediated IR searching. The study from which this research is drawn looks at users trying to answer questions by using the DIALOG system and the use of professional intermediaries. The interaction between user, intermediary, and system is the focal point of her study. Spink raises the question, why bother to research the use of intermediaries in IR searches, when the usage of IR by users who are not taking advantage of intermediaries is exploding? Her response is that we do not yet understand the interactive processes of search and their complexity (Spink. 1997). The primary objectives of the mediated search study are to determine sources and effectiveness of search terms, as well as the process of selecting search terms. The process of selecting search terms is not well understood, and so the construction of search interfaces is hampered. Spink is looking to

understand the interactive process of IR, as opposed to the algorithmic process. In the algorithmic process, systems are considered static, as opposed to an interactive process, which is considered dynamic. Algorithmic research has been conducted for over thirty years, and is exemplified by studies such as TREC (Spink, 1997).

After reviewing and gaining insight into mediated searching, Spink then begins to look at how users search the Web. Her primary source of data for studying user searchers are transaction logs from the Excite search engine. Excite is a major Internet search engine, and according to its web site ([www.excite.com](http://www.excite.com)) claims to have 14 million unique visitors each month. There are two studies that look at Excite data. The first is the "51K study" (Spink, 2001), which consists of transaction data for 51,473 queries from 18,113 users. The transaction data was collected on March 9, 1997. A second study, the "1M study", was also done using Excite data, and consists of transaction data for 1,025,910 queries from 211,063 users on September 16, 1997. The 1M study was done nearly four years after the data was collected. No reason is given for the use of the older data, but as Spink points out, even though web search use is rapidly increasing, the ways in which search is being carried out does not necessarily change as well. Longitudinal studies would have to be done to show if public queries are evolving and changing over time (Spink, 2001). In both studies conducted using the Excite data, users are anonymous. Users are determined by a unique identifier. There are some issues with using only Web transaction logs as the source of analysis. A primary problem is that there are no distinct boundaries for determining when a session begins and when it ends. There are two ways to handle this, one is to set up a maximum number of log entries per session, such as the 100 used in one of the studies (Spink, in press), or to use a temporal cutoff, such as 15 minutes. A related issue is that there is no way to consistently identify individual from non-individual searchers (Spink, in press). An example of a non-individual searcher is an automated agent. Another problem with using only transaction logs is that there is no information from the user on their thoughts during the search, such as when they are frustrated during a search or when they have decided to change search direction. Transaction logs alone cannot always identify such scenarios. Even with such limitations, the studies can show some insight into how search queries are constructed.

A final study looks at European as opposed to American Web searching to find similarities and differences. The source for the European data is again a Web transaction log, from a Web search engine called AllTheWeb. There are over 1 million queries from 200,000 users, but instead of taking place on a single day, the European logs look at two days: February 6, 2001 and May 28, 2002. The logs span the period of a year, which makes it possible to look at how Web searching has changed for users of AllTheWeb. This type of analysis was not possible with the Excite data, since in those studies the data collected was from March and September of the same year.

The Web search studies define a query as "the entire string of terms submitted by a searcher in a given instance", and a term "is any series of characters separated by white space" (Spink, in press). A query can contain logical operators, such as AND, OR, and AND NOT, as well as special modifiers like + and -, which indicate a term must or must not, respectively, appear in a result. Quotes around two or more terms indicate a phrase.

#### Relevance feedback

Spink found that in mediated searching, term relevance feedback (TRF) was most likely used when search terms were exhausted and the user and mediator wanted to continue the search (Spink, 1997). That is, TRF is used late in the search process. Spink also found that although only about one-tenth of search terms came from TRF, TRF accounted for two-thirds of positive document identification (Spink, 1997). Overall, TRF was not used much by mediated searching, but when it was used, TRF was very effective.

In Web search and retrieval, Spink found four distinct session patterns that use relevance feedback (shown in Figure 1). Each session pattern is composed of session states, which are defined as:

- *Initial Query* is the first string of terms that a user entered for a session.
- *Modified Query* is the subsequent query related to the query before it by using same terms or being the same topic.
- *Next Page* is a request to view the next results page.
- *New Query* is a subsequent query by a user unrelated to the previous query.

