A SYSTEMATIC REVIEW OF INTERACTIVE INFORMATION RETRIEVAL EVALUATION STUDIES, 2007-2016

HAIDER ALI HASAN ALYASEEN

AUGUST 2017

A SYSTEMATIC REVIEW OF INTERACTIVE INFORMATION RETRIEVAL

EVALUATION STUDIES, 2007-2016

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED

SCIENCES OF
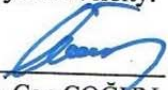
ÇANKAYA UNIVERSITY

BY

HAIDER ALI HASAN ALYASEEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF

MASTER OF SCIENCE

IN

DEPARTMENT OF MATHMATICS

INFORMATION TECHNOLOGY PROGRAM

AUGUST 2017

Title of the Thesis A **SYSTEMATIC REVIEW OF INTERACTIVE INFORMATION RETRIEVAL EVALUATION STUDIES, 2007-2016**

Submitted by: **HAIDER ALYASEEN**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof. Dr. Can ÇOĞUN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Fahd JARAD
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Abdül Kadir GÖRÜR
Supervisor

Examination Date: 17.08.2017
Examining Committee Members

Assist. Prof. Dr. Abdül Kadir GÖRÜR	(Çankaya Univ.)
Assist. Prof. Dr. Reza HASSANPOUR	(Çankaya Univ.)
Assist. Prof. Dr. Bülent Gürsel EMİROĞLU	(Kırıkkale Univ.)

STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name   :   Haider ALYASEEN

Signature   :

Date   :   17.08.2017

ABSTRACT

A SYSTEMATIC REVIEW OF INTERACTIVE INFORMATION RETRIEVAL EVALUATION STUDIES, 2007-2016

ALYASEEN, Haider

M.S., Department of Mathematics
Information Technology Program
Supervisor: Assist. Prof. Dr. Abdül Kadir GÖRÜR
August 2017, 57 pages

Since the last mid-century researchers start investigate in the performance of IIR (interactive information retrieval). Many methods and measures were innovated and used in this field studies. To ensure the replication of maturation researches there are several factors that effect on the maturation of researches such as introducing standards for measurement and analysis, and understand past endeavors. In this study, we analyzed a vast range of papers within the period (2007-2016), where 1110 papers were examined manually and only 78 articles were included. Based on the achieved results in our study, we found that the researchers increased their concentration on IIR evaluation. Due this expansion of researching in this topic, we noticed that researches used some new techniques and datasets for evaluation of IIR systems. Most of the included papers were conducted based on the help of participants and questionnaires.

Keywords: Interactive IR, Systematic Reviews.

ÖZ

ETKILEŞIMLI BİLGİ ERIŞIMI DEĞERLENDİRME ÇALIŞMALARININ SİSTEMATİK INCELENMEŞI, 2007-2016

ALYASEEN, Haider

Yüksek Lisans, Matematik Dalı /Bilgi Teknolojileri Programı

Tez Yöneticisi: Yrd. Doç. Dr. Abdül Kadir GÖRÜR

Ağustos 2017, 57 sayfa

Geçen yüzyılın ortalarından buyana, araştırmacılar IIR'nin (etkileşimli bilgi erişim sistemleri) performansı üzerine araştırmalar başlatmışlardır. Birçok yeni yöntem ve ölçme sistemi bulunmuş ve bu alandaki çalışmalarda kullanılmıştır. Olgunlaşmış araştırmaların çoğaltılmasını sağlamak amacıyla, ölçme ve analiz için standartların belirlenmesi ve geçmişte yapılan çabaların anlaşılması gibi araştırmaların olgunlaşması üzerinde etkiye sahip olan birkaç faktör mevcuttur. Bu çalışmada, 2007-2016 yıllarını kapsayan bir dönemde çok sayıda makaleyi analiz ederek, bunların arasından 1100 adet makale el ile incelenip sadece 78 adeti araştırmaya dahil edilmiştir. Çalışmamızda elde edilen sonuçlara dayalı olarak, araştırmacıların çalışmalarını IIR değerlendirmesi üzerine yoğunlaştırdıkları görülmüştür. Bu konu üzerine araştırmaların artmasından dolayı, araştırmacıların IIR sistemlerinin değerlendirmesi için bazı yeni teknikleri ve veri kümeleri kullandıklarını fark ettik. Çalışmamıza dahil edilen çoğu makale katılımcıların ve anketlerin yardımıyla yürütülmüştür.

Anahtar Kelimeler: Etkileşimli IR, Sistematik lncelenmeşi.

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor Assist. Prof. Dr. Abdül Kadir GÖRÜR of the Computer Engineering Department at Cankaya University, without his helpful advice, valuable comments and guidance this thesis could not be completed. I want to thank my wife Suad and my children for their support during this defficult period. In addition, I dedicate my success to the spirit of my dear father, may God have mercy on him. Also, thanks very much my family especially my mother, my brothers and sisters who encourage me in this period .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Information retrieval system (IR) is fundamentally, considered as one amongst the computer science disciplines that are used to retrieve relevant documents in response to a user's query. The concept and idea of search engines can be clarified and analyzed as a matching process, a user introduces a query to the IR system which stores data, then after matching the process the end results will be listed as a text titles even if the user meant other types of data for instance, video, image etc. The obstacles that associated with retrieving relevant data are hardware and software [1]. The development of IR systems can be improved with smart algorithms that fetch the relevant data which meets user's query. Building smart and efficient algorithms representing a big challenge for IR engineers. IR systems found to fulfill two major tasks, the first task is recognizing and defining relevant information corresponding to user's query. Second task IR systems must response within short time.

Before 60 years ago, engineers and IR designers tried to introduce an IR system that provide an accurate matching for user's query with retrieved results, they realized that they should have some knowledge of user's interaction with IR systems in order to help them to improve their IR's design. IIR system is stand for interactive information retrieval system, which centered around enabling individuals to manage, explore and resolve their data by interacting with their information systems. IIR composed of investigations of individual's information, search conducts, user's utilization of interfaces, search characteristics, and user's collaborations with systems as well. IIR research is additionally focus on arrangement, ordering and recovery systems that are designed to individual user. Many people consider interactive information retrieval as a new field due to the rapid increase of online searching since 1990's.

IIR influences by many factors, mostly these factors are neither easily measurable nor observable. There are vast variety between users in term of the amount of knowledge about the topics that they are interested in, their search experience, and what are the expectations about yielded results [2,3]. The variety of these factors increase the difficulty of building experimental environment that enable all users to have the same opportunity to experience the same, thusly it becomes very hard and simply uneasy to establish a causal relationship. In addition, measuring these factors is not generally effective since there are likely to be an extensive number of factors. As known, there are daily interactions that occur between users and IR systems around the world, the retrieved results do not always make users satisfied. Because IR investigators became flexible with accepting the intrinsically intuitive nature of the IR procedure, and also because the designers of IR system advances to the improvement of end-client frameworks, with the more appreciation about the significance of testing interaction itself was improved.

The results in all IR systems can be classified into two major classes and groups, with results being included and excluded. Due to this variation of responses, many studies were presented to evaluate and examine the performance of these IR systems; the researchers used a vast range of measurements, methods, and datasets for the purpose of evaluation. A systematic review (will be explained in next chapter) would help to collect these methods and determine the sufficient metrics, methods used for IIR evaluation researches. Retrospectively, smart and profound research in the assessment of information retrieval systems started in the mid of 1960s, specifically at Western Reserve University in the USA and at the Cranfield Institute of Technology in the UK. The promptest address emerging in the assessment of these information retrieval systems was step by step instructions to build up criteria and related measures as indicated by which they could be assessed.

1.2 Related Work

Although there are many studies published to evaluate IIR systems, improve some features of IIR systems or other type of studies in this field, only a very small number of

studies that presented a systematic review of IIR systems. In [4] authors presented a comprehensive study between (1967 to 2006). The authors realized that the number of search tool increases rapidly due to the tremendous increase of data volume and because of this increase of IIR tools, many studies introduced IIR assessments by using different metrics, methodologies and in some studies datasets. Additionally, researchers developed new methods of evaluation. Authors prepared the historical outline of 40 years of IIR assessment studies by utilizing the technique for systematic review.

The 2791 of both conference and journal units were manually inspected and 127 articles were chosen for investigation in this review which relies on both the inclusion and the exclusion criteria. The authors coded these articles systematically relies on some features such as researchers, when did they publish the article, where it was published in conference or journal, the most widely recognized sorts of corpora and measures utilized, An extra result of this exploration is a bibliography of IIR assessment look into which can be utilized by understudies, instructors and those new to the range. The authors in this study used systematic review as a method in this study. because the concentrate was on methods and measures. The main steps that authors followed to establish a systematic review as followed:

- Sources determination
- Determine the criteria for article's inclusion and exclusion.
- Approval of manual search and choice procedures.
- Improving coded scheme in order to analyze the chosen articles.
- Apply the improved coded scheme to articles.

The authors meant to reveal the sources that published IIR papers more than other sources. Authors determined inclusion and exclusion criteria as below:

1. Full length research papers were included whereas posters, brief communications, book reviews, short papers, demonstrations should be excluded.
2. If there is an article published in both journal and conference, they include just the journal.

3. The papers that revealed the summary data about a progression of studies could and should be excluded in light of the fact that these papers mostly have no sufficient detail that can be used for coding.

4. Despite there are several types of evaluation in involved with IIR (for instance video, images etc.), but only the evaluation of text-based search should be included.

After the preparation of a list of sources, they asked four of the IIR experts to evaluate the list. The experts were recommended to increase the list of sources and modify them. Authors collected the articles between the periods; (1967 to 2006). The first edition of Annual Review of Information Science and Technology (ARIST) in 1966 contained some chapters that talks about evaluation of man-machine communication which could be representing an IIR studies. Through studying and reviewing 2791 units, authors found that 2667 were journals and 241 were conference units. They tested units, about 97% of journals (2592) and 89% (206) of conference, because some papers were not easy to get them to their hands or too costly.

Authors performed inclusion and exclusion according to the criteria they collected, which match the objective of the type of study which is IIR system. They applied these criteria on a small portion of their collected units, preciously 147 unit, which represents (5%) of the total number of collected units. Firstly, they examined titles and abstracts, then if the decision is difficult to make, they examined the whole paper. Authors used to meet periodically for the purpose of modifying inclusion and exclusion criteria in some conditions, and rechecking the included articles.

Authors considered that one of the most important conditions to include an article, is that users should interacted with IIR systems by requesting some queries and reviewing the results and judged the retrieved results. This criterion preciously was the reason to exclude a vast number of papers, some papers showed that users only searching for queries without evaluating for results, other papers showed the opposite, only results evaluation without human interactions. As mentioned previously, authors relied on traditional text retrieval for two reasons: the first one because the study would be too

large if other types of search (audio, video, images, etc.) were included. The second reason was because the variety of data types need different methods for evaluations, that would make the analysis of results more difficult. Another criterion yielded a large exclusion, the papers should contain an evaluation of at least two IR systems. Studies exhibited search behaviors were also excluded since these studies care only about user's selection behavior but not evaluation for that system. Authors used several queries as listed below:

- The users or user study.
- An interactive information retrieval.
- The user evaluation.
- A human subject.
- An interactive retrieval.
- An interactive IR.
- The user interaction.

These queries were used in all the journals and conferences that used in this study. They found that 127 articles matched the criteria used for inclusions.

