# A Task Completion Framework to Support Single-Interaction IR Research

Anton van der Vegt[1,2], Guido Zuccon[1], Bevan Koopman[1,2], Peter Bruza[1]
[1] Queensland University of Technology, Australia
[2] Australian E-Health Research Centre, CSIRO, Australia
anton.vandervegt@hdr.qut.edu.au, g.zuccon@qut.edu.au,
bevan.koopman@csiro.au, p.bruza@qut.edu.au

**Abstract**

**Purpose:** A conceptual model describes important factors within a system and how they relate to one another. They are important because they help to identify system changes that can yield the greatest improvement. Within Information Retrieval (IR), most research is directed towards multi-document retrieval and a multi-interaction IR (miIR) user scenario. There are few, if any, IR conceptual models supporting minimal or single-interaction IR (siIR) user scenarios, however the need for siIR systems is growing rapidly. This paper takes the first steps towards constructing a task-oriented conceptual model and experimental framework to support siIR research.

**Approach:** A first principles approach is employed to develop a task-oriented conceptual model, called Bridging Information Retrieval (BIR). This model is contrasted with the concept of Relevance, a central factor within IR research.

**Findings:** BIR introduces the central concept of Bridging Information (BI) as the objective of IR systems. BI is the additional information a user requires to complete a task, beyond their innate knowledge. The relationship between BI and relevance is determined.

**Research Limitations:** The theoretical basis of BIR is derived axiomatically, however the resulting system evaluation model is speculative.

**Practical Implications:** The proposed operational framework offers researchers a systematic approach to designing and evaluating siIR systems.

**Originality:** This work contributes a novel task-oriented IR conceptual model and evaluation framework, both centred around the concept of BI for siIR. It also contributes a novel search task classification method.

# 1 Introduction

Accompanying the global proliferation of smart-phones is increased demand for non, or minimal, interaction IR solutions. The small smart-phone screen size hinders complex user interaction that is the norm for desktop internet search. Even in traditional IR environments, such as clinical decision support, the time-pressed clinician doesn't have time to read and process lengthy documents (Sadeghi-Bazargani et al. 2014), but instead requires synthesized summaries or answers (Cook et al. 2013). Allan et al. (2012) recognised at the 2012 SWIRL Strategic Workshop the need for information retrieval over document retrieval as a key challenge facing the IR research community.

However the great majority of IR research lies outside of the study of this kind of minimal interaction IR. Two major Information Retrieval (IR) research areas are Interactive and batch experimentation. Interactive IR system research is typically identified with the study of users operating a search system to fulfil their information need (Kelly 2009) whereas batch-style system testing, such as that employed by TREC programs (Voorhees and Harman 2005), typically evaluates IR systems, without users, on the basis of the relevance of a ranked list of documents, selected by the system in response to a query.

Although significantly different in experimental process, the interactive and batch processes typically share the same intended IR user scenario. This scenario involves multiple user interactions with the IR system, i.e., After the user commences the search with a query (interaction 1), the system retrieves a list of information objects which the user can review (interaction 2), the user can then open one or more linked source documents to read (interactions 3..N) and in some instances provide a new query to continue the search (interaction ...M). In contrast to this multi-interaction IR (miIR) user scenario, is a reduced IR process with minimal interaction: The specification of the query (interaction 1), after which the system returns a single Information Card containing all the information the user requires to fulfil their need. We will refer to this as a single-interaction IR (siIR) user scenario of which good examples are Summarisation (Aslam et al. 2014) and Q&A (Dang et al. 2007).

Since the emergence of the IR discipline, a number of important conceptual research frameworks have been developed to support specific lines of IR research. These models emphasise different aspects of the IR process, including the notion of information relevance (see Mizzaro (1997), Schamber et al. (1990), Saracevic (1975), Huang and Soergel (2013)), cognition and psychology (see Ingwersen (1984), Belkin (1980), Harter (1992)), interaction (see Belkin et al. (1993, 1995)) and integrated seeking and search (see Ingwersen and Järvelin (2005), Saracevic (1996)). However, siIR is considerably different in nature to miIR (see Table 1) and can represent a complete reversal of role responsibilities from a user-lead search to a system-lead recommendation, which puts into question whether the same conceptual IR frameworks are applicable and/or suitable to both types of IR research.

In this article the authors examine a number of important existing IR frameworks and suggest that a more specific model is required for siIR research. In Section 3

| # | System Characteristics | Multiple Interaction IR | Single Interaction IR |
|---|---|---|---|
| 1 | Information retrieval unit | Corpus information unit, e.g. document or web page | Sub-corpus information unit, e.g. text/picture(s) |
| 2 | Intra document selection | No | Yes |
| 3 | Inter document selection | No | Yes |
| 4 | Basis of system performance measurement | Ranked list of documents | Answer Card content |
| 5 | User interface pattern | SERP | Answer Card |
| 6 | Primary responsibility for fulfilling the user's need | User | System |

Table 1: IR System Characteristics that distinguish between multiple interaction and single interaction IR systems

a first principles approach is employed to derive a task-oriented IR framework that distinguishes between siIR and miIR. In addition a novel task classification approach is detailed to provide a framework for research into a more complex range of siIR problems than Q&A. Although considerable research is now conducted into siIR systems, the authors have been unable to find a suitable foundational framework to support these research activities. This paper provides the first steps towards such a framework, which can support a more systematic approach to siIR research including system evaluation and results interpretation.

# 2 Historical Context

In this Section the bias towards multi, rather than single, interaction IR processes is investigated from a research and commercial stand-point. The historical basis for this miIR bias is identified and finally the suitability of relevance, a central notion within IR research, is questioned in relation to siIR.

## 2.1 *miIR Bias in Commercial and Test IR Systems*

In current commercial search engines, a user typically provides a search query and the system generates a Search Engine Results Page (SERP) with a, often ranked, list of linked relevant information sources for the user to decide between, select, read or re-query[1]. The SERP user-interface pattern is inherently multi-interactive by design. It also reflects the uncertainty of the system designers in being able to precisely select the information required to meet the user's information need, as specified by the user's query. In other words, in many instances, miIR is essential to overcome system limitations, or query ambiguity.

Within experimental IR systems research, two major IR system evaluation approaches are premised on miIR:

[1]See for example www.google.com, www.bing.com or domain specific services such as PUBMED

1. Interactive IR system testing, which explicitly targets a miIR user model. In this model, users are provided with search tasks and their interaction with the IR system is recorded and evaluated (see Kelly and Sugimoto (2013), Over (2001)), and;

2. Batch-style system testing, which originates from the ASLIB Cranfield Project (Cleverdon 1960, Cleverdon et al. 1966), implicitly targets miIR. In this approach, and that of subsequent research groups (Voorhees and Harman 2005, Tsikrika et al. 2013), IR systems are typically evaluated on their ability to generate a ranked list of documents that are relevant to a notional user's need. Although the user has been abstracted out of the test process, the list of documents as output and evaluation based on more than just the first document, reflect an *intended-multi-interaction* user scenario, where a user has to select and read through one or more documents in order to fulfil their information need.