- *Relevance Feedback* is the use of relevance feedback (e.g., selecting the hyperlink "More Like This").
- *Previous Query* is the subsequent query that is exactly like the previous query.

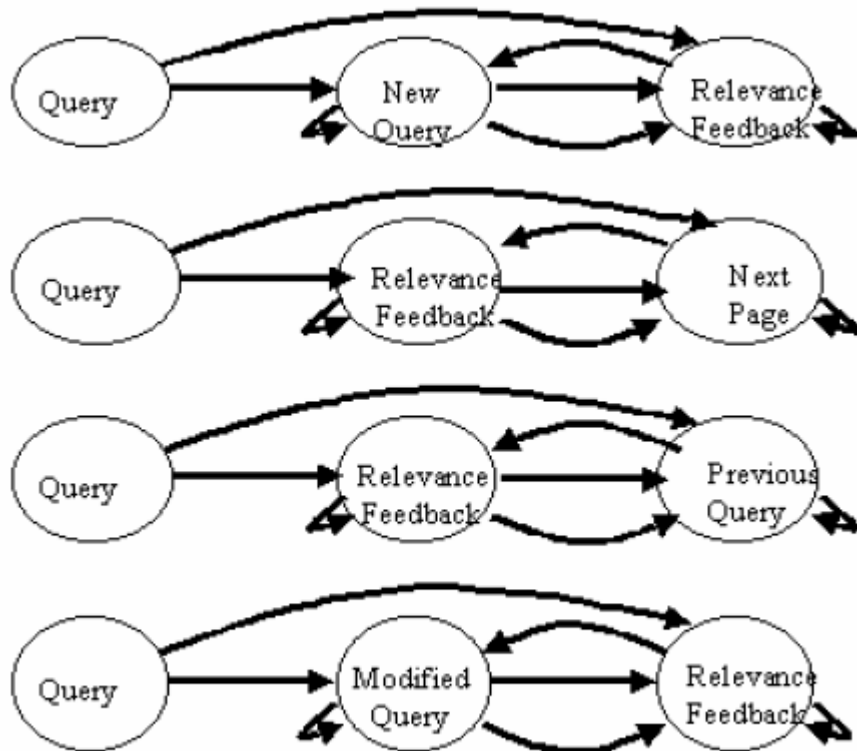


Figure 1 - Relevance Feedback Session Patterns (Spink, 1999)

The implication of identifying these states is that IR system designers should accommodate these states and patterns that searchers may find themselves in. In the 51K study, of 51,473 queries from 18,113 users, only 1,597 queries from 823 users were TRF queries (Spink, 1999). This indicates that most users are not taking advantage of TRF in their search. Also, users of TRF were more likely to have longer search sessions and were more determined in their search efforts than average search users. Interestingly, Spink's analysis shows that while TRF is successful 60% of the time it is used, very few web searchers take advantage of the feature (Spink, 1999). Spink believes this may be true because the failure rate of 40% is still too high for most users to invest the effort to use the feature. A potential way

to improve Web IR system design is to automatically retrieve and present TRF documents so that TRF is not user-initiated (Spink, 1999).

Topical relevance is a standard measure utilized in information retrieval to evaluate the effectiveness of a query based on the documents retrieved (Saracevic, 1975). It is a direct topic match between user's search terms and the terms in the retrieved document, rather than a user relevance indication as part of an information seeking goal (Spink, in press). In her studies, Spink asks users to use a binary value for calculating relevance (either a document is or is not relevant). She finds that users, despite not taking advantage of search engine features, are finding relevant documents about 50% of the time. She also finds that users spend little evaluating the relevancy documents, and often rely on the abstract that appears next to document retrieval listing to make a relevance judgment (Spink, in press). This has implications more for content providers than IR system designers: content should be easy to view and abstracts should be concise and to-the-point.