Authors found that the period between 1997 -2006 is the golden age of the publication, where (70%; n=89 out of 127) whereas only 27% of papers were published between 1986-1996. The minorities of publications were between 1977-1986, less than 4% were published during that period. Figure (1) below shows the distribution of publications between 1967 and 2006. between 1967 and 2006
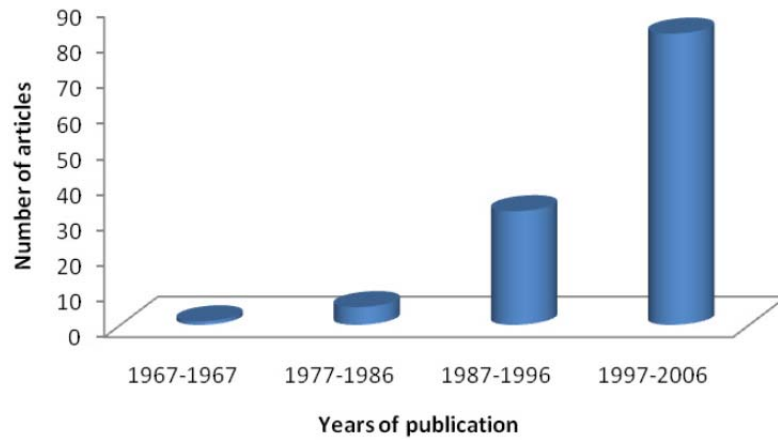
Figure 1 Published articles by years [4].

Authors have found out; that about 50 % of the papers were published in 3 publications: Namely, JASIST, IP&M and the SIGIR Proceedings, whereas the other 50% of papers were distributed between other 21 papers as shown and illustrated in figure (2).
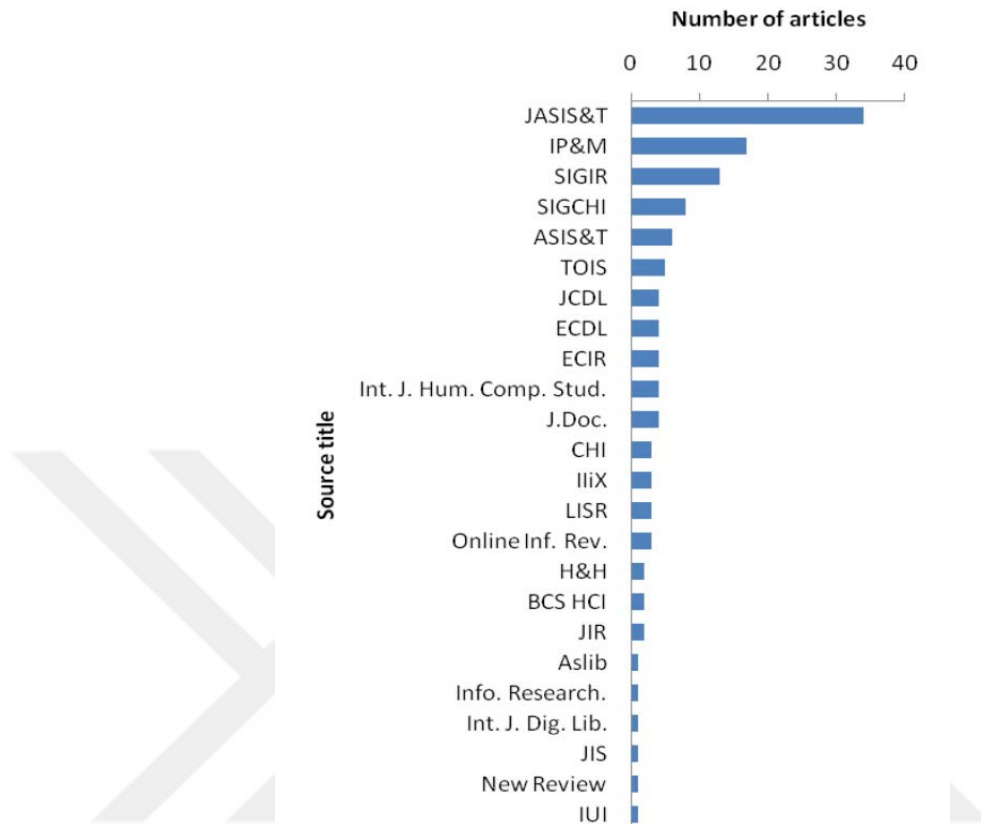
Figure 2 Publications versus number of articles [4].

In term of number of publishers, the highest number of authors was 9 and lowest was 1 author, 36% by 2 researchers, 24% by 3 and 22% were composed by 4 to 9 researchers. The majority of authors (83,1%, n=212) participated with only one publication, whereas only (13%, n=33) participated with only two papers, (3%, n=7) and (2%, n=4) participated with three and four publications respectively. The smallest ratio (1%) represents authors who participated in 5,6, and 7 publications where (n=2 and 1). Authors found that around 6% participated in three or more publications.

Authors found that the articles were represented by only 20 countries, United State came in the first place in term of number papers country representation where (n=77), second place goes to United Kingdom (n=23) as shown in figure 3.
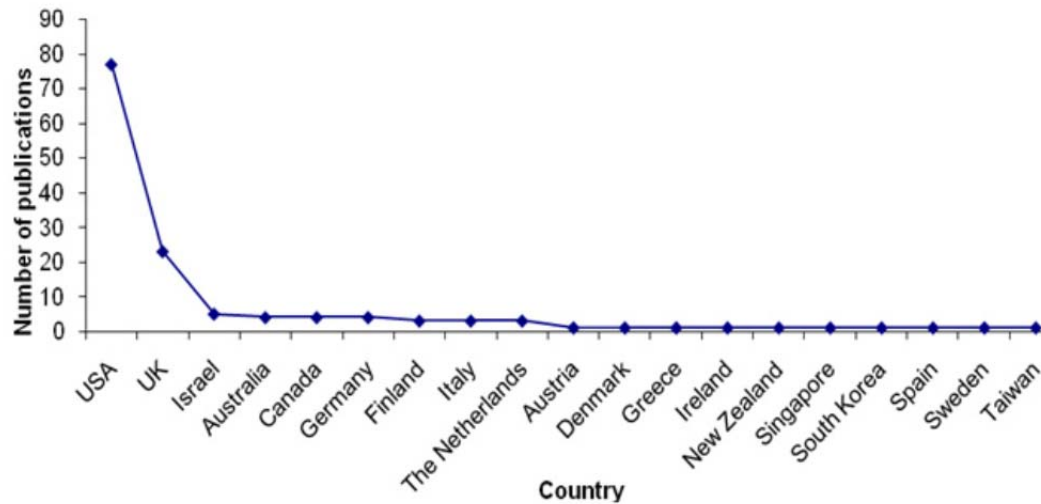
Figure 3 Countries versus number of publications [4].

Authors also tested the articles whether they contained explicit research questions or explicit hypothesis or research question and a hypothesis. The found that (19.3%, n=29) of articles contained explicit research questions, about (10.7%, n=16) of these articles contained explicit hypothesis, whereas (4.7%, n=7) contained research question and a hypothesis. The majority of articles about (65.3%, n=98), didn't contain explicit research questions or explicit hypothesis or even research question and a hypothesis. Authors also revealed that there were eighty-eight unique research questions. They clustered these questions using key terms as in the table 1.

In term of theory use, they found that the majority (64%, n=81) of articles didn't contain a clear use of theory, over (31%, n=40) of articles used one references from the behavior of information-seeking literature. Only (3%, n=4) of studied articles referenced from other area of theories, while (1%, n=2) of articles used both references (the behavior of information-seeking literature and other fields).

Table 1 Explicit research questions-key words [4].

Term	Number of questions
Effectiveness	17
User preference and opinion	14
User behavior and interaction	12
Usability	11
Relevance	7
System performance	5
Time	5
Queries	4
User training and knowledge	3
Quality	3
Accuracy	1
Comparison	1
Reliability	1
Satisfaction	1
Theory	1
Unassigned	9

Authors frequently used; Provalis Research's WordStat/QDA Miner to reveal the highest repeated terms as illustrated in table 2, where the highest 15 are repeated terms.

Table 2 Objective statements-Key words [4].

Word	Frequency
Search	95
System	76
User	58
Interface	41
Compare	34
Performance	33
Information	33
Effectiveness	31
Subject	29
Use	25
Retrieval	25
Evaluate	23
Query	21
Base	21
Investigate	19

Subjects

Authors revealed the number of subjects in the 127 articles, the average number of subjects were (37.07) per study, (min=4 subjects/study, max=238 subjects/study).
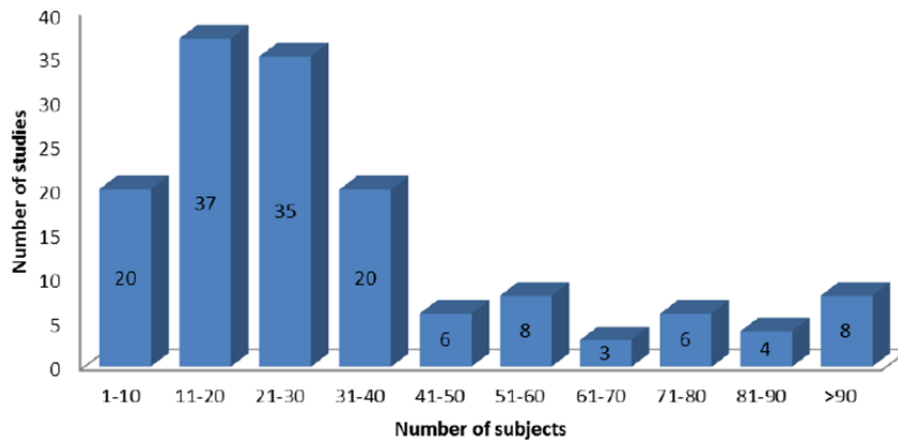
Figure 4 No. of participants in each study [4].

Search tasks

The search tasks were inspected on the bases of whether or not they were allocated to the subject or were of the subject's decision. The tasks were additionally inspected for indicated sorts or topics and the tasks number from each subject were made by a request to finish. Figure 5: Shows and illustrates the number of articles versus tasks number.
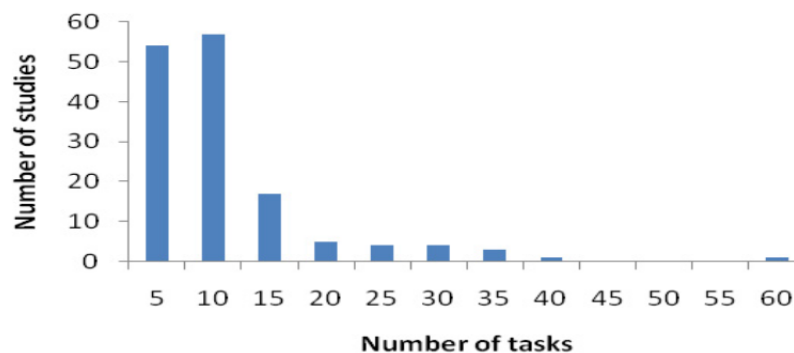
Figure 5 Tasks versus number of studies [4].

In spite of the fact that time was utilized as a measure in a considerable way according to the reviews, just 38.7% of articles announced a correct time restrict.

Study Design and Method of Analysis

Most of the studies were performed inside the labs or other was controlled in settings. 149 studies out of 150, indicated that these studies were within-subjects or between the subject studies. The ratio of studies within-subject were about (69%, n=103), and (31%, n=46) of studies were between the subjects. Commonly one gets more prominent power with less number of subjects when utilizing a within-subjects design, which may be the reason this sort of design is utilized vastly.

Most papers did not include a particular area describing strategies for examination, all things considered outcomes were analyzed to figure out which sorts of factual tests were directed for discussion.