Table 2 uses the TREC program to exemplify the variety of IR systems experimentation; most tracks can be classified as either miIR or intended-miIR.

| **TREC Program** | **Evidence** | **Example citation** |
|---|---|---|
| **Multi-Interaction IR:** | | |
| Interactive | Real users employed | Over (2001) |
| ciQA | User interaction supported | Dang et al. (2007) |
| Filtering (Adaptive) | User feedback on relevance | Robertson and Soboroff (2002) |
| Dynamic Domain | User feedback on relevance | Yang et al. (2015) |
| Open search | Real users employed | http://trec-open-search.org/ |
| Hard track | User clarifying form | Allan (2003) |
| **Intended-Multi-Interaction IR:** | | |
| Ad-hoc | Ranked list of documents | Harman (1995) |
| Session | Ranked list of documents | Carterette and Hall (2013) |
| Web,non-interactive | Ranked list of web pages | Craswell et al. (2003) |
| Novelty | Ranked list of sentences | Soboroff and Harman (2003) |
| Genomics (primary) | Ranked list of documents | Hersh and Bhupatiraju (2003) |
| Robust | Ranked list of documents | Voorhees (2003) |
| Enterprise | Ranked list of emails | Soboroff et al. (2006) |
| Precision Medicine | Ranked list of articles | http://trec-cds.appspot.com/2017.html |
| Hard track | Ranked list of passages | Allan (2003) |
| Contextual Suggestion | Ranked list of suggestions | Dean-hall et al. (2014) |
| **Single Interaction IR:** | | |
| Genomics (secondary) | Produce a GeneRIF | Hersh and Bhupatiraju (2003) |
| QA | Precise factual answer | Dang et al. (2007) |
| Complex Answer Ret. | Synthesized knowledge article | http://trec-car.cs.unh.edu/ |
| Live Q&A | Answer response | Agichtein et al. (2015) |
| Temporal Summarization | Sentence selection | Aslam et al. (2014) |

Table 2: TREC lines of experimentation and their targeted interaction bias

## 2.2 *Historical Emergence of the miIR Bias*

As the discipline of IR emerged in the 1950's, the bias towards miIR was implicit in both the objectives of the discipline and the measures of system success. Luhn clarified the objective; to find " *documents within a collection which have a bearing on a given topic*"(Luhn 1957, pg309) and Kent et al. (1955) defined the duel measures for IR system effectiveness: Recall and pertinency factor (today known as precision), which are still in use today, although often in some derivative form.

Both of these measures, like the objective, encourage multi-document retrieval, except where precision @1 is set as the sole measure of performance - a rare test case. Multi-document retrieval presumes a miIR user scenario in which users must select which documents to read and must search within a document or across documents to find relevant information to fulfil their need. In contrast, had the bias been towards siIR, the objective of IR and measures of success would be aligned to the retrieval of a single piece of information to resolve the user's need; something Brookes prophesised of 3 decades later, "*The day will come when present documentary data bases become real information systems offering their users information directly rather than lists of documents to be located and read*" - (Brookes 1980, pg5).

The batch-style, experimental work of (Cleverdon 1960) applied these measures and became known as the system oriented approach to IR. From the 1980's onwards various alternative approaches to IR experimentation and research arose, however although each framework or model emphasised different factors of the IR process, the targeted miIR user scenario was generally presumed or reinforced. Provided here are some important historical models to support this assertion.

Often referred to as the counter-point to the system oriented approach is the user-oriented approach, proclaimed by Schamber et al. (1990). This model emphasises the user, their dynamic state of cognition, their multi-dimensional assessment of relevance and the interactive nature of the search task within the broader information seeking process. The user-oriented approach studied real users in multi-interaction search scenarios to better understand the IR process.

Cognition and interaction have their roots in the early 1980's when Belkin (1980) and later Belkin et al. (1982) report on a cognitive perspective of IR called Anomalous State of Knowledge (ASK). In this hypothesis an information need arises when a user recognises an anomaly relating to their knowledge concerning a situation. The (interactive) IR system's purpose is therefore, to resolve these anomalies. Ingwersen (1984) explored the psychological and cognitive nature of information seeking and highlighted its iterative and interactive nature while Harter (1992) re-oriented the psychological relevance work of Sperber and Wilson (1986) to the problem of IR and reinforced the dynamic and changing context of the human mind as new documents are processed.

Information search interaction has its roots in library search and information seeking, potentially with expert intermediaries (Ingwersen 1984). Information search that is focused on the use of IR systems as the intermediaries, is a branch of IR called Interactive Information Retrieval (IIR). Belkin considered IR as an, "*inherently interactive*

*process*" (Belkin et al. 1993, pg325) and sought to better understand its interactive nature. In particular he explored and classified user information seeking strategies, through custom built IR system interfaces, such as BRAQUE (Belkin et al. 1993) and Merit (Belkin et al. 1995), the idea being to capture the user's goal/intent and support their interactive search 'dialogue'; a dialogue between the user and the IR system.

Ingwersen and Järvelin (2005) attempted to integrate information seeking and information retrieval (IS&R) into a single framework. Their conceptual framework is, "*founded on the holistic cognitive viewpoint*"(Ingwersen and Järvelin 2005, pg259), meaning that the individual information seeker's perception is the focus of the model. The user, or cognitive actor, is central, however the nested contexts within which the user operates are carefully identified as components of the model. Although depicted as a static model (Ingwersen and Järvelin 2005, Figure 6.1), interaction is referred to as a *vital process*, and it is broken down into types, including short-term, session-based and longitudinal.

## 2.3 *Multi Interaction IR Bias and Relevance*

The notion of relevance is central to IR (see Saracevic (2016)). The concept was incorporated early on by Cleverdon (1960), Cleverdon et al. (1966) within the document relevance assessment step of their IR system evaluation methodology. Saracevic (1996) proposed a relevance framework called the Stratified Model of IR Interaction. As the name suggests, user interaction was central to this framework and is represented as one of the *general attributes* of Relevance. Like Belkin, Saracevic saw the IR interaction as a dialogue between participants - user and 'computer'(Saracevic 1996, pg9).

Within a miIR user paradigm, finding relevant information for a user's query is a well acknowledged and intuitive objective for an IR system, however if the user's information need demands a single answer, and the system is capable of such an answer, then in this siIR user scenario providing a relevant answer is insufficient and it is no longer intuitive. A user asking '*What is the deadliest disease in the world?*' expects a single answer response (*Coronary Artery Disease*), not a set of documents containing that information nor even a single response that is relevant, for example information on heart diseases. This distinction suggests two possible conclusions: Either (1) for the single response 'type' of information needs and 'Answer response' is a manifestation of relevance or (2) An 'answer response' is a distinct notion to a 'relevant response'. In either case, a more nuanced or new model of understanding surrounding'Answer responses' would be highly beneficial to the study of siIR.

**In summary,** the bias towards researching miIR is found across virtually all of the major research movements and frameworks in IR's short history. This is not surprising given that information needs suited to siIR represent a small proportion of the infinite array of human information needs. Also, early in the history of the IR discipline, technical and system limitations would have further limited the set of needs that could have been resolved in an siIR user scenario. However, with the advancement in both

system power and technical capability, the problems that can now be resolved directly are significant and this, together with a growing demand for siIR solutions, present considerable impetus to warrant an IR framework that distinguishes between and supports siIR research.

# 3 A Task-Based Conceptual Model of IR

In this Section we construct a first principles conceptual framework for IR built upon a task-oriented perspective. A task-oriented view is not new to IR research, as discussed in Section 3.3.3. Incorporating task within the conceptual framework may support a classification of search tasks on the basis of a requirement for interaction: something that may expose the factors that determine the nature of IR interaction and therefore help researchers to distinguish between minimal and multi interaction search tasks.

The purpose of a conceptual framework, according to Engelbart and English (1968) is to orient the important factors within a system and document how they relate to each other so that one can deduce the types of changes within those factors that might yield the greatest performance benefits and therefore the most productive lines of research. In particular, the authors are interested in understanding:

- Whether the IR process can be cast from a task-oriented perspective
- How the conceptual factors change when the IR process shifts from interactive to non-interactive IR problems
- How such a conceptual model can better support the siIR user scenario

There is currently no model that specifically delineates between siIR and miIR. Thus we construct a conceptual framework from scratch. We call this process a first principles approach, which consists of: (1) Identifying the foundational elements of a task-oriented IR approach; (2) Validating the elements against accepted IR thinking and practice; (3) Using this foundation, establishing whether miIR and siIR can be differentiated. This approach, originating in the field of Philosophy by Aristotle (Graham 1999), has widespread use in Mathematics (see the axiomatic method of Potter (2004)) and Physics and Chemistry (see the ab initio method of Navrátil et al. (2016)).