#### Advanced search techniques

There are several advanced search facilities that are available for Web searching. Spink finds that users generally do not use these facilities. For example, in the 51K study less than 10% of queries used Boolean operators, while in the 1M study less than 5% of queries used Boolean operators (Spink, 2001). Also, in the 51K study, 9% of queries contained a + or - modifier, and only 6% of queries consisted of a phrase (two or more terms surrounded by quotes). More interestingly, for searchers using Boolean operators, 50% misused the AND operator, 28% the OR operator, and 19% the AND NOT operator. For the AND operator, the most common mistake was in not capitalizing the AND as required by the search engine (e.g. "information and visualization" as opposed to the correct "information AND visualization"). Another problem encountered with the Boolean AND operator is user confusion of 'and' as a conjunction and 'and' as an operator. Users seem to have trouble making a distinction between the two.

Modifiers are another advanced search technique. Modifiers are generally + (plus), - (minus), and quotes. In the 51K study, the + and - modifiers were used in 1 of 9 queries but in 75% of the cases were used incorrectly (Spink, 1998). The use of quotation marks to indicate a search phrase was use in only 1 of every 20 queries, but only 8% of queries mistakenly used quotation

marks. The most common problem with modifiers is in their specification. The + and - modifiers must appear at the front of a term and not have a space separator. This is different than Boolean operators, which require a space between search terms. Another common problem with + and - modifiers is that users use their knowledge of mathematical notation and construct queries with + and - after a term and not before it. For example, 'news+weather rather than the correct '+news +weather'. Even though most users do not take advantage of advanced search features, Spink found that some users issued queries that attempted to take advantage of features that were not available on Excite. Examples given are using the NEAR proximity operator, and the usage of '\*' and '?' as stemming operators.

Spink believes that Web search users and advanced searching currently do not work well together. Web search users do not take advantage of operators and modifiers, and when they do, mistakes are usually made. Web IR system design changes need to be made to make operators and modifiers more consistent in usage. Alternative interfaces are also another possible way to help users construct queries. Spink (1998) specifically mentions Venn diagrams as a user interface for constructing Boolean queries.

#### Term Characteristics

Since Spink's focus is on search terms, she characterizes them in extensive detail throughout her research in Web searching. The number of terms per query in the 51K study was 2.32 while in the 1M study there were 2.4 terms per query (Spink, 2001). The number of terms per query percentage wise in the 1M study was 26.6% with one term, 31.5% with two terms, 18.2% with three terms, and dropping off to less than 1.8% of queries having more than seven terms.

The distribution of terms is shown in the rank-frequency plot of Figure 2. A Zipf distribution follows a straight line, so the plot shows that the term distribution is not quite Zipf and probably requires a more sophisticated model (Spink, 2001). Modeling the distribution of the linguistic characteristics of Web queries is important because, as Spink points out, the user query language can be anticipated and supported.

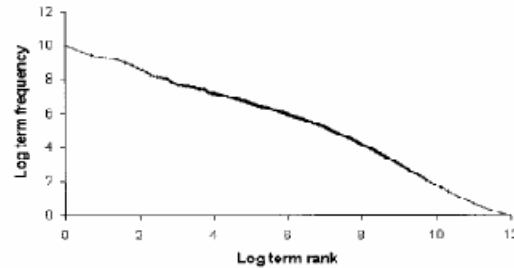


Figure 2: Rank-frequency distribution of terms used within unique queries (Spink, 2001)

In her 51K and 1M studies, Spink identifies the most frequently occurring terms. In the 51K study, the top 63 terms are identified, while in the 1M study the top 67 terms are identified. These top terms account for approximately 10% of all terms in all queries (Spink, 2000). An interesting aspect of term distribution is the term use opposite the most frequent - the terms used only once. Terms with a frequency account of one accounted for nearly 9% of all terms in all queries. This long tail in the term distribution is not well understood and requires further research.