Cited Works

Authors compiled all the 127 papers in an excel spreadsheet, they found that 3128 references were cited in those 127 articles. The average of citation was 24.6 references for each article. Minimum number of references was 3, whereas the maximum numbers were 77 references. 43% of references were journal articles, while 37% were conferences, and only 12% were monographs, additionally, authors also analyzed the year of publications for all the 3128 references, as shown in figure 6.
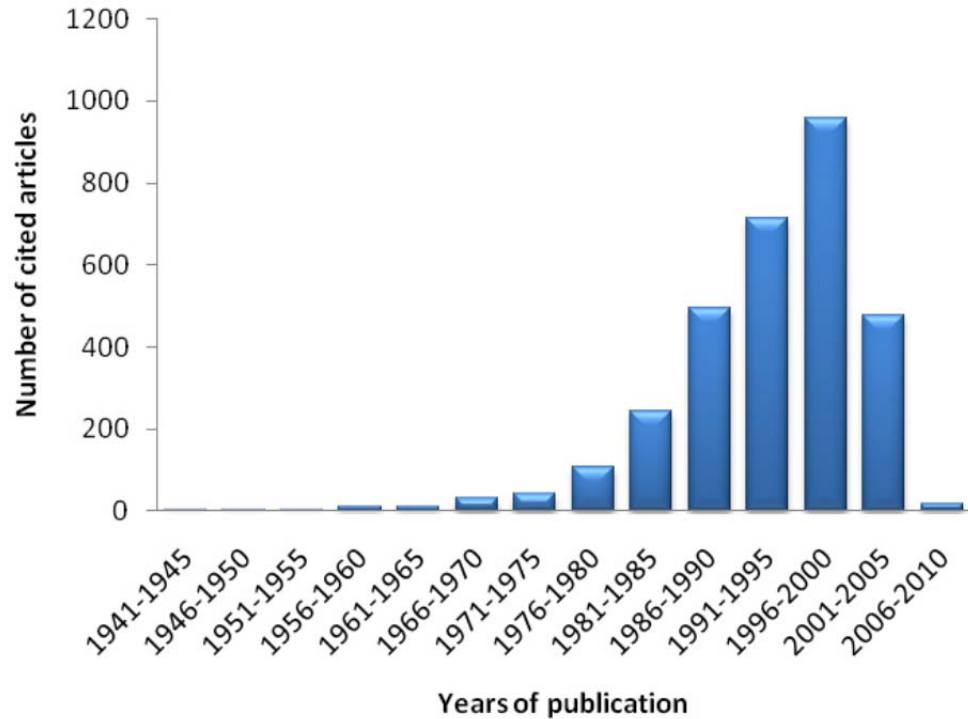
Figure 6 Number of cited articles versus years of publications [4]

There are other types of references such as webpages, report, and technical reports. Webpages were characterized under the heading "Web", while reports and specialized reports were under the "Report" class. A total of 730 remarkable sources were perceived among the referred to works. They were situated in two ways:

1. The quantity of one of a kind articles refers to that source.
2. The aggregate number of references.

The third investigation was utilized to exhibit which sources have contributed expansive quantities of articles to the information base, instead of the fewer very cited to articles. The larger part of the sources were cited by just a single paper (68%), which were cited to just once (65%), and contributed to the one or less one of a kind of articles to the rundown of cited works (75%).

Authors concluded that although there may be many studies which were published to evaluate IIR systems and with methods, measurements, but there is still too much to do

in the term of the development of the methods and the measures. As more researchers concentrate on the plan and change of structures that help IIR, there is an extended prerequisite for the directions about how to lead evaluations and an extended necessity for suitable and dependable measures. One essential result of this exploration is the IIR assessment examines a list of sources. This bibliography can be utilized as a perusing list for graduate understudies or those new to the range or as a starting stage for an instructor aggregating a perusing list for IIR. While Michael Unterkalmsteiner et al. [5] have tended to assessment and estimation of Software Process Improvement by evaluating 148 articles distributed within the period 1991 to 2008. In [6] the authors, Magne Jørgensen with Martin Shepperd, finished a systematic review on Cost Estimation by looking over 304 articles distributed in 76 journals. In [7] authors systematically distinguished and analyzed all the available papers on pair programming. In [8] another systematic review study which was presented in term of public health. The authors depict the techniques created for surveying investigate on individuals' points of view and experiences close by trials inside a progression of audits on youngsters' psychological wellness, physical action, and adhering to a good diet. Authors Tracy Hall, et al [9] presented articles talk about motivation in the field of software engineering. In [10] authors found 29 essential studies, were published within the period January 2011 and June 2014. A large portion of them concentrate on programming improvement, and to a lesser degree, requirements, and other bolster regions. The authors essentially found 1079 papers related to their topic, after removing duplicates titles, they eventually found 29 papers as in figure 7.
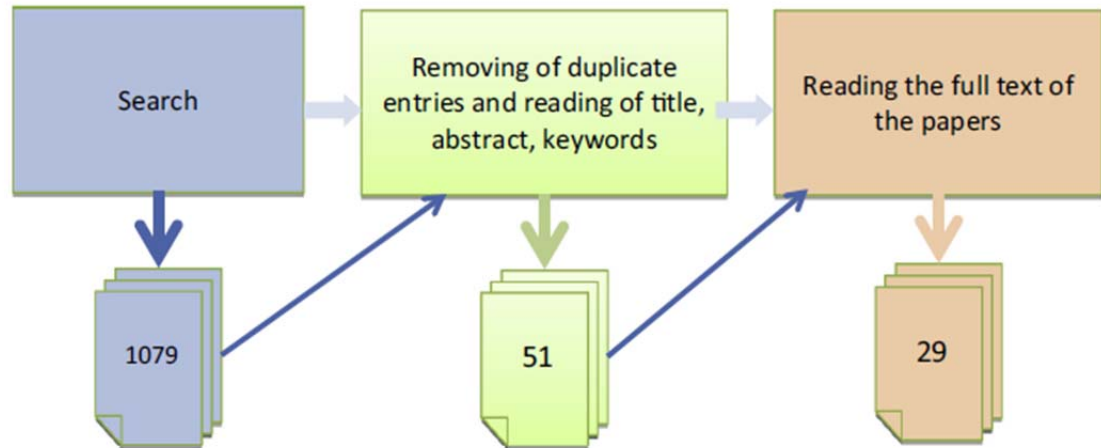
Figure 7 Observed results from search process [10].

1.3 Purpose of the Study

Our study considers an extension for the study in [4], we intend reveal the state of art of evaluation interactive information retrieval systems for the period between (2007-2016). Our study should present for the researchers a clear guide if they work in this discipline in term of using methods, datasets etc.

CHAPTER TWO

BACKGROUND

2.1 What is Systematic Review?

Systematic reviews are a sort of literature survey that gathers and fundamentally analyze different research studies or papers, utilizing strategies that are chosen before at least one research question is formulated, and after that discovering and analyzing related studies. In other word is the summary of several studies that stat unambiguous questions, utilize systematic a strategy to distinguish, select, and basically assess pertinent reviews, and to gather and analyze information from them. A systematic review is in this way a bit of research in its own particular right and, by its tendency, can address considerably more extensive inquiries than single exact reviews ever can. To be sure, systematic reviews sit most importantly other research plans at the highest point of the 'chain of command of confirmation' since they can possibly give the most critical practical implications. A giant number of research is published each year, routinely with conflicting revelations. These between-study differentiation may be a result of study differences, deformities or good fortune (testing assortment). In such conditions, it is not for the most part clear what the general picture is, or which comes about are most strong and should be used as the purpose behind training and approach decisions. A systematic survey endeavors to collect all observational prove that fits pre-determined qualification criteria keeping in mind the end goal to answer a particular research address. There are several key features of systematic review [11]:

- Several clear targets with pre-characterized qualification criteria for ponders.
- An unequivocal, reproducible procedure.
- A precise hunt should meet the qualification criteria when it is identifying all studies.

- An assessment of the legitimacy of the deciding the included investigations.
- A methodical introduction of the qualities and discoveries of the included investigations.

2.2 Systematic Review Conducting

For the purpose of conducting a systematic survey, many guidelines were published. The aim of systematic review is addressing issues by identifying, critically evaluating and integrating all the discovered studies and addressing research questions. The successful systematic review must fulfill:

- Build up to what degree existing exploration has progressed towards illuminating a particular issue.
- Distinguish relations, inconsistencies, and abnormalities in the written work, and examine clarifications behind these.
- Formulate general explanations or an all-encompassing conceptualization.
- Comment on, assess, extend, or develop theory.
- Give description for the future work.

We conclude that a systematic review is a sort of researches since it addresses questions more than empirical studies. Systematic reviews are described by being objective, efficient, straightforward and replicable. There are several steps used to conduct a systematic review as follow:

1. Identify clarify the aim of the research.
2. Determine the studies that will be included in the systematic review, this step depends on the first step.
3. Identify sources to search for the studies.
4. Choose the included studies and collect the data within these studies.
5. Evaluate the risk that could occur from biasing within the included studies.
6. Analyze the collecting data.
7. Introduce the final results.

2.3 Interactive Information Retrieval System (IIR)

In this section, we presented a detailed background on the: IIR systems. We began with the more extensive picture of data searching and then we narrowed it down to the picture to concentrate on the information retrieval system. A depiction of their contributing components is additionally given. These contain:

- The data seeking process.
- The Relevancy.
- The query formulation.
- The presentation of results.
- The visualization.

For many years, information retrieval researches have two main directions:

1. The system-oriented approach

   This deals with user behavior by taking a simplified view on it. It can be simply explained; when users submit his query after that the retrieved results will be checked in term of ranked items. So, the aim of the systems is to list the relevant retrieved items on the top of the results.

2. Cognitive approach

   Concentrates on the user; in light of observational studies they designed models of the client's intellectual systems while recovery. Up until this point:

   The system-oriented perspective of IIR systems has been tested on several fronts. These incorporate data needs, relevancy that can be described as non-binary, data searching and the need to consider the collaboration and human association in the assessment. As mentioned in [12] that TREC interactive track was a try to confirm the assumptions fundamental the system-oriented approach. Shockingly, the eventual outcomes of this evaluation illustrated that differences in term of performance of a system disappeared in interactive retrieval. The reason stands behind this results that user can easily identify the relevant entries within a document list [13].

2.3.1 Information Searching Process

Information behavior can be described as a person's engagement in the activities while determining their information needs [14]. In case if users fail to find their information, the searching will be continued. The continued process of seeking relays on the relevancy of the retrieved data, if the retrieved results are relevant, then the information seeking process will be finished. There are many reasons that might lead to the needs of information. Different aspects that effect on information needs were discussed and explained in several studies.

2.3.1.1 Information Requirements

While, information seeking or triggering can be the most basic factor sought through information that the user can't act upon. Taylor in [15], Dervin, in [16] and Belkin et al. [17] summarizes the different aspects of information. Marchionini, [18] explains the following requirements as well. Taylor, in [15] defines four levels of information needs:

- Visceral: The visceral level recognizes some defects, but there is no cognitive definition.
- Conscious: At the level of consciousness, the information searcher describes the defect and limits it.
- Formalized: And been able to express this problem; Albeit ambiguous at the formal level, one can clearly articulate the explicit statement of the problem (e.g. English).
- Compromised: A level is a formal statement that is presented by the search system as a form constraint limit. Taylor's work laid the foundation for a deeper concept.

Understand the motivations of information search; as a result, we can have multiple types of data needs, for instance fuzzy, and many more.

Dervin, in [16] has been paying special attention to user's needs. The model understands the world according to people's needs. The model assumes the user through three stages to understand the world, that it, faces and solves their data problem. The first stage defines the contents of the data needs, named the situation. In some cases, people see the

understanding and the gap between them which needs to be understood at the current situation. These gaps are reflected in the problem. And then uses the answers or assumptions of these gaps to move to the next case. The use of the situation is more suitable than the information to seek more general human conditions. Belkin and his colleagues developed an information model; Seeking a knowledge state (ASK) that focuses on information seekers' anomalies [19].