## 3.1 *Elements of the IR Problem*

What is the underlying problem that information retrieval tackles? The problem can be constructed from its elemental constituents and axiomatic assumptions:

1. Information: All verbal, written, digital, pictorial forms of information.
   (a) Assumption 1.1: There exists a quasi-infinite quantity of information available in the world.

2. Person: A natural person.

   (a) Assumption 2.1: A person has a changing and limited bank of innate knowledge.

   (b) Assumption 2.2: A person has a limited cognitive capability to process information.

   (c) Assumption 2.3: A person has limited time to perform any task and such limitations may be self or externally imposed.

3. Task: The broadest sense of a task performed by a person. It may be self-imposed or set by others, mental and/or physical. It includes activities performed for work and/or play.

   (a) Assumption 3.1: Tasks require knowledge to complete.

   (b) Assumption 3.2: Tasks are time bounded, i.e., no individual task is everlasting.

## 3.2 *The IR Problem Statement*

To complete a *task* that requires a *person* to draw upon more knowledge than they currently possess will often not be possible because it will take more time than they have, or the *task* requires, to manually search and process the available *information* to provide them with the missing knowledge they need to complete the *task*.

The problem, as stated above, is essentially one of time and human limitations. One can imagine a person sitting at a desk. On the one hand there is a written task to perform and on the other hand there is an imposing pile of unordered documents including much of humankind's written and pictorial record. Upon reading the task the person realises there are some things she does not know so she starts reading the documents to her right, one-by-one until the knowledge she needs is found. In all but the most fortunate cases, the person will never find the knowledge and the task will never be completed — therein lies the central IR problem.

## 3.3 *Elements and Assumptions Discussion*

### 3.3.1 *Element 1: Information*

The first element of the problem is information, taken as meaning all accessible written and pictorial data stored digitally or printed to a medium. Assumption 1.1 highlights the scale and ever growing quantity of information that is accessible today through the Internet, libraries and corporations. It is self-evident that from a human perspective the amount of information available for reading and processing is effectively infinite.

### 3.3.2 *Element 2: Person*

The consideration of human cognition as a key, and sometimes central, element within the IR and information seeking processes has long been recognised. Earlier, in Section 2.2, some of the important cognitive IR research was mentioned, including the ASK hypothesis of Belkin (1980) and the psychological and cognitive considerations of Ingwersen (1984) and Harter (1992). From an information seeking perspective, Dervin (1983) proposed a sense-making approach in which a person, in the context of a problematic situation, can only progress by seeking information to bridge their cognitive gaps. Kuhlthau (2004) documented a staged information seeking process in which the user constructs their own understanding of the task and answer during the search process. The answer construction process is highly subjective, based on the user's initial cognitive constructs. In the first principles framework, cognition is represented by the 'person' model element and is a similarly central factor.

Assumptions 2.1 and 2.2 are rooted in cognitive science. The *limited bank of innate knowledge* is a reference to people's limited memory, in particular long term declarative memory which is the memory a person draws from consciously and intentionally (Cohen and Squire 1980). Tulving and Donaldson (1972) delineated a number of types of memory of which *semantic* memory relates specifically to the storage of general knowledge. According to Bulletin and Voss (2009) there is no current answer to the question of how much information can be stored in memory, however despite this, it is currently self-evident that people are currently unable to store and access unlimited semantic memories.

Assumption 2.2 is grounded in the "*information processing*" metaphor of Miller (1956) whereby all biological organisms, including humans, are limited in how much information they can process at any point in time. The final assumption (2.3) is a more generally self-evident assumption that for a human, time is always limited, if not by the length of one's life, as measured in time, but more usually by self imposed time limits such as sleep or having other tasks to attend to, or by external time limits such as those imposed by work, or family.

### 3.3.3 *Element 3: Task*

In the first principles approach, we employ a natural definition of task, defined by Vakkari (2003) as: "*an activity to be performed in order to accomplish a goal*". This subsumes the work task definition of Li and Belkin (2008) as it includes any task, personal, work or otherwise. This approach is in-line with Järvelin (1986) who asserts that work tasks are the central driver for IS&R activity and then Ingwersen and Järvelin (2005), who extend this to include all tasks. In the detailed Sections to follow, Task is confined to a search task level to correspond with the operational level of IR systems. We apply the definition of Li and Belkin (2008) that a search task particularly employs the use of an IR system to locate information.

In IR, a user's task is usually considered a contextual or situational factor that affects relevance (Schamber et al. 1990, Saracevic 1996, Huang and Soergel 2013). The

first principles model asserts task as an independent element rather than a contextual factor, or rather than being incorporated into the user element as a perception of task, as asserted by Ingwersen and Järvelin (2005). This is because tasks can be defined and task completion assessed independently to a user. It is, therefore, an independent factor within the IR problem.

Reid (2000, 1999) proposed a similar task-oriented approach incorporating task as an independent variable. Reid's framework incorporates the concept of document task relevance, as viewed by the user (task performer) after task completion, and an additional evaluation criteria, called information value, which is a measure of document relevance, also assessed by the user, after feedback from the task setter. This latter measure represents the notion of the contribution, of the information, to the task outcome. In Reid's later explication of the operational framework, information value is dropped and document relevance assessments are performed by the user after both task completion and external feedback. By having multiple users perform the same task, document relevance is weighted for a test collection and then standard precision and recall measures utilised within a standard batch evaluation approach. The centrality of task is common to the model proposed here. However, Read's proposed incorporation of task is realised through the evaluation of documents for task-relevance by users, i.e. how helpful each document is towards completing the task. In this sense, task is not independent to the user and it is accounted for by the situational manifestation of relevance explicated by Saracevic (1996). This is significantly different to the handling of task in the first principles model, as detailed in Section 3.4.

With respect to assumptions 3(a) and 3(b), it is self-evident that tasks require knowledge to be completed, whether innately or externally sourced and it is also self-evident that all tasks are time limited for an individual.

## 3.4 *The role and Objectives of the IR System*

The IR system is the 4th element in the first principles model. The IR system is typically depicted as an interface between the user and the available information objects (corpus). The user enters their request into the IR system which in-turn executes search algorithms to select the most relevant information, as assessed by the system and relative to the user's request, from the corpus. The IR system then returns information objects, typically ordered in some way, such as by decreasing relevance. The user can preview these information objects and select those to read in full. The process, as defined here, is interactive in nature.

Instead, in this first principles model, the IR system represents the technological response to the IR problem, as articulated in Section 3.2. The *role of the IR system* is therefore:

For a given *task*, to select Bridging Information (see below) for the *person* so they can complete their *task* within a suitable timeframe.

Where Bridging Information (BI) is information, selected from the corpus, that provides

the additional knowledge, or a means to deduce the knowledge, the person requires over and above their innate knowledge to complete the task.

### 3.4.1 *Bridging Information and Relevance*

In conventional IR, relevant information is the target of search. Re-setting the objective of the IR system to bridging information requires clarification and explanation.

The defining characteristic of BI is that it must enable task completion. Although information in the corpus may be stored as discrete files or documents of data, BI is best considered as a, possibly ordered, set of statements or paragraphs or images, each of which are called BI elements and each of which may be sourced from the same or different files or documents. Together, this composite set of elements represents the BI that enables a user to complete their task. The sequence of elements is important if it effects the user's ability to complete the task.