Spink also uses search terms, as well as queries, in order to find out what types of information people are looking for. She first uses a co-occurrence of terms analysis to determine the term pairs that occur most often in unique queries. The result was a list of the fifty most frequently occurring term pairs. Although Spink developed an arbitrary classification for the most frequently appearing single search terms, she did not do so for the co-occurrence terms. Instead, a single list was developed from which you can draw your own classifications. Spink did, however, classify entire queries using a random sampling of queries. The classification was done based on work done at Rutgers to develop a Web classification system (Spink, 2001). The result of classifying the queries is shown in Figure 3. The subject classifications are 1-Entertainment, 2-Sex, 3-Commerce, 4-Computers, 5-Health, 6-People, places, things, 7-Society, 8-Education, 9-Performing/fine arts, 10-Government, 11-Unknown.

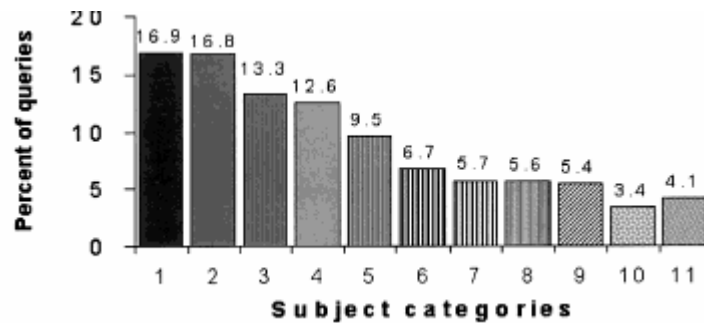


Figure 3: Distribution of sample queries across subject categories (Spink, 2001)

There are several interesting items about classifying terms and queries. The first is that what content is available on the web is not what people are looking for. Spink cites a study that describes the distribution of content on the web as 83% commercial, 6% educational, 3% health, 2% personal, 1% pornography, and 5% not described (Spink, 2001). Figure 3 shows an example of content and searching: close to 10% of searches are for health topics while just 3% of sites are health related. An implication for Web IR system designers is that the public finds classifications very understandable (Spink, 2001), and so the use of classification may be a good thing to design into a Web IR system.

Comparison of American and European Web Search and Retrieval

American and European Web searchers share some commonalities. European searchers use three or fewer search terms for approximately 85% of their queries (Spink, in press). American searchers used three or fewer search terms in 80% of queries. Searchers from both areas also predominantly use English terms for searching. Although this is not unexpected for American searchers where English is the primary language, it is for European searchers since for the searchers studied, English is a secondary language (most searchers from the European study are from Norway and Germany) (Spink, in press). Both American and European Web searchers spent little time make relevant judgments. The majority of searchers spent 15 minutes or fewer viewing retrieved documents. Also, information seeking sessions of both searchers are generally short (Spink, in press, 2001).

There are differences between American and European searchers as well. The top three categories for information seekers in Europe are People-places--things, Computers, and Commerce (Spink, in press). For American information seekers, the top three categories are Entertainment, Sex, and Commerce (Spink 2001). American searchers spent more time searching for e-commerce sites than did their European counterparts (Spink, in press).

Unfortunately Spink did not examine the use of advanced search features in the European study. It would have been interesting to see if there is any consistency across cultures for the avoidance and/or misuse of search techniques such as Boolean operators and query modifiers. If so, this would be further encourage IR system developers to develop better interfaces for taking advantage of these features. Also, Spink did not examine the use of relevance feedback in the European study. Again, it would be interesting to see if whether or not European searchers take advantage of the relevance feedback facility built into most Web search engines. Since the European study is recent and the paper researched is in press, looking at these additional areas may be Spink's intention. For the Excite study data, Spink researched different aspects of American searchers from 1997 - 2003. So, Spink may just be beginning to examine European searchers and how they compare to American searchers.

#### Conclusions and Future Research

One of the implications of Spink's work is that there is a difference in how mediated search in traditional IR works as contrasted by Web search. Mediated searchers use an average of 7-15 terms for search terms as opposed to less than three terms for Web searchers (Spink, 2001). For Web searching, interactions are short and number of search terms small. Mediated searches also rely heavily on mediators, particular in term relevance feedback (TRF). In mediated searches, the TRF provided by professional mediators are responsible for generating two-thirds of relevant search terms, while in Web searching only 11% of search terms come from TRF. In both mediated and Web searching, most users do not take advantage of TRF event though it seems highly desirable (Spink, 2001). Mediated searchers tend to use Boolean operators extensively as compared to Web searchers, who rarely used Boolean operators (Spink, 2001).