In this model, the information searcher is worried about an issue, however the issue itself and the data expected to take care of the issue are unclear. For data searchers; It is obvious that the search request must be expressed by the process of clarification which means that the search system should support iterations and interactive conversations with the user must go through a process of clarification.

This model is intended to explain the general openness of information, not directly applicable to the actual search type of the problem or to pursue information experts in one field. The ASK display is the hypothetical premise of data plan and profoundly intuitive framework. Taylor's inside organs and cognizant data levels need to compare to what German says of a "gap", and the " unordinary condition of learning" that Belkin and his colleagues asserted. Marchionini in [20] stated that Information problems have been removed from defects: A person's mental capacity has some ideas, events or objects.

2.3.1.2 Tasks

In general, the user's information search is designed to solve the problem and complete the task. Although the tasks in the information search study is barely taken seriously. Vakkari mentioned in [21]: it is generally believed that data looking for is assignment situated. Others may state that; this is the undertaking of motivating the action. In this manner, it is important to consider the assignment. In the event that we need to completely comprehend the human data conduct, in general a task can be depicted as an action to be accomplished. An objective anyway, it is helpful to concentrate on look as far as inquiry conduct task. Search quests are natural and stand out from the real actors' tasks [21]. The simulation task is a man-made modification. This endeavor to give the

searcher a more hearty depiction of the data issue. These sorts of assignments can be utilized for research facility assessment to give search. Evaluate the scene of the search system or interface feature set. Investigate the relationship between the complexity of different tasks and the search of information in some studies.   In [22] Campbell reviewed the task complexity of few researches were classified as follow:

1. The main task of the psychological experience of the implementation.
2. The interaction between the person who perform a task and that task.
3. The functional objective task characteristics.

Bystrom and J-arvelin studied the influence of task complexity on the type of data, (seek and use). Their classification helps in defining five task-based complexities based on a priori [23].

In [24] Borlund, presented the use of simulated work tasks to construct more realistic search task. The simulated work task is not only a brief description of the narration. A narrator needs information, but also needs for the situation a task information. The simulation task is interesting in providing a search contents for the searcher by which the searchers can evaluate.

Toms et al [25] evaluated the impact of task domain on search. These consists of: consumers travel, shopping and health. They found a big difference with search methods used in several areas.

Bell in [26] combined the five categories of Bystrom and J-arvelin. Divided to 3 categories, and tested whether they could predict the degree of complexity of the predictability a manual search task.

2.3.1.3 Relevancy
Relevance considers a basic term in IR science. The previous view was focused at the semantic level was described by Glover et al in [27], relevance alludes to the binary state whether the document is the same as the subject of the query. Schamber et al. in [28] re-examined the literature of relevance and noticed by asserting: Is a dynamic and multi-

dimensional cognitive concept. This is a complex but systematic one measurable phenomenon Saracevic, distinguished 5 correlations:

1. Systems or algorithms
2. Themes
3. Targeted or cognitive
4. Context
5. Motivations

Algorithm correlation is objective, it is the same, no matter how many times the searcher tries. The other 4 kinds describe the relevance: As a subjective idea that relies upon the searcher and its data setting. The subject relevance describes the degree to which the searcher believes that; there is a match document substance and its data should be required. To get such relevance, records out to thought process positive opinions, for example, fulfillment, accomplishment and achievement [29]. Borlund in 2003 [30] separated relevance into 2 major kinds:

- Systematic-oriented
- Subjective.

These framework driven methodologies regard relevance as static and target in contrast with the intellectual approach that acknowledges pertinence to be a subjective customized mental experience.

2.3.1.4 Models and Empirical Studies

By checking the information search model, it is important because the searcher is actually doing the real task by looking for data, as this might be altogether different from what others consider searcher. The model under this category describes the model that the user uses to locate and get data resources. Kuhlthau in [31] stated that based on some longitudinal studies, simulate information Student and library user search the process. It identified many different stages in the information search process. These included launching, selecting, exploring, and developing, through: Collection and presentation. It felt suspicious, anxious and depressed feeling with information seeking. The data

seeking procedures spans the information searches for activities in a conversation, not a single moment in time [31]. This is like the information seeking behavior model presented by Ellis in [32]. The following features: start, link, browse, distinguish, monitor, extract, verify and so on and so forth. Within the meeting, the searcher's knowledge state is not static but close to dynamic; as the search progresses, it is changes. Any of the steps in the process do not have to be: As the searcher can skip or repeat the steps. Kuhlthau's model is very similar to that of Eisenberg and Berkowitz [33].

Choo et al. [34] developed a network information search model which collect between both: Search and browse. They suggested that many of models of Ellis have been executed. The components can be found now in the Web browser. The searcher can start from the site. (Link), link to information resource (link), bookmark page (distinction).

Broder [35] classifies network queries into 3 kinds: navigation, information, and transactional. Based on the survey results, around 73% of the request are enlightening. About 26% were cruising, and an expected 36% were exchanging. A few inquiries have a place with various classifications. As indicated by the log investigation, more extensive announced that 48% of the inquiries were data, 20% route and 30% arrangements.

2.4 Information Search

Data inquiry can be viewed as a mix of IIR. classic IR, not exclusively to consider the learning of the searcher, also considers the underlying model to match the information needs. The collection of searches; therefore, we first consider IIR then the classic IR.

2.4.1 Interactive Information Retrieval

Information search and behavioral description are interactive features [36]. IR as information search behavior is the behavior of the "micro level". The searcher is used to interacting with various information systems. It includes all interactions with the frameworks, both at the level of human-computer interaction, for instance, using the mouse device and clicking on the link or at the intelligence level which is also involving mental behavior, for instance, judging the relevance of information or related data. Bates, (1989) presented information on the "Berry Picking" model (shown in Figure). It

is assumed that the user needs to change while viewing the retrieved document. Thus leading to new unexpected directions. Through the search, the user collects the relevant items retrieved through different queries (berry-pick)[37]. This method has also been supported by other studies in [32] and [38].
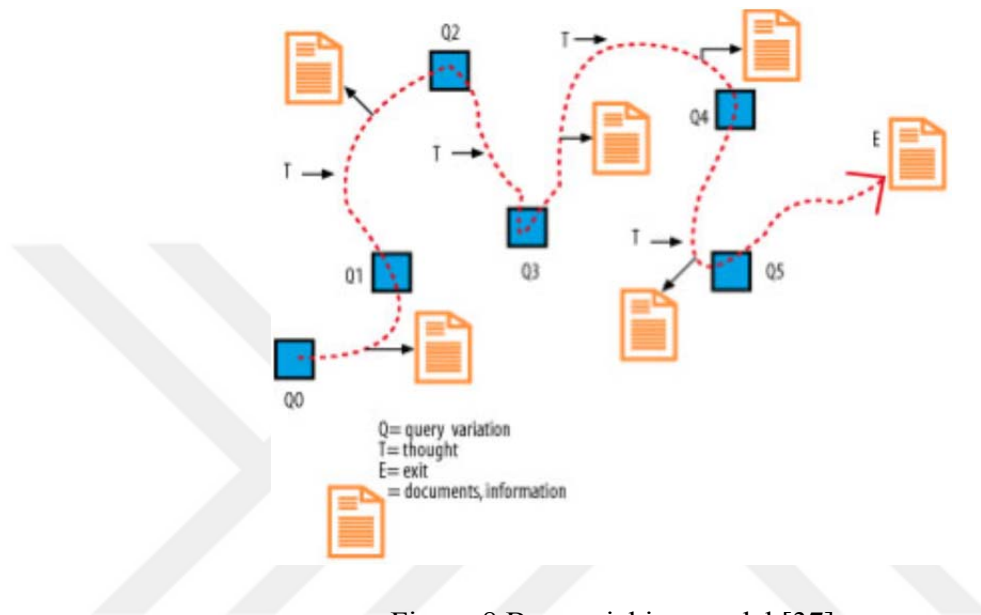
Figure 8 Berry-picking model [37]

In the strategic model, author asserted that; the user may adopt different strategies and methods. It defines the interaction with information, for example, improved search. Bates made one including 4 levels of search actions; move, strategy, strategy, and strategy [39]. Belkin's "plot pattern" in [40] considers the user's interaction with the information retrival systems. As a series of several interactions with the information search plot. This model is an action performed in the search of information of 4 binary values size:

1. Method (scan or search)
2. Interactive target (learning choice)
3. Retrieval mode (identification - specification)
4. Considered resources (information-meta information)

The mix of these measurements prompts 16 unique snippets of data looking for system. As per Belkin et al., Any single ISS (data looking for system) can be depicted by their

position along these four measurements. So as to conquer "ASK ". The situation, they presented the script or plan such a script, for example, based on and abstract The observation of the person involved in the search for information can be utilizes as an average for structured human-computer interaction, designed to achieve the objectives of a particular ISS.

In Saracevic's hierarchical model complex processes are broken down into hierarchies in order to conduct more detailed research and interdependence at each level. It treats the process as a surface level that involves the computer and the user meeting during a surface level. Interface, with several different levels or stages. For the user, the level of hypothesis is cognitive, emotions and contexts. These levels represent the user's explanation, motivation and level of calmness. For PCs, the prescribed level is designing, handling, and substance level. These levels relate to equipment, preparing, and information structures, individually. At that point the association is the cooperation between these distinctive levels [41].

The intuitive input and seeking process demonstrate was created by Spink Spin k in [42] by expecting the cyclic idea of the information retrieval interaction: This model is gotten from an observational research. It recognizes a portion of the parts of the inquiry procedure when a man cooperates with an IR framework. These include users judgment, search strategy or mobile, interactive feedback loop. In Spink units each search strategy may contain 1 or more cycles. Each cycle may contain one or more interactive feedback. The input can also indicate the movement within the search strategy and can be considered as a search strategy for further search. Each move is made up of users input or query request system output.

2.4.2 Information Retrieval (IR)

Information retrieval is a science that identifies and retrieves information from a collection. It responds to the information needs of the searcher. Lancaster in [43] elaborated on this definition: The information retrieval is not notified as the information retrieval system (i.e., the change is made through Knowledge) the user asks. It just informs the existence (or does not exist) and the whereabouts of the documents related

24

to their request. The information usually needs to be represented by a string of words, and the IR system uses it to match the mechanism to determine the relationship between the document and the subject of the survey. There are many possibilities for defining the basic search units considered for matching. Mechanism: complete document, part of the document, XML element or sentence can be seen as atomic units. For instance, when a document is considered to be the entire components the document matches the query. By retrieving the document part of the match, such as segment and paragraph. The motives behind this approach are two aspects described by the pioneer. This method (Salton et al in [44]: efficiency and effectiveness. Efficiency comes from the user because they did not face a lot of information; effectiveness has larger units with larger pieces of information which are easier to retrieve. İn term of models, the research of the retrieval model has been very independent of cognitive work. The classical methods of the above methods are Fuzzy and boolean search, The probability model and the vector space model are still the main practical applications even in the current study to find. However, most of search studies the model concentraint on 2 major extensions of the probabilistic method, the probability reasoning and language models.

The logical view of the IR system is introduced, where the search is interpreted as uncertain reasoning; Rijsbergen proposed the concept of the probability of instability: For this reason: Let Q signify the present inquiry and a report, at that point the framework should: Is designed to estimate the probability . The main advantage of this approach is its ability also consider the complex reasoning process. However, the model does not give a specification of how to derive its parameters from the actual data [45].

2.5 Query Formulation

There is no point by point information of the arrangement and recovery of the earth, Most clients think that its hard to develop very much composed inquiries reestablish. Website page internet searcher perceptions demonstrate that clients frequently adjust their preparatory request [46]. The first query should be treated as one Guess [47]. Some methods can help users in this case to query inaccuracy: These incorporate non-interactive and interactive techniques for question expansions. They can be based on

user participation level which is a comparison of two methods. Non-interactive method work without user intervention and extend the query at the algorithm level. In another case, suggests a list of users that they can identify and select. These terms are considered to be more relevant to the tasks on their hand.