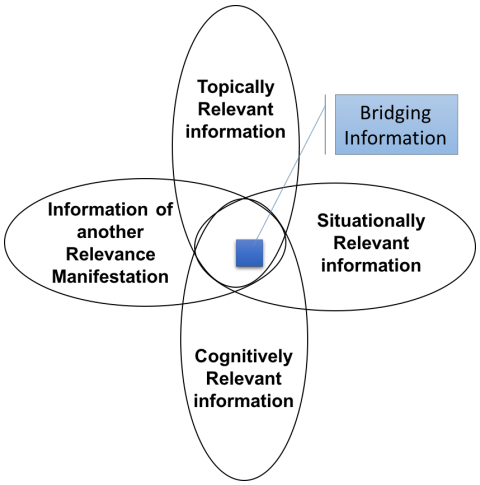


Figure 1: Venn diagram depicting the information objects, within a corpus, and how they relate to each relevance dimension for a given user and task. Bridging information is depicted as a, possibly ordered, subset of information objects within the intersection of information objects across all relevance dimensions.

How does BI relate to relevant information, theoretically? This question will be addressed in the context of Saracevic's Stratified Model and in particular his discussion of manifestations of relevance (Saracevic 1996, pg12). In Saracevic's model, each relevance manifestation represents a relation to the information object. Is BI an existing or new manifestation of relevance or something else entirely? If we propose a new hypothetical manifestation, called Bridging Relevance, then the relation is between each BI element and the knowledge the user needs to acquire to complete their task. The criteria by which Bridging Relevance is inferred is task completion. As created, Bridging Relevance is a composite manifestation of both cognitive and situational relevance. This leads us to suggest that rather than being a new dimension (manifestation) of relevance, it is actually a specific pattern or instance of relevance, i.e., Bridging Information is a specific subset of the intersections of a number of dimensions of relevance.

Figure 1 demonstrates this theoretical understanding of BI. For a given task and user, each petal of the flower represents the set of information retrieved that contains

the relation of that relevance manifestation. So for example the 'situational' petal contains all information objects that relate to the task, and the 'cognitive' petal contains all information objects that relate to the user's state of knowledge and cognitive information need. The centre of the flower represents the intersection of each dimension such that information objects within the intersecting area are related from a task and cognitive perspective (and topical, etc). BI represents a subset of this intersection, delineated by the additional criteria that the set of BI elements must also enable task completion. Information may be useful, novel to the user (i.e., not seen before) and topical, however these relations alone do not guarantee task completion. In addition the BI elements may be required to be presented in a specific sequence in order to enable task completion.

### 3.4.2 *Practical Limitations of BI Retrieval Systems*

There is also an important practical constraint for BI Systems. A retrieval system has no way of knowing the contents of a user's cognitive resources in order to assess the gap it must bridge with the knowledge required for the task. Therefore, a practical assumption of BI retrieval is that a common set of BI can be provided for a group of people such that each person within the group, when provided with the same BI, can complete the same task. It is a fair assumption that underlies all instruction sets targeting an audience greater than one, e.g., board game instructions.

The segmentation of users into groups, which are aligned to the completion of specific tasks, is a central premise for the design of successful BIR systems, as discussed next.

### 3.4.3 *The Objectives of BI Retrieval Systems*

To establish the objectives of BIR systems, consideration of system performance is required, which is typically measured in terms of effectiveness, i.e., the ability of the system to achieve its intended purpose, and efficiency, i.e., the throughput of the system per unit of both time and work (see Kelly (2009), Kelly and Sugimoto (2013), Hornbaek (2006)). For BIR systems, effectiveness is measured by task completion and the size of the user group able to complete a given task. The population size is a factor because higher effectiveness is indicated if a broader user population is able to complete the same task with the same BI. For example, given a medical search task, it is more effective if 10 physicians and 10 nurses can complete the same task when provided with the same BI, than just the physicians. Efficiency is measured by the time and cognitive load it costs users within the user group to complete the task. Both efficiency and effectiveness are functions of the:

1. Number of BI elements selected: To resolve a task, there exists a minimum set of BI elements, below which no user can complete the task. Adding extra elements beyond this minimum may increase the opportunity for other users to also complete the task, however it is also likely to decrease efficiency

2. Representation of the BI elements: BI elements can be any digital representation that conveys information including sentences, paragraphs, pictures, graphs and diagrams

3. Sequence of the BI elements: The correct ordering of the BI elements for human processing is likely to improve system performance for more complex tasks

Incorporating system efficiency, effectiveness and practical system limitations, the BIR system objectives can be clarified: For a given search task, to provide the best available bridging information to enable task completion for the largest user group possible, in the shortest time possible and requiring the least cognitive load for the users.

## 3.5 *Distinguishing Multi from Single Interaction IR*

In the first principles framework, the role of Bridging IR (BIR) systems is to retrieve BI to enable users to complete their tasks. It is at this point in the derivation of the conceptual model that the level of user interaction can be clearly distinguished on the basis of the nature of the task, i.e., whether users want and/or need to perform multi-interaction user search in order to complete the task.

### 3.5.1 *Requirement for Multi-Interaction Search*

There are some tasks that require multi-interaction search for resolution because the search activity is integral to the user's task completion process. Exploratory search is an example, typified by the Berry-picking principle of Bates (1989). In this model an information seeker selects promising information for their original need and upon processing the information identifies new ideas and potentially reformulates the original need and takes new seeking directions. Using this interactive and somewhat serendipitous approach to information seeking, the task is able to be completed. It is the inchoate nature of the task that is critical in these cases, and for such tasks, BI cannot be formulated until later in the seeking process when the task definition becomes clearer. Therefore, initially at least, these tasks are unsuitable for siIR.

Task clarity is often associated with task complexity. Campbell (1988) developed a task complexity classification model incorporating a number of attributes of complexity including the presence of uncertainty. Campbell identifies four major task classifications, of which *Fuzzy tasks* are examples of those that require multi-interaction search because the outcome is unclear at the start of the task and there is, "*minimal focus for the task-doer*"(Campbell 1988, pg48). To complete such a task, the task-doer must firstly clarify the outcome by exploring the possibilities.

In relation to task outcome, Byström and Järvelin (1995) propose a task complexity classification scheme based upon apriori determinability of the: (1) information need (task inputs); (2) process and (3) result(outcome). The assertion of determinability is a subjective assessment by the user. Factor (3) directly correlates with tasks that require multi-interaction search because until the user knows what is required of the task, i.e.

the task outcome, the task cannot be resolved. Factors (1) and (2) do not imply a requirement for multi-interaction search because in these cases the task outcome is known, but the user may not be aware of the inputs or process required to complete the task. Kuhlthau (2004) similarly identifies this distinction between complex search tasks that generate confusion and uncertainty for users at initiation and more routine search tasks, more akin to Q&A that avoids the need to construct the answer through the seeking process.

A second class of tasks that require multi-interaction search for resolution are those where personal preference decisions are needed for their resolution. Examples of such tasks include the search for recipes or home appliance selection, where personal tastes or criteria weightings are required throughout the task completion process. For these tasks, interactive selection and personal judgement of information is essential. Without access to such personal preferences, siIR systems are unable to provide an appropriate decision in order to select the correct BI for the user without further user interaction.

### 3.5.2 *Desire for Multi-Interaction Search*

Bates suggested that, "*There are times when many people want to do their own searching*"(Bates 1990, pg575), in response to what she saw as the general direction of IR at that time towards a fully automated search process. There are tasks for which the user desires multi-interaction search, and therefore the retrieval of a single set of BI is an inadequate solution for the user to complete the task, e.g., search for entertainment, such as looking up available movies in the area. A BI retrieval system could provide a list of available movies, but a user may prefer to browse a number of movie review sites. Conversely, there are many search tasks for which the user is not interested in performing the search themselves. For these tasks a multi-interaction search process is burdensome, costing time and mental resources.