After examining the research thread of Web searching over several years, the following general conclusions can be drawn from Spink's work in Web search and retrieval:

- *The number of terms in a web query averages about two;*
- *Relevance feedback is rarely used;*
- *Boolean operators and modifiers are not frequently used, and when they are used, most users find difficulties in using them correctly;*
- *Users do not spend much time making relevancy judgments;*
- *The term frequency distribution is highly skewed: a few terms are used quite often while many terms are used only once;*

Based on the conclusions, there is future research that can look into the issues raised by Spink's work. The research might include:

*Improve the use of advanced search techniques:*

As Spink writes, "The low use of advanced searching techniques would seem to support continued research into new types of user interfaces, intelligent user interfaces, or the use of software agents to aid users in a much simplified and transparent manner" (Spink, 1998).

*Improve use of relevance feedback:*

Term relevance feedback is not widely used in mediated and Web searching. It may be possible to have IR systems automatically incorporate TRF into query results, so that the returned document set includes the results of TRF searches as well as the user-specified term search.

*Improve classification of search results:*

Spink reports that users find classification of query results easy to understand (Spink, 2001). Classification provides an overview of results that allow users to drilldown into areas which may contain more relevant documents. User interfaces based on classification may be more effective than simply providing an unordered list of potentially relevant documents.

*Better understand the language use in Web searches:*

Since the term distribution is highly-skewed, consisting of a few widely used terms one end and many single-use terms on the other, it would be useful to study how users are using language to perform their search. IR system designs

can then be adapted to account for how searchers use language when interacting with the system.

*Examine cultural differences in Web search*

Spink begins to examine how users from different cultures search the Web. Additional research is needed to examine issues with term relevance feedback and advanced search techniques. These areas have not been examined by Spink in prior research.

In summary, Spink finds that Web search is different in many ways from the traditional IR searching. Web IR system designers need to consider the ways in which searchers use Web interfaces for searching. Traditional IR systems are not as effective with Web searches. Most Web searchers focus on high-precision, their sessions are short, and their term use is also short. Understanding how Web search is utilized is important not just for improving search engines for the general public, but in understanding how interfaces to future digital libraries may be constructed.

References:

Budd, J. (2000). Scholarly productivity of U.S. LIS Faculty: An Update. *Library Quarterly*, 70(2), 230-245.

Ding, Chowdhury and Foo (1998). Mapping the intellectual structure of information retrieval studies: An author co-citation analysis, 1987-1997. *Journal of Information Science*, 25(1), 67-78.

\*Jansen, B. J., and Spink, A. (in press). An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management*.

\*Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*. 36(2), 207-227.

\*Jansen, B. J., Spink, A. and Saracevic, T. (1999). The use of relevance feedback on the web: Implications for web IR system design. In *1999 World Conference on the WWW and Internet*, Honolulu, Hawaii.

\*Jansen, B. J., Spink, A., and Saracevic, T. (1998). Failure analysis in query construction: Data and analysis from a large sample of Web queries. *Proceedings of the 3rd ACM Digital Libraries Conference*, 289-290.

Saracevic, T. (2004). Personal web site at Rutgers University, School of Communication, Information and Library Studies. <http://www.scils.rutgers.edu/~tefko/> as of March 4, 2004.

Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society of Information Science*, 26(6), 321-343.

Spink, A. (2004). Personal web site at University of Pittsburgh, School of Information Sciences. <http://www.sis.pitt.edu/~aspink/> as of March 4, 2004.

\*Spink, A. (2003). Web Search: Emerging Patterns. *Library Trends*, 52(2), 299-306.

\*Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

\*Spink, A. and Saracevic, T. (1997). Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science*, 48(8), 741-761.

\* indicates a core paper used in the stream of researched analyzed.