2.5.1 Terms and Conditions

The term relationship can be built up from various distinctive resources around the world or neighborhood level. The worldwide approach alludes to the count of the term relation that is considered every one of the reports from the whole corpus, while the nearby method is constrained the underlying retrieved sets of documents is receptive to the query [48]. The author in [49] incorporated the idea of global analysis into a local analysis practice. The conventional method of terminology similarity is based on statistical methods. Ie G. Common frequency, mutual information and chi-square.

By considering the complete file, paragraph sentence, or fixed-size window, the words that appear in the text can be estimated in a number of ways [50]. İn [51] extracts terminology and constructs the concept hierarchy from the search results and uses the term to calculate the concept of relationship. A new alternative to the term is to identify the relevant query terms in the collected log [52].

2.5.2 Relevance Feedback

Relevance feedback , express and certain has ended up being a compelling method. Enhance search results [53]. Relevant feedback techniques need to obtain information about the relevance of the retrived seeking results and submitted to the searcher. Initially the associated costs are considered to be user-oriented where the user must mark documents related to their information that needs to be in hand. Later this concept extends to a two-way process, both the user and the system can respond to another in IIR [54].

Empirical studies have shown that IIR system users want explicit relevance feedback characteristics[55]. Recessive feedback and implicit feedback technology does not unambiguously infer the information needs of a search behavior, and can be utilized to personalize system responses and build system user models. As the main application

areas, hidden feedback technology has been developed for the recommendation and filtration systems.

Many of the behaviors have been described as potentially relevant in the literature. Feedback indicator. Nichols in [56] developed an observable classification scheme. The behavior is focusing on its application in the information filtering system. It presents a list of potentially observable acts; increases the purchase, evaluation, reuse, reference, mark, catch, association and query to those mentioned above. Oard and Kim expanded the work of organizing two observable behaviors: Behavioral axis refers to the fundamental purpose of behavior. Further subdivided into 4 categories: check, keep, reference and annotations [57].

The annotations are searchers engaged in actions that intentionally add personal values. Information objects such as markup, rating and organization documentation. Horizontal axis: "minimum range refers to the smallest unit associated with the behavior". The segment level includes the operation of the smallest extent that is part of the object (For example, a section is part of the document). The object is a self-contained item (such as a document). A class is a set of objects.

2.6 Results are Presented and Visualized

After the background matches, the search engine in turn will returns the results list in a way that is less likely. Both results and relevance are introduced to the searcher in an organized form of a document proxy. This one is the most advanced search engine results which presents the main way. Experience studies have shown that query-based sentences can help to assess relevance search results; they are more effective files rather than by the most advanced search engine presented document fragments [58]. Visual representation is an effective method of communicating information. Hearst in [59] divided current visualization techniques as follows: highlighting, brushing and linking, Translation and zoom, focus plus context, magic lens and overview plus detail [60].

2.7 Interactive Information Retrieval System Evaluation

With the dawn of IIR systems, the need to evaluate the vast range of different IIR systems appeared. A few criteria and measures have been proposed and utilized as a part

of assessing IIR performance. There is no clear idea about what is an effective IIR execution or which are the best available assessment measure(s).

2.7.1 System-Driven Evaluation

The evaluation of the system based on the model of Cranfield to test the quality of the IR system. By considering the test set. The main purpose of this assessment is to evaluate the algorithm [61]:

1. How good is indexing technology?
2. How good is the ranking algorithm?
3. How good is it?

Relevance feedback: This kind of assessment does not require user involvement It can be done in a controlled lab. The test set consists of three components:

1. A group is different from thousands of titles to TB bytes,
2. Are usually collected by creators and occasionally from the actual query,
3. Contains the relevance of the information to judge

Related or unrelated files to respond to each query. relevance is gotten in various methods by enrolling assessors, or through various methods for gathering the work together. Most collectibles are too large to be fully evaluated to find all relevant documents. Therefore, the summary is performed before the correlation judgment of each subject is obtained. The fundamental idea is constraint just around those documents that are well on the way to be important. The IR framework runs a similar subject to get a list of the most important documents. A fixed gets the quantity of best level reports from each run and union them into a pool. Then read each document and evaluate its relevance. Keeping in mind the end goal to assess the execution of a specific calculation, two measures were utilized: recall and precision. Accuracy reports the proportion of relevant and recalled retrieval documents, measured the proportion of relevant documents retrieved. High recall refers to retrieval everything is relevant, but the accuracy is high and the precision is high, which means retrieval (Which may be small) a set of highly relevant files. The system is usually evaluated in a variety of ways Recall

level. The F metric (Equation 2.1) combines precision and regression into a number. People can adjust the index according to the accuracy and recall interest.

F-measure $\alpha = (1 + \alpha) \bullet P \bullet R\ \alpha \bullet P + R$ (2.1)

The Clanfield method's assumptions are often criticized because of [62]:

1. Related documents assuming that they are independent of each other
2. All documents are similarly significant
3. Emphasizes high memories
4. Interactions are ignored.

2.7.2 User-Centered Evaluation

User-oriented measures to evaluate the system as a whole include algorithms and interfaces. The components of these assessments are experimental subjects, search tasks, systems and collections. This assessment is carried out in a relatively controlled environment.

The TREC interaction track was created to develop a better assessment method

IIR system. The method of track use is criticized for adaptability. Interactive TREC does not process information needs, but processes pre-built information Request, binary correlation assessment Hybrid evaluation [63].

2.7.3 Mixed Assessment

In 2003 Borlund proposed a hybrid approach for IIR system evaluation. Taking into account the information and the relevance of the information needs of the searcher and experimental control. They proposed the "half-life and the relative relevance" of measures. The measures of the effectiveness of IR systems. These measures were based on the subject and the relevance of objective types [30].

2.7.4 Operational Evaluation

The fourth is the operation of the whole system in the actual situation when the assessment within are not controlled settings. The searchers use their own work and they decide when to stop; searching with no preparation, it is hard to clarify the results, however they are more sensible. Longitudinal assessment have a few likenesses with this

sort of data accepting that the issue holds on finished a more drawn out time, for example, days, weeks, months, or months even a couple of years. A portion of the studies along these routes concentrated on data look conduct [31].

CHAPTER THREE

RESULTS

3.1 Method

Our study considers an extension for the study in [4], the same techniques were used to build our systematic review within the period (2007-2016). In this study, we can evaluate the used methods and measures in IIR evaluations in the last ten years and if there were any improvements on these methods or measures while evaluating the performance of IIR systems. We followed several steps to build our systematic review, we highlighted on these steps in this chapter in details as below:

1. Sources identifications: the purpose of sources identification is to determine in which source IIR studies were mostly published. Since our study consider an extension to study [4], we used the same sources used in [4] to collect our dataset (9 journals and 6 conferences). Only the studies that included the assessment of text-based searching were considered so this implied it was pointless to include the journals and conferences of multimedia. As known that there is more than one type of academic works such as posters, demonstrations, brief communications, short papers and books, all these types of studies were excluded only full-length papers were included. If there is a paper published in both journal and conference, just the journal article considered to be included. There are some papers which presented an overview of several studies, this kind of studies were excluded due to the shortness of some details used for coding.

Table 3 Publications examined

Type	NAME	1967-2006	2007-2016
Journal	ACM Transactions on Computer-Human Interaction (CHI)	52	49
	ACM Transactions on Information Systems (TOIS)	96	56
	Aslib Proceedings	356	127
	Information Processing & Management (IP&M)	234	64
	Information Research: an international electronic journal	47	11
	Journal of Documentation	177	54
	Journal of Information Retrieval, The (JIR)	32	79
	Journal of the American Society for Information Science & Technology (JASIS&T)	340	101
	Online Information Review	172	31
Conference	Conference on Human Information Interaction and Retrieval (CHIIR)	27	23
	ACM International Conference on Information and Knowledge Management (CIKM)	14	126
	ACM Special Interest Group on Information Retrieval Conference (SIGIR)	30	197
	ACM/IEEE Conference on Digital Libraries (JCDL)	18	103
	Information Interaction in Context (IIiX)	1	44
	IEEE Conference		45

We included in this study 572 journal units (each unit represents an issue), and 538 conference unit (each unit represents a proceeding). The problem that we faced while examining the units that it was difficult to have access to all articles due to high cost for some of them or they were unavailable. The table below (3.1) shows the number of examined paper in each source versus the results of study [4]. We can notice in table 3.1 that the total number of retrieved conference units in our studies greater than the number of retrieved conference units in study [4].

According to the results that we have got, the most journal that retrieved results was Aslib Proceedings, and for conference was ACM Special Interest Group on Information Retrieval Conference (SIGIR).

2. Inclusion and exclusion criteria development: The objective of this progression was to create and assess inclusion and exclusion criteria that could be utilized to systematically choose articles for the study. The improvement of the inclusion and exclusion criteria started with an aggregation of criteria that coordinated the objective study sort: IIR assessments. We relayed on several criteria which in turn were refined couple times, these criteria are listed below:

- Publication year: as we mentioned previously that this study an extension to study [4], the studies before 2007 were excluded whereas studies between 2007-2016 were tested.
- Study goal: studies should be an IIR evaluation and what we mean by evaluation is using scientific methods and metrics to compare at least two systems with a direct interaction where user enter a query and retrieve a result. We also included some papers that contain a creation of models and frameworks that used for IIR evaluations, we considered these kinds of studies presented new methods of IIR evaluation.
- User interaction: only studies that contain end user interactions were included.
- Type of query: only IIR systems that deals with textual documents were included, other types of IR systems such as (video, audio, images etc.) were excluded. we used several formulas of queries to search for IIR evaluations.

All these criteria were taken in consideration while testing the articles, we followed almost the same criteria that were used in [4]. We excluded studies that discus improving search behavior or fetching related results or classifying IIR systems based on some surveys such as in [64] or only discussing methods used for the purpose of evaluating IIR systems without applying these methods on systems and discussing the yielded results such as the study in [65], because these kinds of studies don't present

metrics or measurements to evaluate these IIR systems. We applied the criteria that we stated previously to all the 1,110 units from both journals and conferences. Only 7.2 % (n=78) paper matched the criteria and considered as included papers. It is not safe to say that this number of articles (78) represents the exact number of the all the published papers of IIR system evaluation within the last 10 years but we can say that the included papers represent the population of the IIR evaluation systems.

3. Validation of manual search process

The hardest part in this study, is the manual review of the articles. As known queries have a direct influence on the retrieved results, we used several queries that should expand the range of the retrieved results related to our top (IIR evaluation). We noticed that the ACM database is not working correctly, when we used the queries ((user or "user study" or "interactive information retrieval" or "user evaluation" or "human subject" or "interactive retrieval" or "interactive IR" or "user interaction") we got a vast range of results with many duplicates and out of range. Another problem was noticed in ACM database, when we select conference to look through only conferences, we get results from several journals. ACM database also presents a feature which enable users to downloads results in CSV files, but we found that the number of the results in the CSV file differs from the results shown on the screen for the same query, that happens in some cased where the number of results is big number.

4. Coding scheme

We coded our dataset manually according to several features that used in [4] which in turn will ease understanding this study for the readers. These features are common between all the included papers, additionally it will give us an idea about the year of publications, number of authors, the purpose of studies and their objectives, the methodology used in each study, and methods in each study and some other features.