### 3.5.3 *Task Classification*

Figure 2 depicts a task classification matrix based on the binary user factors: requirement and desire for multi-interaction search. Using this model, search tasks can be classified into three groups.

1. Hunting search tasks are those for whom users *must* perform multi-interaction search because the search is integral to the user's task completion process. This can arise because: (a) The task itself is unclear and interactive search is required to define the task before completion is possible, e.g. exploratory search, or; (b) Task completion requires input of personal preference decisions within the interactive search process, e.g. selecting an appliance for purchase

2. Entertaining search tasks are those for whom users *want* to perform multi-interaction search, because of the entertaining or stimulating nature of the search process, rather than because of any necessity to perform interactive search, e.g. browsing

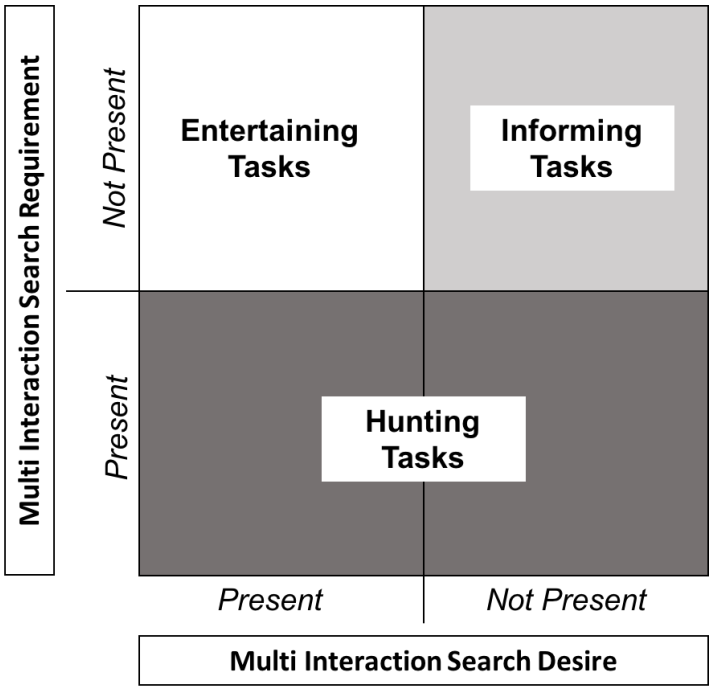


Figure 2: Task Classification Matrix. Classification into 3 groups based on two binary factors: (1) the need for multi-interaction search to complete a task and (2) the user's desire for multi-interaction search to complete a task

3. Informing search tasks are those for whom users neither want nor must they, perform multi-interaction search in order to complete the task. Examples include word definitions, how-to procedures, Q&A, problem solving and other search tasks where the outcome is clear beforehand.

**In summary,** siIR is confined to the resolution of Informing search tasks whereas miIR is suited to Hunting and Entertaining search tasks. Figure 3 depicts the process for classifying a search task and selecting a suitable mode (multi or single interaction) of IR operation. The focus of the remainder of this paper is the single-interaction branch of BIR (siBIR).

# 4 Single Interaction BIR Experimental Framework

In this section elements of an experimental framework for siBIR is proposed. An experimental framework can help the researcher to consider key decisions in the experimental process. The framework described here targets siBIR system evaluation, i.e., the retrieval of an Answer Card to enable a user group to complete a task. The framework addresses the following questions:

1. What is the search task and user group?
2. How do you derive the BI for each search task?

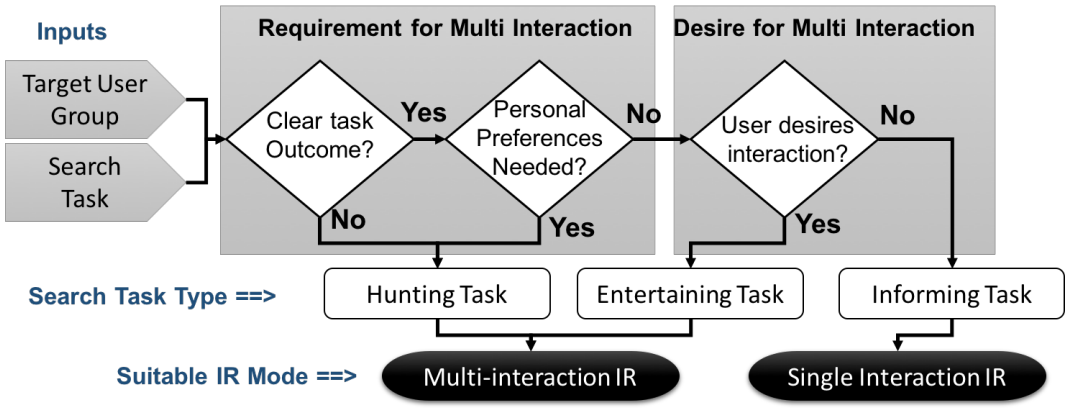


Figure 3: Task Classification Process. Process flow diagram showing the process inputs, decision points, classified search task outputs and suitable mode of interaction for the IR system

3. What measures are suitable for system evaluation?

## 4.1 *What is the Search Task and User Group?*

Prospective search tasks and user groups can be checked for suitability using the classification process proposed in Figure 3. This confirms the task is an informing search task, and therefore suited to siBIR.

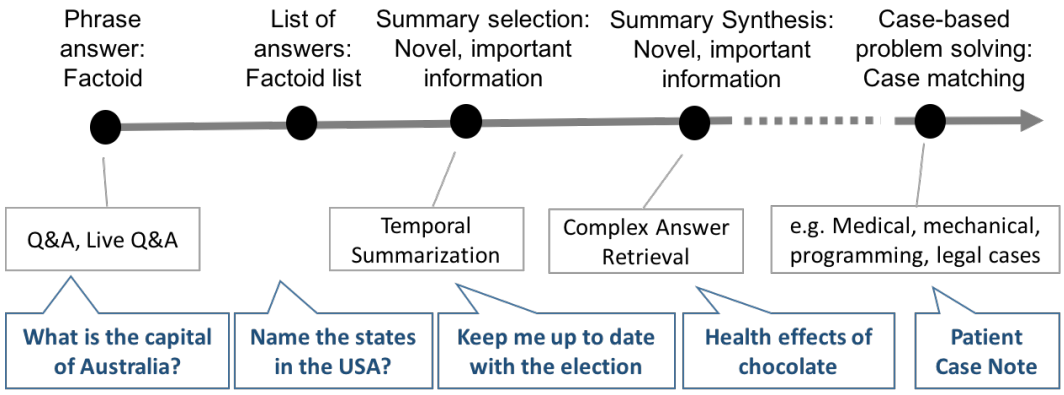


Figure 4: Informing Search Task Complexity Continuum. Moving from left to right, task complexity increases. Above the line are listed specific types of tasks and below the line are the corresponding examples of each type to task and the below this a query example. The dotted line denotes a shift from the present to the future.

Although historically there has been a bias towards miIR experimentation (see Section 2), considerable research has also targeted siIR, some of which is listed in Table 2. By applying the complexity framework of Byström and Järvelin (1995), the apriori determinability of task inputs and process are helpful indicators of the complexity of the informing search tasks. We can use these factors and intuition to place existing siIR search tasks on a continuum of complexity, here depicted in Figure 4.

**Extending Search Tasks to Problem Solving** The authors posit that problem solving tasks are a good fit for informing search tasks and a logical next step in the

evolution to more complex siIR research. A problem has a clear outcome: Resolution, and in many cases a person does not necessarily want to perform the search, they simply want a solution. On the other hand the inputs required and process necessary to resolve the problem may not be apriori obvious to the person, suggesting it is a complex informing search task. This is where an IR system can fill the necessary knowledge gap to suggest the inputs and provide the process.

A problem is defined as, "an intricate unsettled question or a source of perplexity, distress, or vexation"[2]. This is a broad definition, and clearly not all problems are suitable candidates for resolution by IR, however one class of problems, notably case-based problems, may be appropriate. Case-based problems present with a sample case exhibiting non-ideal or malfunctioning behaviour. Many important and prevalent classes of case-based problems exist, including medical diagnoses, programming bugs, legal cases and equipment failure. Resolution of case-based problems can often be resolved by finding similar historical cases that have already been resolved or by looking to best practice and technical documentation, both of which require information retrieval.

A relevant example of a potential medical case-based problem resolution IR system is by Goodwin and Harabagiu (2016) who developed a system to firstly diagnose a case, called *Answer Discovery*, before retrieving suitable Medline articles. The system was evaluated on the generation of a ranked-list of medline documents, i.e. an intended-interactive use-case. The system produced inferred average precision results 40% higher than state-of-the-art solutions for the TREC 2014 CDS track(Simpson et al. 2014). Yet perhaps the true value of this system was its ability to generate answers directly for clinicians, without exploratory search.

**Search Task Representation** Referring back to Figure 4, as the complexity of the search task increases, the representation of the task, as query, may also increase to accurately capture the need. In the case of phrase/list answers, regular questions suffice, e.g., *List of states in USA*, whereas at the other extreme, case-based problems may require the entire case as query, e.g., In the medical domain, the patient admission note can be taken directly as the query, as required in 2016 TREC CDS track (Roberts et al. 2016).

## 4.2 *How Do You Derive the BI for Each Search Task?*

For siBIR, an Answer card is retrieved, rather than documents, and the basis of evaluation is whether the user can complete the search task, using just the Answer card. BI enables task completion, so therefore, all BI elements must be present on the Answer card. Deriving the BI elements is the most important step in the experimental process, as the output of all systems will be evaluated against it.

No matter which method is used to derive BI, the underlying principle remains the same: For a given search task, when provided with the minimum BI, any member of

[2]https://www.merriam-webster.com/dictionary/problem

the target user group must be able to complete the search task, without further search interaction.

**Factoid Q&A BI** is the simplest siBIR case. The search task for the user is to be able to provide the answer to a specific factoid question. In the TREC 2007 Q&A track (Dang et al. 2007), there is just a single BI element, the answer phrase, although the systems were further tested on the provision of evidence (source document) as well. The provision of evidence is further discussed below. Because the answers are non-ambiguous, known facts, the BI is straight-forward to develop.

**Summarisation BI** is considerably more complex and introduces a number of issues. The TREC Temporal Summarisation track (Aslam et al. 2014) is used as the working example. In this example the search task is to provide a monitoring system for an event, providing new and important updates relating to the event as they arise. The track employs the concept of information 'nuggets', which are *atomic pieces of information relevant to the query.* Are nuggets the same as BI? To answer this question, the task definition must be examined.

The difficulty with this track is that task completion is ambiguous, i.e. the task outcome, is unclear. The search task is to produce a summary, yet the definition of the summary may differ for different user groups. One group of users may want essential updates only, and others may want as much information as possible. The track organisers resolved this ambiguity by classifying the nuggets into (a) those of key importance and (b) those of any importance. This is similar to other tracks using nuggets graded as 'vital'/'non-vital'. Unlike nuggets, BI is indivisible. If the minimum set of BI is present, then task completion is enabled, otherwise it isn't. In this particular case, the difficulty of task definition can be resolved by splitting the search task into two: One for essential-information users and one for detail-information users. When this happens the BI is equivalent to either only vital nuggets or all nuggets, respectively. Using a graded system evaluation allows systems to perform well despite meeting the needs of neither target user group, i.e., by providing some vital and some non-vital nuggets. Using a BI approach will clarify which systems meet the needs of each target user group.

Grading is also used for other purposes. Within the TREC 2007 Q&A track (Dang et al. 2007), a five level graded judgement was applied to each test system response. Between not-correct and correct were three extra grades: (1) 'Not supported' indicated that the source document provided with the answer did not support the answer; (2) 'Not exact' meant that the correct answer was provided although it may have been missing a bit or had extra bits and (3) 'Locally correct' meant that the answer was correct, but the assessor felt a better, contradictory answer existed elsewhere in the corpus. From a BIR perspective, each of these grades represents a specific, independent issue, rather than a continuum of performance. If the search task demands evidence (grade 2), then both the answer and the evidence are BI elements to enable task completion. Evaluation measures would then capture this failing. Similarly item (2) either represents missing

| **Performance Factors** | **System Effectiveness Measurement** | **System Efficiency Measurement** |
|---|---|---|
| IIR Performance | Completion | Ideal Interactive Efficiency |
| BI selection | Are all BI elements present including direct task-completion elements and evidential elements? | |
| BI representation | | How easily can people process the format (text/image/presentation) of the BI? |
| BI sequence | Is the BI depicted in the right order? | |
| BI duplication | | Are BI elements repeated? |
| Non-BI | | Is there extraneous information? |

Table 3: Measurements required for evaluating the efficiency and effectiveness of siBIR systems

BI elements or extraneous non-BI, both of which can be captured through independent measures. Grade 3 reflects uncertainty surrounding the task. Is a local answer required or a global answer? These are separate tasks and provision of the alternate answer should not indicate a level of success because the user will not be able to complete their task.

Whether to grade system responses is a key question within the experimental IR process. Within the siBIR framework each BI element is essential for task completion, otherwise the information is not BI. The need for grading can be removed by clarifying either (1) the task definition including the user group or (2) the use of independent measures to capture different modes of failure.

**Case based problem solving BI** may need to be derived, rather than looked up. For example to develop the BI for a medical diagnosis based on the provision of case symptoms, may require expert search first, to initialise the gold standard BI for that search task. For case-based problem solving tasks, it is quickly apparent that evidence is often an essential accompaniment to the BI, so in the case of a diagnosis determination, how that diagnosis was derived will also be essential if other people in the user group, i.e., other clinicians, will be prepared to accept the system response.

#### 4.2.1 *What Measures are Suitable for siBIR System Evaluation?*

BI systems are evaluated for effectiveness and efficiency with consideration towards the influencing factors(Section 3.4.3). Table 3 summarises the measurements required. Note that for siBIR systems, the Answer card is the basis of evaluation.

**Interactive IR Measures** Two key overall IIR performance measures applicable to siBIR evaluation are identified by Kelly (2009), Hornbaek (2006): (1) Completion (system effectiveness) is binary task completion, task completion accuracy or expert

assessment of the quality of outcomes, and; (2) Completion time (system efficiency) for a given task. Completion is a percentage with optimal value of 100% and can be measured and aggregated across tasks and users, but completion time is not a percentage. To enable comparison between search tasks and systems, time-to-complete must be normalized. We call this normalized measure Ideal Interactive Efficiency (IIE); It is calculated by comparing the ideal time-to-complete (defined below) with the actual time-to-complete, for the task, i.e.,

$$IIE(SearchTask) = \frac{T_c ideal}{T_c actual} \tag{1}$$

where $T_c ideal$ is the ideal completion time for a given search task. It is established during test collection preparation when users view a gold-standard Answer Card, and then go on to complete the search task successfully, so there is no interaction time. $T_c actual$ is the actual time it takes the user to complete the search task. Task timings commence from the moment the user is given the task and end when the user has finished their task completion response.

**BI Selection** Assess whether all BI elements are present

$$= \frac{\text{Actual number of distinct BI elements selected}}{\text{Required number of BI elements}} \tag{2}$$

This is equivalent to Q&A Instance Recall (Dang et al. 2007, pg 7)

**BI Representation** Assess the time and cognitive load required to process the BI elements before the person is able to complete the task. One method strongly akin to evaluating the quality of BI representation is the measurement of understandability of text. Zuccon (2016) proposed an understandability biased measure for document retrieval that could be adapted to BI retrieval. In an IIR scenario, timing how long the user takes to process the BI before completing the task may also provide an indicative measure. A variant of Time Biased Gain Smucker and Clarke (2012) could be utilised for batch system testing.