Table 4 User coding features

Class	Item
Publication	Year
	Source
Contributors	Name and affiliation
	Country
Study Purpose	Research Questions, Hypothesis and use of theory
	Objectives
Method	Subjects: number, type, label, compensation
	Corpus: type of documents, document source
	Search task: type, number, time
	Study design, type of data analysis
Measures	Output measures: conceptual and operational definitions
Cited Work	Genres
	Years
	Source titles
	Item titles
	Cited authors

3.2 Results

According to the features that we set, we found that 78 papers matched the features that are specified for inclusion purposes. We coded all these 78 papers in next sections. The titles of the included papers found in this link (https://goo.gl/v1PpLE).

3.2.1 Publications Features

In this section, we are about to reveal the year that witnessed the highest number of published papers that included in this study. We found within the last ten years (2007-2016).
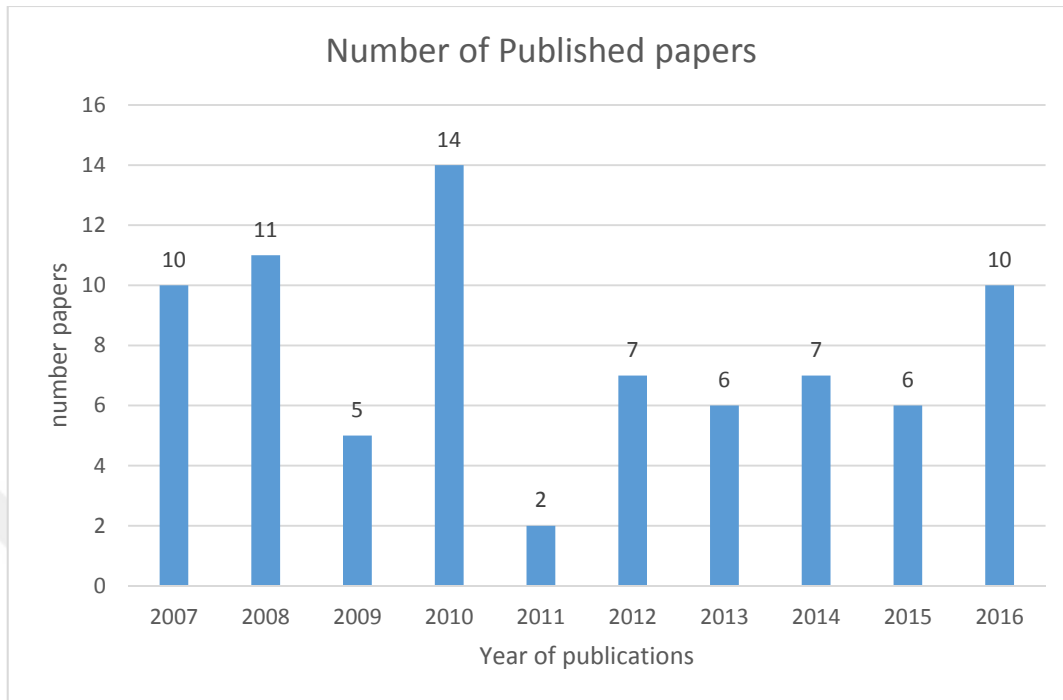
Figure 9 Number of published papers within last 10 years

As shown in figure (9) it is clearly that the highest number of published papers were distributed between 2007,2008 and 2010, about 44.8% (n=35) of papers within these three years were published. During 2011 only 2 papers were published. the average of publication in each year were about (7.8) papers.

The most included papers were found in ACM Special Interest Group on Information Retrieval Conference (SIGIR) where 17 papers out of 78 included papers were found in (ACM SIGIR), ACM IIiX came in second place where 11 papers were found in it. Figure 10 shows the number of included papers versus their sources. The majority of the included papers were shared between ACM (SIGIR) and ACM IIiX as mentioned previously, some of the sources used in [4] did not contain studies that matched our criteria for instance, Canadian Journal of Information and Library Science, interacting with computers journal and International Journal of Human-Computer Studies. Although in study [4] the majority of the included papers were found in Journal of the Association for Information Science and Technology (JASIS&T), but our research in this journal yielded only 8 papers that matched our criteria.
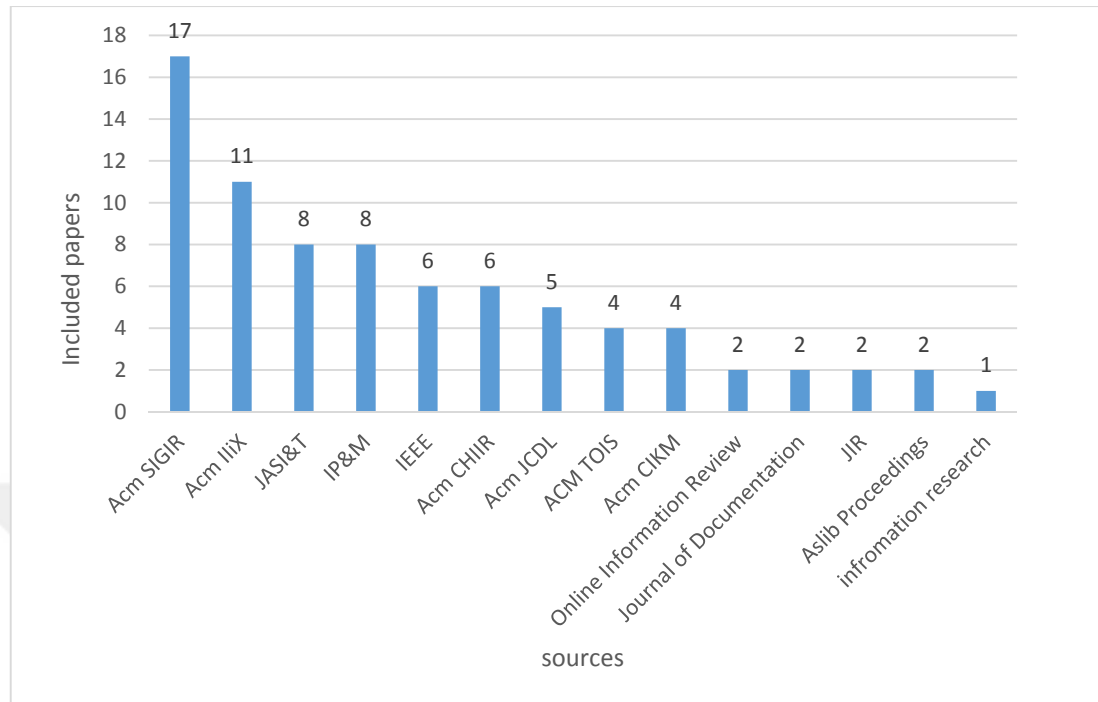
Figure 10 Number of included papers from each source

3.2.1.1 Authors

The number of authors differs from one article to another; we found that 254 authors contributed in the 78 articles that we included in this study. The maximum number of authors was 14 and the minimum number of authors was 1 authors, about 60.2% (n=47) of papers were written by 2-3 authors, 29.4% (n=23) were written by (4-6) authors, one study was written by 9 authors and one study was written by 14 authors.

Figure 11 shows the distribution of countries publications, where 14 countries contributed in the included 78 studies, USA came in the first place with 36 papers, UK in second place with 11 papers. 9 countries published between (1-3) papers which were included in our study.
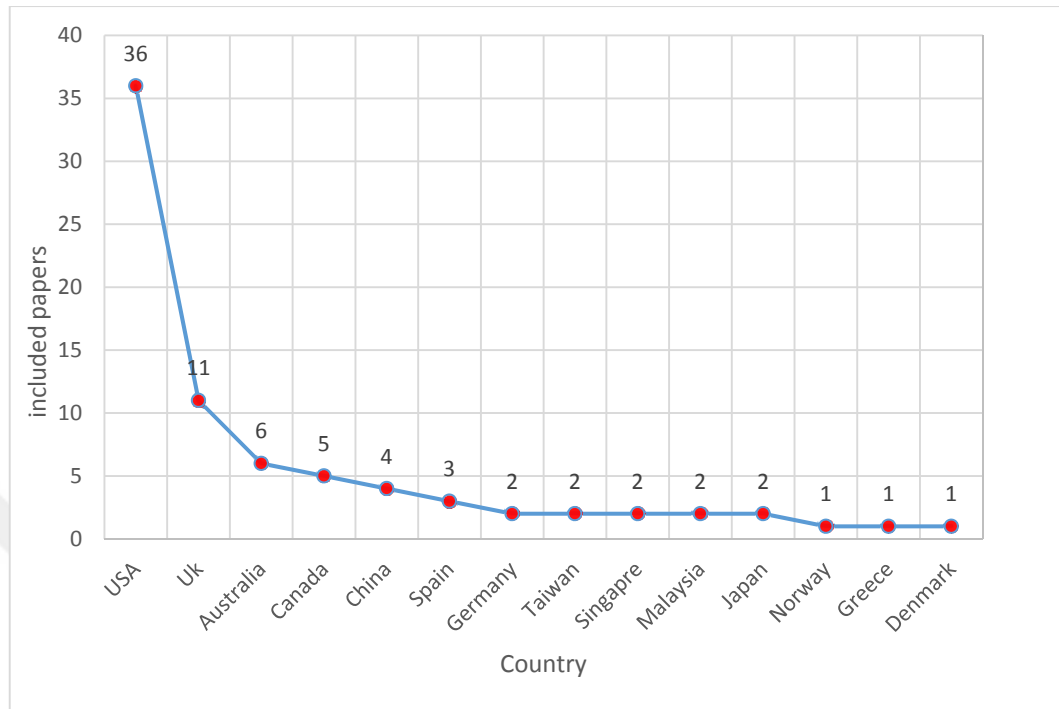
Figure 11 Countries contribution

3.2.2 Study Purpose

In this section, we coded several items such as research questions, objectives, hypothesis and finally explicit use of theory. The questions were recognized as targets and purposes expressly expressed as questions; hypotheses were recognized as those things expressly marked as theories, (for example, "it was hypothesized" or "the study was based on the accompanying hypotheses"); also, objectives were recognized as both casual and formal statements. Papers were classified in term of theories into three classes: papers without any theory, papers with theories from data searching, and papers with theories from other study.

3.2.2.1 Hypotheses, Research Questions

We discovered that 33.3% of papers (n=26) contain explicit hypothesis, 32% (n=25) included explicit questions, and 19.2% (n=15) included both hypotheses and explicit questions. We grouped these questions in a table based on some terms as shown in table. We should mention that these terms were classified according to the existence of these the terms in the questions. The unassigned questions were 22.

Table 5 Terms mentioned in research questions

Term	Number of questions
Effectiveness	6
User preference and opinion	2
User behavior and interaction	9
Usability	9
Relevance	14
System performance	3
Queries	5
User training and knowledge	2
Accuracy	1
Comparison	5
Reliability	6
Time	3
Test collection	2
Ranking	2
Unassigned	22

3.2.2.2 Objectives

After identifying the purpose statements within the 78 papers, we used RapidMiner software to look for the highest frequently occurring terms, we listed the highest 15 as shown in table 6.

Table 6 Key words used in objective statements.

Term	Frequency
Information	26
Retrieval	19
Evaluation	17
Interactive	14
Precision	8
Average	6
Search	6
User	6
Methods	5
Performance	4
Relevance	4
IR	4
Measure	3
Performance	2
Analysis	2

3.2.3 Methods

In order to analyze the number of subjects in the corpus, we looked through all the 78 papers. The figure (3.4) shows the distribution of subjects within each study. About 96.1% (n=75) of the included papers was conducted with a direct participating of subjects. There are two studies which don't include any participants, instead these two studies presented two different frameworks used to evaluate information retrieval systems. Although one of the criteria required a direct interaction of subjects with IR systems, but in [4] the authors hoped to find studies with frameworks used for IR evaluation that's why we included these two studies. In 54 studies students participated, we couldn't determine the number of students in all the 78 studies. The maximum number of subjects was 139 which was found in two studies, the minimum number of subjects was 6 in only one study.