**BI Sequence** Assess whether sequential dependencies between the BI elements have been taken into consideration. The TREC temporal summarisation track took into consideration dependencies between information by excluding potentially relevant information that depended on other information which was not present (Aslam et al. 2014, 14). From a BI perspective, this means only counting distinct BI elements in equation 3 where all dependent BI elements are also present.

**BI duplication & presence of non-BI** Assess whether duplicate BI elements and/or non-BI elements are included on the Answer card, which would reduce the efficiency of processing the card by the user.

$$= \frac{\text{Actual number of distinct BI elements selected}}{\text{All Answer Card elements selected}} \quad (3)$$

This is equivalent to Q&A Instance Precision (Dang et al. 2007, pg 7). This has also been calculated as a verbosity measure and included in a discounting function (Aslam et al. 2014, pg 11).

# 5 Conclusion

The IR research community has traditionally focused on the miIR user search scenario in which multiple documents are retrieved for a specific query and the user is responsible for selecting and reading documents and possibly reformulating the query to fulfil their need.

The siIR search scenario, however, is rapidly growing in importance with the proliferation of personal digital assistants and hand-held devices, which favour single-shot query-answer systems.

Although many conceptual models have been developed to support IR in general, few, if any, are specific to siIR and many do not distinguish between siIR and miIR. A central concept in many of these IR models, Relevance, may not be suited to siIR and as suggested in Table 1, system design is likely to differ markedly between siIR and miIR systems. For these important reasons a new underlying conceptual model for siIR was sought. Despite much existing experimental research into siIR, a valid conceptual model would help ground and interpret this research and promote more systematic development, evaluation and measurement of siIR systems. This paper takes the first steps towards achieving this.

The proposed model, called Bridging Information Retrieval (BIR), was derived using a first principles approach from a task-completion perspective. The central premise of the model is that the purpose of IR is to provide Bridging Information to people to bridge the knowledge gap between the person's innate knowledge and that required for task completion. To delineate between miIR and siIR, a novel search task classification process was proposed based on the requirement and/or desire for multi-interaction search. This process enabled the development of a single-interaction BIR experimental framework to support siIR system evaluation.

**What value is there in a single interaction IR Framework?** The proposed framework helps researchers to:

1. Clarify which search tasks are suited to siIR as well as to encourage research into more complex tasks, such as case-based problem solving

2. Employ bridging information and task completion to evaluate Answer Cards, rather than using relevance, which is suited to multi-document retrieval

3. Explain the importance of user-group and search task selection and how together they impact system evaluation and the need for graded assessment

4. Identify the underlying siIR system objectives and performance measures, from which current measures in use are derived

In supporting the siIR researcher, it is expected that new and improved systems can be developed targeting more complex problems.

**Future Work.** BIR and the siBIR operational framework encourage many new lines of research. Most importantly the exploration of the relationship between Bridging Information and IR system performance including how the number, presentation and sequence of BI elements impacts system efficiency and effectiveness. Ascertaining how evidence fits within the model will be important - is evidence BI or a separate concept and how does evidence relate to the user acceptance of system solutions for different types of informing search tasks? Developing a more robust search task complexity continuum including the factors affecting the level of complexity would support a more systematic approach to task selection for TREC style experimentation. Also exploring the overlap between miIR and siIR tasks and assessing whether siIR systems provide better user satisfaction over interactive ones might expose user groups that have natural orientations to either system. Finally, case-based problem solving was touched on to exemplify the future direction of siIR systems - much research is needed here to develop viable solutions to these important problems.

# References


Agichtein, E., Carmel, D., Pelleg, D., Pinter, Y. & Harman, D. (2015), Overview of the TREC 2015 LiveQA Track., *in* 'TREC 2015', pp. 1–9.

Allan, J. (2003), HARD Track Overview in TREC 3: High Accuracy Retrieval from Documents, *in* 'Proc. 14th Text Retr. Conf. - TREC '05', pp. 1–17.

Allan, J., Croft, B., Moffat, A., Sanderson, M., Aslam, J., Azzopardi, L., Belkin, N., Borlund, P., Bruza, P., Callan, J., Carman, M., Clarke, C. L. a., Craswell, N., Croft, W. B., Culpepper, J. S., Diaz, F., Dumais, S., Ferro, N., Geva, S., Gonzalo, J., Hawking, D., Jarvelin, K., Jones, G., Jones, R., Kamps, J., Kando, N., Kanoulas, E., Karlgren, J., Kelly, D., Lease, M., Lin, J., Mizzaro, S., Murdock, V., Oard, D. W., Rijke, M. D., Sakai, T., Scholer, F., Si, L., Thom, J. a., Thomas, P., Trotman, A., Turpin, A., Vries, A. P. D., Webber, W., Zhang, X. J. & Zhang, Y. (2012), 'Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne', *SIGIR Forum* **46**(1), 2–32.

Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V. & Sakai, T. (2014), TREC 2014 Temporal Summarization Track Overview, *in* 'Proc. 23rd Text Retr. Conf.', pp. 1–15.

Bates, M. J. (1989), 'The design of browsing and berrypicking techniques for the online search interface', *Online Rev.* **13**(5), 407–424.

Bates, M. J. (1990), 'Where should the person stop and the information search interface start?', *Inf. Process. Manag.* **26**(5), 575–591.

Belkin, N. J. (1980), 'Anomalous states of knowledge as a basis for information retrieval', *Can. J. Inf. Sci.* **5**, 133–143.

Belkin, N. J., Cool, C., Stein, A. & Thiel, U. (1995), 'Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems', *Expert Syst. Appl.* **9**(3), 379–395.

Belkin, N. J., Marchetti, P. G. & Cool, C. (1993), 'BRAQUE: Design of an interface to support user interaction in information retrieval', *Inf. Process. Manag.* **29**(3), 325–344.

Belkin, N. J., Oddy, R. N. & Brooks, H. M. (1982), 'ASK for information retrieval: Part I. Background and theory', *J. Doc.* **38**(2), 61–71.

Brookes, B. C. (1980), 'Measurement in Information Science : Objective and Subjective Metricai Space', *J. Assoc. Inf. Sci. Technol.* **31**(4), 248–255.

Bulletin, P. & Voss, J. L. (2009), 'Long-term associative memory capacity in man.', *Psychon. Bull. Rev.* **16**(6), 1076–81.

Byström, K. & Järvelin, K. (1995), 'Task complexity affects information seeking and use', *Inf. Process. Manag.* **31**(2), 191–213.

Campbell, D. J. (1988), 'Task Complexity: A Review and Analysis.', *Acad. Manag. Rev.* **13**(1), 40–52.

Carterette, B. & Hall, M. (2013), Overview of the TREC 2013 Session Track, *in* 'TREC 2013', pp. 1–12.

Cleverdon, C., Mills, J. & Keen, M. (1966), 'Factors Determining the Performance of Indexing Systems Volume 1. Design', *ASLIB Cranf. Proj. Cranf.* **Vol 2**, 37–59.

Cleverdon, C. W. (1960), 'ASLIB Cranfield Research Project - Report on the first stage of an investigation into the ccomparative efficiency of indexing systems', *Aslib J. Inf. Manag.* **12**(12), pp421–431.

Cohen, N. J. & Squire, L. R. (1980), 'Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that.', *Science (80-. ).* **210**(4466), 207–210.

Cook, D. A., Sorensen, K. J., Hersh, W., Berger, R. A. & Wilkinson, J. M. (2013), 'Features of Effective Medical Knowledge Resources to Support Point of Care Learning: A Focus Group Study'.