3.2.3.1 Corpus

In term of test collections and dataset, 41.2% (n=32) described collections within the 78 studies. The most frequently utilized collection within these studies was TREC, where (29.4%) of studies used TREC, whereas INEX and CLEF which are test collections were used in 1 study in each. In addition, we found in two studies two datasets (synthetic dataset, OHSUMED3).

3.2.3.2 Search Tasks

We analyzed the 78 studies to find out the way that these studies deal with tasks in term of the type of the task whether it is assigned or un assigned to subjects, the number of tasks in each study and time consuming to perform that task. We discovered that 56 studies out of 78 included tasks which were performed by known number of subjects in some cases and unknown number numbers in other cases. In addition, some studies didn't state details about tasks in term of assignment. the total number of known tasks is 1590 task. Studies without precious number of tasks was 7 study, 22 study without no tasks, about 34.5% (n=28) of these tasks were assigned to subjects. In term of time, some tasks were performed in hours, and minutes. The maximum time was 120 minutes, and the minimum time was 5 minutes.

3.2.3.3 Study Design and Method of Analysis

Before we start discussing the results yielded from the analysis of the studies, we should know that there are two types of study design as below:

- Within study design: considers a kind of experimental design in which all members are presented to each treatment.
- Between Subjects Design: each participant take place in only one group and the results of group A will be compared with results with group B.

About 82% (n=64) of studies were within-study design and 11.1 % (n=9) were between study design. It is clear that the most used design in the 78 studies is the within design rather than between design that is because between design is more prevalent in search behavior studies rather than evaluations studies.

In term of the used analysis methods, t-test method was used in 17 study (The t-test evaluates whether the methods for two gatherings are factually not quite the same as each other. This investigation is proper at whatever point you need to think about the methods for two gatherings), Wilcoxon (The Wilcoxon test, which alludes to either the Rank Sum test or the Signed Rank test, which compares two groups.) test used in 3 studies, ANOVA test (utilized for comparing between the average of more than two unrelated groups to find out if there are statistically different between the means of these groups) was found in 24 studies which consider the most frequently used measures in all the 78 studies, and finally chi-square (The Chi-Square test of Independence is utilized to decide whether there is a significant relationship between two nominal (categorical) factors) was found in only 2 studies.

3.2.4 Measures

It was time consuming with a lot of effort to determine the used measures within the all the 78 studies some of studies. There were several measures used in the studies such as Precision which was the most frequently used measure, where 19.7% (n=16) study used this measure whereas recall used in 19.2% (n=15) of studies, mean average precision was also used in 2 studies, also f-measure was found in 4 studies. We also found 6 studies used frameworks. Mean Squared Error was found in only one study.

Although there are several types of measures, but the most used type in the 78 studies was performance measures such measures are measures that describe how far the subject is in achieving a particular task, precision and recall are good example of performance measures. Retrieved precision can be defined as the number of retrieved relevant results over the total number of retrieved results (relevant and non-relevant results); therefore, retrieved precision considers a session-based search instead of being query-based measures that in turn complicate the ability to implement an interactive situation. The only different between the retrieved precision and retrieved precision is that the subjects should view the results before they judge whether it is relevant or not, and the relationship for the viewed precision is that the number of the unique relevant results divided on the total number of viewed results. The calculation of precision in system-

centered assessment is clearer since it depends on whether a result is recovered or not; in client user-centered, results must be retrieved, seen and labeled as important by a subject. This shows a few issues utilizing system-centered assessment measures in IIR situations. Subjects regularly miss some of the retrieved results in the query items list that are viewed as relevant; that includes these unviewed results in the calculation of precision would not give a precise measure. Subjects additionally settle on choices about which reports seeing in view of what they have as of now observed; results that show up topically pertinent may be skipped since they don't give any new data. As a result of these troubles, specialists regularly detailed two arrangements of exactness measures; those in light of subjects' significance judgments and those in view of standard importance appraisals. In the last case, regardless of whether subjects saw the results were immaterial. Recall measure showed low variation. The most common utilize recall measure is the standard recall measure of the extent of related results retrieved by subjects. Obviously, there is as yet the issue of deciding if it is sufficient for a subject to retrieve a result or whether they have to see or even save the result for it to be incorporated into the calculation. Recall was accounted in term of usability more than precision which is less utilized.

The second type of measures is process measures, which interested in the interaction between subjects and systems, there are 26 studies that used process measure since the number of queries and number of clicks was related to the evaluation process within these 26 studies. Only 5 studies that used clicks only as a factor involved with the evaluation process.

The third type of measures is usability measures, where subjects use systems and gives their opinions for example answering questionnaires and satisfaction degree for users who used that system. Our analysis indicated that there are 53 studies that used this method while evaluating the performance of systems within these 53 studies.

In order to summarize the results of the used measures based on their types, we illustrated the results of the used types of measures in figure (3.4). Most of the articles that reported one or several measures, where 28.2% (n=22) study used performance

measures, process measures were used 33% (n=26), whereas usability measures used only in 67.9% (n=53) studies.

Figure 12 The distribution of number of article used each types of measure.

3.2.5 Cited Works

We follow the same methods in [4] to analyze regarding reference and what is related with it by compiling all the information within all the 78 studies such as year, source type, and source title. We build a database for reference to analyze the information related with references by using Microsoft access. The number of references found within these studies was one thousand eight hundred fifty-eight (1858). About 13.1 average references in each study, table 8 shows types of cited reference.

Table 7 Type of references

Type	Number	Percent of cited work
Conference	1028	55.32%
Journal	572	30.78%
Book	79	4.25%
Web	45	2.42%
Review	24	1.29%

Workshop	24	1.29%
Tech. Report	11	0.59%
Theses	7	0.37%
Other	68	3.65%
Conference	1028	55.32%

According to the table above, the most dominant cited works were conferences (55.3%), whereas journals in second place with (30.7%) of cited works.

Figure 13 Publication Year of cited work
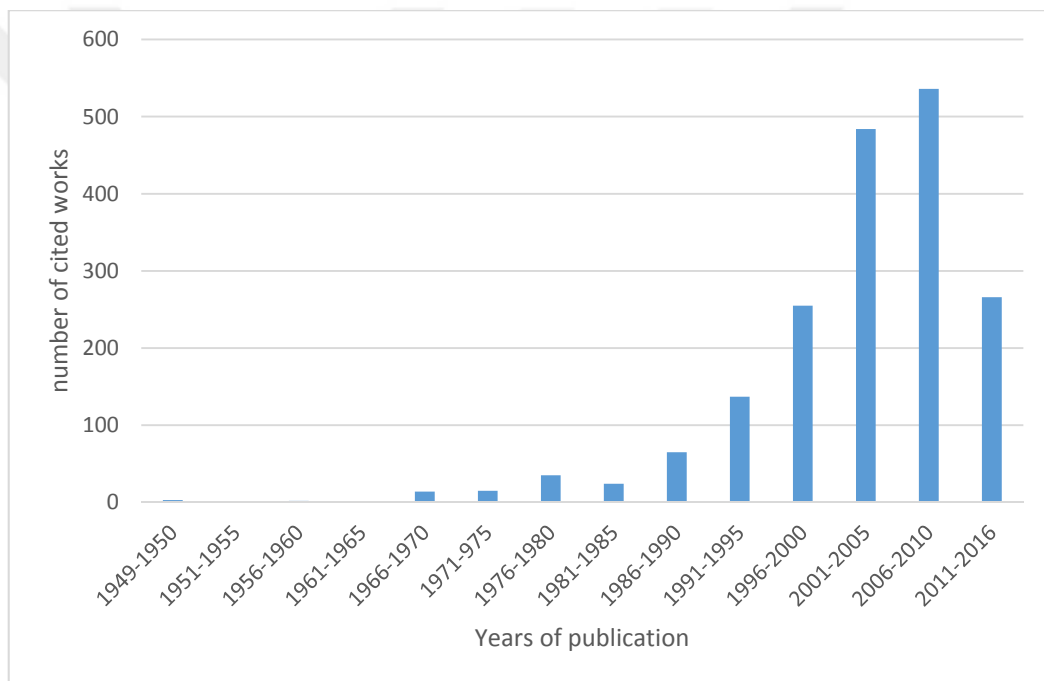
As shown in figure 13 the majority of cited articles relies between (2006-2010), where the highest number (n=129) cited articles were published in 2005 and the same number in 2009.

3.2.5.1 Source Titles

As in study [4] we followed the same procedure to classify our references based on the type of each reference, for instance journals were grouped with journals, webs were

grouped with webs and so on. The table below shows the most frequently occurrence of source where the references published in. We listed several sources with highest number of occurrence. SIGIR came in first place with 280 occurrences in all the references.

Table 8 Top sources by number of citations references

Source	TYPE	Number of Articles Citing	Percent of All Articles
SIGIR	Conference	280	15.06
JASIST, JASIS, American Documentation	Journal	166	8.93
IP&M, Information Research, Information Retrieval	Journal	87	4.68
Information and Knowledge Management Conference	Conference	78	4.19
TREC	Conference	74	3.98
SIGCHI	Conference	49	2.63
CIKM	Conference	49	2.63
ACM Transactions on Information Systems (TOIS) Journal	Journal	47	2.52
WEB	Webpages/Sites	39	2.09
Journal of Documentation	Journal	35	1.88
ACM Transactions on Information Systems (TOIS) Journal	Journal	47	2.52
WEB	Webpages/Sites	39	2.09
Journal of Documentation	Journal	35	1.88

3.2.5.2 Citations of Included Papers

With the assist of Google scholar we listed the number of citations of the included papers. In table 10 we illustrated the most frequently cited papers, in some papers we noticed that many researchers cited from these articles which mean that such articles consist valuable and useful information.

Table 9 Most cited articles

Title	number of citations
Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search	479
A Comparison of Query and Term Suggestion Features for Interactive Searching	101
Affective Feedback: An Investigation into the Role of Emotions in the Information Seeking Process	97
Evaluating Topic Models for Digital Libraries	92
Distribution of Cognitive Load in Web Search	79
Model-driven formative evaluation of exploratory search: A study under a sense making framework	77
A Survey of Patent Users	52
The Dynamics of Interactive Information Retrieval, Part II: An Empirical Study from the Activity Theory Perspective	49
Questionnaire mode effects in interactive information retrieval experiments	48
Novelty and Topicality in Interactive Information Retrieval	45

3.2.5.3 Cited Authors

The total number of the unique cited authors is (2366), about 72.4% (n=1715) of authors received a single citation and about 13.7% (n=325) associated two citations.

Table 10 Most cited authors

ID	Author Name	# of articles citing author
1	D. Kelly	53
2	K. Jarvelin	49
3	P. Borlund	33
4	N. Belkin	30
5	L. Azzopardi	29
6	A. Spink	28
7	J. Allan	23
8	P. Ingwersen	21
9	T. Saracevic	21
10	M. Sanderson	20
11	W. Hersh	19
12	C. Cool	19
13	H. Keskustalo	18
14	R. W. White	18
15	I. Ruthven	17
16	E. Voorhees	16
17	R. White	16
18	S. Dumais	16
19	A. Turpin	15
20	F. Scholer	15
21	G. Marchionini	15
22	N. J. Belkin	15
23	P. Vakkari	15
24	D. Elsweiler	13
25	C. Zhai	13

CONCLUSION

As mentioned previously that our study considers an extension to study [4], where many articles were reviewed to analyze these studies systematically. We concluded that within the last ten years the number of studies used to evaluate interactive information retrieval systems has increased rapidly where 78 studies were included and analyzed in this study. We concluded that conferences became more interested for searchers in such studies were 49 articles out of 78 were included all these papers were published in several conferences and 29 articles were journals whereas in [4] journals were more than conferences. ACM SIGIR retrieved the majority of our relevant sources. In term of measures we noticed that authors in [4] showed that there were no frameworks used to evaluate IIR systems. We found in two studies frameworks were used to evaluate the performance of these systems. In term of datasets in our study we found that authors used several datasets such as (synthetic dataset, OHSUMED3) and the most common used one is TREC. A large portion of the collections utilized was produced to support system-centered search and future work may create collections particularly intended for interactive searching. In term of research tasks, there is still much work to be done to create errand framework for IIR assessments for creating tasks since there is no particular rule to make investigate tasks. The majority of studies were set up as within-subject design, since the between-subject design required more subjects. Most of studies included questionnaires which represent usability measure of IIR which reflect the opinions of subjects of the tested systems. According to the results we found that there is some progress in such studies in term of using frameworks as tools for evaluation, the number of studies within the last ten years. Most of users are primarily from USA and UK even if there are many countries were represented in these studies. Research questions were to a great extent worried about issues of the user rather than the system. Research queries would move the concentration from assessment to interactive standards and perhaps take into consideration more research profundity in the field. The largest

demographic of subjects were university students. The majority of references types were conference and journals since both conference and journals present huge number of articles which give authors more flexibility to gather ideas about their topics.

REFERENCES

1. Robins, David. "Interactive information retrieval: Context and basic notions." InformingSciJ 3 (2000): 57-62..

2. Ingwersen, Peter, and Kalervo Järvelin. "Introduction." The Turn: Integration of Information Seeking and Retrieval in Context (2005): 1-21..

3. Michel, Dee Andy. "What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information." Journal of the American Society for Information Science 45.7 (1994): 498..

4. Kelly, Diane, and Cassidy R. Sugimoto. "A systematic review of interactive information retrieval evaluation studies, 1967–2006." Journal of the Association for Information Science and Technology 64.4 (2013): 745-770.

5. Unterkalmsteiner, Michael, et al. "Evaluation and measurement of software process improvement—a systematic literature review." IEEE Transactions on Software Engineering 38.2 (2012): 398-424..

6. Jorgensen, Magne, and Martin Shepperd. "A systematic review of software development cost estimation studies." IEEE Transactions on software engineering 33.1 (2007)..

7. T. Dyba, E.Arisholm, D.I.K. Sjoberg, J.E. Hannay, F.Shull , "Are Two Heads Better than One? On the Effectiveness of Pair Programming", IEEE Software, vol. 24, Issue 6, (2007), 12-15.

8. Harden, Angela, et al. "Applying systematic review methods to studies of people's views: an example from public health research." Journal of Epidemiology and Community Health 58.9 (2004): 794-800.

9. T. Dyba, E.Arisholm, D.I.K. Sjoberg, J.E. Hannay, F.Shull , "Are Two Heads Better than One? On the Effectiveness of Pair Programming", IEEE Software, vol. 24, Issue 6, (2007), 12-15.

10. Pedreira, Oscar, et al. "Gamification in software engineering–A systematic mapping." Information and Software Technology 57 (2015): 157-168.

11. Higgins, Julian PT, and Sally Green, eds. Cochrane handbook for systematic reviews of interventions. Vol. 4. John Wiley & Sons, 2011.

12. Voorhees, E. and Harman, D. Overview of the eighth Text REtrieval Conference (TREC-8). In The Eighth Text REtrieval Conference (TREC-8), NIST, Gaithersburg, MD, USA, (2000), pages 1–24.

13. Turpin, Andrew H., and William Hersh. "Why batch and user evaluations do not give the same results." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

14. Wilson, Tom D. "Models in information behavior research." Journal of documentation 55.3 (1999): 249-270.

15. Taylor, Robert S. "The process of asking questions." Journal of the Association for Information Science and Technology 13.4 (1962): 391-396..

16. Dervin, Brenda. "Useful theory for librarianship: Communication, not information." Drexel library quarterly 13.3 (1977): 16-32..

17. Belkin, Nicholas J. "Anomalous states of knowledge as a basis for information retrieval." Canadian journal of information science 5.1 (1980): 133-143..

18. Marchionini, G. Information seeking in electronic environments. Cambridge University Press, New York, NY, USA, 1995.

19. Belkin, Nicholas J., Robert N. Oddy, and Helen M. Brooks. "ASK for information retrieval: Part I. Background and theory." Journal of documentation 38.2 (1982): 61-71.

20. Marchionini, Gary. "Information-seeking strategies of novices using a full-text electronic encyclopedia." Journal of the American Society for Information Science 40.1 (1989): 54.

21. Vakkari, P. Task-based information searching. 37: (2003), 413–464.

22. Campbell, Donald J. "Task complexity: A review and analysis." Academy of management review 13.1 (1988): 40-52.

23. Byström, Katriina, and Kalervo Järvelin. "Task complexity affects information seeking and use." Information processing & management 31.2 (1995): 191-213..

24. Borlund, P. Evaluation of interactive information retrieval systems. page 276. PhD dissertation. 2000.

25. Toms, Elaine G., et al. "The effect of task domain on search." Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research. IBM Press, 2003..

26. Bell, David, and Ian Ruthven. "Searcher's assessments of task complexity for web searching." Advances in Information Retrieval (2004): 57-71..

27. Glover, Eric J., et al. "Web Search---Your Way." Communications of the ACM 44.12 (2001): 97-102.

28. Schamber, Linda, Michael B. Eisenberg, and Michael S. Nilan. "A re-examination of relevance: toward a dynamic, situational definition∗." Information processing & management 26.6 (1990): 755-776.

29. Saracevic, T. Modeling interaction in information retrieval. In Proceedings of the American Society for Information Science, volume 33, (1996), pages 3–9.

30. Borlund, Pia. "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems." Information Research. An International Electronic Journal 8.3 (2003).

31. Kuhlthau, Carol C. "Inside the search process: Information seeking from the user's perspective." Journal of the American society for information science 42.5 (1991): 361.

32. Ellis, David. "A behavioural approach to information retrieval system design." Journal of documentation 45.3 (1989): 171-212.

33. Eisenberg, Michael B., and Robert E. Berkowitz. "Information Problem-Solving: The Big Six Skills Approach." School Library Media Activities Monthly 8.5 (1992): 27.

34. Choo, Chun Wei, Brian Detlor, and Dan Turnbull. "Information seeking on the Web: An integrated model of browsing and searching." first monday 5.2 (2000)..

35. Broder, Andrei. "A taxonomy of web search." ACM SIGIR forum. Vol. 36. No. 2. ACM, 2002..

36. Wilson, Tom D. "Models in information behaviour research." Journal of documentation 55.3 (1999): 249-270.

37. Bates, Marcia J. "The design of browsing and berrypicking techniques for the online search interface." Online review 13.5 (1989): 407-424.

38. O'Day, Vicki L., and Robin Jeffries. "Orienteering in an information landscape: how information seekers get from here to there." Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. ACM, 1993..

39. Bates, Marcia J. "Where should the person stop and the information search interface start?." Information Processing & Management 26.5 (1990): 575-591.

40. Belkin, Nicholas J., et al. "Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems." Expert systems with applications 9.3 (1995): 379-395.

41. Saracevic, Tefko. "The stratified model of information retrieval interaction: Extension and applications." Proceedings of the Annual Meeting-American Society for Information Science. Vol. 34. LEARNED INFORMATION (EUROPE) LTD, 1997.

42. Spink, Amanda. "Study of interactive feedback during mediated information retrieval." Journal of the American Society for Information Science 48.5 (1997): 382-394.

43. Lancaster, F. W. Information Retrieval Systems: Characteristics, Testing, and Evaluation. Wiley, New York et al, 1968.

44. Salton, Gerard, James Allan, and Chris Buckley. "Approaches to passage retrieval in full text information systems." Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1993.

45. Van Rijsbergen, Cornelis J. "A non-classical logic for information retrieval." The computer journal 29.6 (1986): 481-485.

46. Spink, Amanda, et al. "Information seeking and mediated searching study. Part 3. Successive searching." Journal of the Association for Information Science and Technology 53.9 (2002): 716-727.

47. Efthimiadis, Efthimis N. "Query expansion." Annual review of information science and technology 31 (1996): 121-187.

48. Attar, Rony, and Aviezri S. Fraenkel. "Local feedback in full-text retrieval systems." Journal of the ACM (JACM) 24.3 (1977): 397-417.

49. Xu, Jinxi, and W. Bruce Croft. "Improving the effectiveness of information retrieval with local context analysis." ACM Transactions on Information Systems (TOIS) 18.1 (2000): 79-112.

50. Terra, Egidio, and Charles LA Clarke. "Frequency estimates for statistical word similarity measures." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

51. Sanderson, Mark, and Bruce Croft. "Deriving concept hierarchies from text." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

52. Beeferman, Doug, and Adam Berger. "Agglomerative clustering of a search engine query log." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

53. Salton, G., and C. Buckley. "Improving retrieval performance by relevance feedback in Journal of the American Society for Information Science." Volume 14 (1990): 288-297.

54. Spink, Amanda, and Robert M. Losee. "Feedback in information retrieval." Annual review of information science and technology 31 (1996): 33-78.

55. Belkin, Nicholas J., et al. "Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience." TREC. 1999.

56. Nichols, D. Implicit rating and filtering. In Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering, ERCIM. (1998): pages 31–36.

57. Oard, Douglas W., and Jinmook Kim. "Modeling information content using observable behavior." (2001): 481..

58. Beaulieu, M. and Gatford, M. J. Interactive okapi at trec-6. In Text Retrieval Conference (TREC) TREC-6 Proceedings. (1998): pages 143–167.

59. Hearst, Marti A. "User interfaces and visualization." Modern information retrieval (1999): 257-323.

60. Bier, Eric A., et al. "Toolglass and magic lenses: the see-through interface." Proceedings of the 20th annual conference on Computer graphics and interactive techniques. ACM, 1993..

61. Declerck, Thierry, et al., eds. Semantic Multimedia: 5th International Conference on Semantic and Digital Media Technologies, SAMT 2010, Saarbrücken, Germany, December 1-3, 2010, Revised Selected Papers. Vol. 6725. Springer Science & Business Media, 2011., Revised Selected Papers (Vol. 6725). Springer Science & Business Media.

62. Malik, S. Interactive Information Retrieval with Structured Documents (Doctoral dissertation, Universität Duisburg-Essen, Fakultät für Ingenieurwissenschaften» Informatik und Angewandte Kognitionswissenschaft). 2010.

63. Borlund, Pia. "Experimental components for the evaluation of interactive information retrieval systems." Journal of documentation 56.1 (2000): 71-90.

64. Ghorab, M. Rami, et al. "Personalised information retrieval: survey and classification." User Modeling and User-Adapted Interaction 23.4 (2013): 381-443.

65. Kelly, Diane. "Methods for evaluating interactive information retrieval systems with users." Foundations and Trends® in Information Retrieval 3.1–2 (2009): 1-224.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: ALYASEEN, Haider
Nationality: Iraqi
Date and Place of Birth: 03 Jun 1978, Baghdad
Marital Status: Married
Phone: +90 537 873 7599
email: haider.sh06@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
Higher Diploma	University of Technology, Computer Science Department	2003
BSc.	University of Technology, Computer Science Department	2001
High School	AL Jazera High School	1997

WORK EXPERIENCE

Year	Place	Enrollment
2010- Present	Administrative and Finance Office/Commission of Integrity (COI)	Assistant Manager
2005-2010	Administrative and Finance Office/Commission of Integrity (COI)	Senior Programmer
2004-2005	General Directorate of Educational Planning/ Ministry of Education	Programmer

FOREIGN LANGUAGES
English

HOBBIES
Reading, Travel, Movies, Sports.