Craswell, N., Hawking, D., Wilkinson, R. & Wu, M. (2003), Overview of the TREC 2003 Web Track., *in* 'TREC 2003', Vol. 3, pp. 1–15.

Dang, H., Kelly, D. & Lin, J. (2007), Overview of the TREC 2007 Question Answering Track., *in* 'TREC 2007', Vol. 7, p. 63.

Dean-hall, A., Clarke, C. L. a., Thomas, P. & Voorhees, E. (2014), 'Overview of the TREC 2014 Contextual Suggestion Track Adriel', *Proc. 21st Text Retr. Conf.* .

Dervin, B. (1983), *An overview of sense-making research: Concepts, methods, and results to date*, The Author.

Engelbart, D. C. & English, W. K. (1968), A research center for augmenting human intellect, *in* 'Proc. December 9-11, 1968, fall Jt. Comput. Conf. part I', ACM, pp. 395–410.

Goodwin, T. R. & Harabagiu, S. M. (2016), Medical Question Answering for Clinical Decision Support, *in* 'Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.', ACM, pp. 297—-306.

Graham, D. W. (1999), 'Aristotle: Physics, Book Viii'.

Harman, D. (1995), Overview of the fourth text retrieval conference (TREC-4), *in* 'TREC-4', Vol. 4, pp. 1–24.

Harter, S. (1992), 'Psychological relevance and information science', *J. Am. Soc. Inf. Sci.* **43**(9), 602—-614.

Hersh, W. & Bhupatiraju, R. (2003), TREC Genomics Track Overview., *in* 'TREC 2003', pp. 14–23.

Hornbaek, K. (2006), 'Current practice in measuring usability: Challenges to usability studies and research', *Int. J. Hum. Comput. Stud.* **64**(2), 79–102.

Huang, X. & Soergel, D. (2013), 'Relevance: An improved framework for explicating the notion', *J. Am. Soc. Inf. Sci. Technol.* **64**(1), 18–35.

Ingwersen, P. (1984), 'Psychological aspects of information retrieval', *Soc. Sci. Inf. Stud.* **4**(2-3), 83–95.

Ingwersen, P. & Järvelin, K. (2005), *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, Springer Science & Business Media.

Järvelin, K. (1986), On information, information technology and the development of society: An information science perspective, *in* 'Inf. Technol. Inf. use Towar. a unified view Inf. Inf. Technol.', Taylor Graham Publishing, pp. 35–55.

Kelly, D. (2009), 'Methods for evaluating interactive information retrieval systems with users', *Found. Trends Inf. Retr.* **3**(1-2), 1–224.

Kelly, D. & Sugimoto, C. R. (2013), 'A systematic review of interactive information retrieval evaluation studies, 1967–2006', *J. Am. Soc. Inf. Sci. Technol.* **64**(4), 745–770.

Kent, A., Berry, M. M., Luehrs, F. U. & Perry, J. W. (1955), 'Machine literature searching VIII. Operational criteria for designing information retrieval systems', *Am. Doc.* **6**(2), 93–101.

Kuhlthau, C. C. (2004), *Seeking meaning: A process approach to library and information services*, Libraries Unltd Incorporated.

Li, Y. & Belkin, N. J. (2008), 'A faceted approach to conceptualizing tasks in information seeking', *Inf. Process. Manag.* **44**(6), 1822–1837.

Luhn, H. P. (1957), 'A statistical approach to mechanized encoding and searching of literary information', *IBM J. Res. Dev.* **1**(4), 309–317.

Miller, G. a. (1956), 'The magical number seven, plus or minus two: some limits on our capacity for processing information.', *Psychol. Rev.* **101**(2), 343–352.

Mizzaro, S. (1997), 'Relevance: The whole history', *J. Am. Soc. Inf. Sci.* **48**(9), 810–832.

Navrátil, P., Quaglioni, S., Hupin, G., Romero-Redondo, C. & Calci, A. (2016), 'Unified ab initio approaches to nuclear structure and reactions', *Phys. Scr.* **91**(5), 53002.

Over, P. (2001), 'The TREC interactive track: an annotated bibliography', *Inf. Process. Manag.* **37**(3), 369–381.

Potter, M. (2004), *Set theory and its philosophy: A critical introduction*, Clarendon Press.

Reid, J. (1999), A New Task Oriented Paradigm for Information Retrieval: Implications for Evaluation of Information Retrieval Systems., *in* 'CoLIS', Vol. 3, pp. 97—-108.

Reid, J. (2000), 'A Task-Oriented Non-Interactive Evaluation Methodology for Information Retrieval Systems', *Inf. Retr. Boston.* **2**(1), 115–129.

Roberts, K., Demner-Fushman, D., Voorhees, E. M. & Hersh, W. (2016), Overview of the TREC 2016 Clinical Decision Support Track, *in* 'Proc. Twenty-Fifth Text Retr. Conf. TREC 2016', pp. 1–14.

Robertson, S. & Soboroff, I. (2002), The TREC 2002 Filtering Track Report., *in* 'TREC 2002', pp. 1–5.

Sadeghi-Bazargani, H., Tabrizi, J. S. & Azami-Aghdash, S. (2014), 'Barriers to evidence-based medicine: a systematic review', *J. Eval. Clin. Pract.* **20**(6), 793–802.

Saracevic, T. (1975), 'Relevance: A review of and a framework for the thinking on the notion in information science', *J. Am. Soc. Inf. Sci.* **26**(6), 321–343.

Saracevic, T. (1996), Relevance reconsidered, *in* 'Proc. 2nd Conf. Conceptions Libr. Inf. Sci.', pp. 201–218.

Saracevic, T. (2016), 'The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?', *Synth. Lect. Inf. Concepts, Retrieval, Serv.* **8**(3), i—-109.

Schamber, L., Eisenberg, M. & Nilan, M. (1990), 'A re-examination of relevance: toward a dynamic, situational definition', *Inf. Process. Manag. An Int. J.* **26**(6), 755–776.

Simpson, M. S., Voorhees, E. M. & Hersh, W. (2014), Overview of the trec 2014 clinical decision support track, Technical report.

Smucker, M. & Clarke, C. (2012), Time-based calibration of effectiveness measures, *in* 'Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.', pp. 95—-104.

Soboroff, I., de Vries, A. & Craswell, N. (2006), Overview of the TREC 2006 Enterprise Track., *in* 'Trec 2006'.

Soboroff, I. & Harman, D. (2003), Overview of the TREC 2003 Novelty Track., *in* 'TREC 2003', pp. 38–53.

Sperber, D. & Wilson, D. (1986), *Relevance: Communication and cognition*, Vol. 142, Harvard University Press Cambridge, MA.

Tsikrika, T., Larsen, B., Müller, H., Endrullis, S. & Rahm, E. (2013), The scholarly impact of CLEF (2000–2009), *in* 'Int. Conf. Cross-Language Eval. Forum Eur. Lang.', Springer, pp. 1–12.

Tulving, E. & Donaldson, W. (1972), *Episodic and semantic memory*, Academic Press.

Vakkari, P. (2003), 'Task-Based Information Searching', *Annu. Rev. Inf. Sci. Technol.* **37**, 413–464.

Voorhees, E. (2003), Overview of the TREC 2003 Robust Retrieval Track., *in* 'TREC 2003'.

Voorhees, E. & Harman, D. (2005), *TREC: Experiment and evaluation in information retrieval*, MIT Press, Cambridge, MA.

Yang, H., Frank, J. & Soboroff, I. (2015), TREC 2015 Dynamic Domain Track Overview, *in* 'TREC 2015', pp. 1–28.

Zuccon, G. (2016), Understandability biased evaluation for information retrieval, *in* 'Eur. Conf. Inf. Retr.', Springer, pp. 280–292.