2
Information Exploration

Humans are explorers by nature, the term "Exploration" refers to "the act of travelling through a place in order to find out about it or look for something in it" (HOMBY; WEHMEIER, 2004). In a similar way, in the information context, exploration can be seen as a journey of acquiring new knowledge through the exposure to, interpretation, and analysis of an information space (BELKIN, N.J.; ODDY; BROOKS, 1982; KUHLTHAU, 1991; VAKKARI, 2010), preferably supported by an efficient computational system. The idea of enhancing human ability for manipulating information spaces in order to promote intellectual growth is not new. In 1945, Vannevar Bush had already envisioned a device called "memex" (memory extender) (BUSH *et al.*, 1945) for storing and retrieving books, creating and following links, and annotating contents in a private library. This device would work as a prosthetic enhancement for the individual's memory. Bush's "memex" paved the way for the elaboration of the Hypertext concept (NELSON, 1965), and the World Wide Web (BERNERS-LEE *et al.*, 1994). However, with the spreading of the Web and the subsequent development of semantic technologies, cloud computing, and Big Data, the information landscape has increased exponentially both in size and complexity, presenting a big challenge for information explorers.

People usually explore information spaces through interaction with an Information Retrieval System (IRS), where a query is specified and submitted to the system and the system returns a set of documents that match the original query. Figure 1 shows a visual representation of this model. This is the basic interaction model implemented in systems such as Google[3], Yahoo![4], and Microsoft Bing[5], and database management systems. If on one hand such systems

[3] google.com
[4] https://www.yahoo.com/
[5] https://www.bing.com/

are easy to use, on the other hand it models any information-seeking task as a sequence of isolated query-responses.

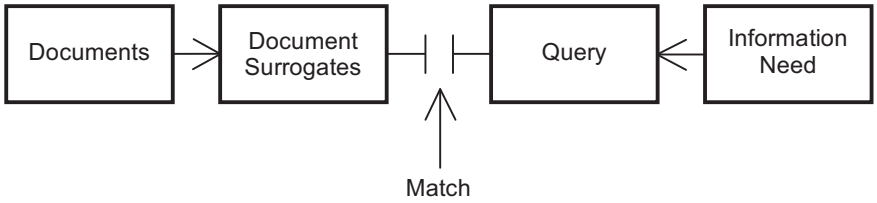

Figure 1 - Information Retrieval model (BATES, 1989)

The query-response model has been strongly criticized in the last years due to its poor semantics for describing complex information tasks (BATES, 1989). One can have a reasonable success rate for fact-finding and question-answering tasks, where the user accurately specifies the query that precisely matches the desired items with minimal examination of the results. This scenario is unrealistic for the majority of real-life search tasks (ROSE; LEVINSON, 2004), where, besides data lookups, the user navigates, filters, gathers, examines and compares result set items. This way, lookup actions can be seen as an attempt to approximate the desired information items that will be further processed over multiple interaction sessions over time (MARCHIONINI, 2006). Examples of such complex tasks are to find trends in patenting behavior, write an essay of alternative treatments for a disease, or tracing the profile of a research institution.

The focus of the IR field lies on researching data representations and matching algorithms, where precision and recall are the main evaluation measures (BAEZA-YATES; RIBEIRO-NETO, 1999). It usually does not include the human factors, such as, expected outcomes and task characteristics and context in their evaluations. Interactive Information Retrieval (IIR) (RUTHVEN, 2009) studies how people interact with information using IR systems, where the focus goes beyond the quality of the matching algorithms and also includes researching of novel kinds of interactions with IR systems and human-centered evaluations.

Information Visualization aims at presenting visual representations of large collections of data to aid its analysis and interpretation. Although it is a relevant technique to leverage exploration processes, its main focus is not on information-seeking processes (WHITE; ROTH, 2009).

Information Seeking (IS) is a general definition of information tasks that aims at describing any kind of task employed to fill knowledge gaps on the mind

of the seeker (BELKIN, N, 1995). Exploratory Search and Exploratory Data Analysis (TUKEY, 1977) are specializations of Information Seeking. The distinction between exploratory search tasks from other information seeking tasks is the characteristics of both the search context and the search process, as explained by (WHITE; ROTH, 2009). The search context describes the users goal and expectations, the knowledge state and preferences, and also the emotions involved, where uncertainty and anxiety are common feelings (KUHLTHAU, 1991). The exploratory search process can be ultimately described as a combination of querying and browsing activities.

In this work we extrapolate the common notion of exploratory search processes. Besides querying and browsing activities motivated by a desire of learning, we also approach activities targeting the management of the knowledge acquired along the process, as well as reuse and sharing of exploration solutions, preferably leveraged by a formal exploration model. For this reason, we refer to the process of exploration of (semi) structured datasets as *Information Exploration*, which is considered as a generalization of exploratory search tasks (WHITE; ROTH, 2009). The next sections describe in detail the characteristics of exploration tasks and behavioral models that give us the background to approach exploration processes in the remaining chapters.

2.1. Exploration Tasks

Exploration tasks are usually composed of querying and browsing activities to foster intellectual development. Nonetheless, there are characteristics that particularly elicit exploratory behavior in the execution of information tasks. The work in (WILDEMUTH; FREUND, 2012) carries out a literature survey addressing such characteristics and presents a summarized list:

Cognitive:

- They have learning and investigation as acceptable goals;
- The problem is general rather than specific;
- They involve some degree of uncertainty;
- The problem is ill-structured, i.e., the definition does not contains detailed information of sub problems or the aspects to be tackled;
- The task definition is dynamic and evolves over time;

- The problem is multi-faceted, requiring the investigation of multiple concepts, or multiple dimensions of a single concept;
- The problem has a higher degree of complexity or difficulty;
- The task requires cognitive processes such as analysis and sense-making, decision-making or other.

Behavioral:

- The task is open-ended, tending not to finish with a clear and punctual answer;
- The task target may be multiple items instead of a single item;
- The tasks occur over time, usually through multiple iterations and search sessions.

While the first group presents the cognitive challenges that are usually faced in exploration tasks, the second group focuses on how users behave during the task resolution process. The first cognitive characteristic identifies the learning goal of the tasks. Lack of knowledge has been appointed as key motivator of exploration tasks. BELKIN *et al*. (1982) has characterized the situation where the user has a problem and does not have adequate knowledge for deriving a precise description of the documents in the form of a IR query as an "anomalous state of knowledge". KUHLTHAU (1991) defines the information seeking process as a sequence of stages mostly characterized by the reduction of uncertainty, proportional to the gain of contextual knowledge, at each stage. MARCHIONINI (2006) also presents *learning* and *investigate* as acceptable goals for exploration tasks. Here, we define knowledge as an information type generated as a result of cognitive information processing, such as analyzes, interpretations, and comparison processes (KRATHWOHL, 2002). Thereby, learning can be considered a reasoning function that maps information sources to new knowledge states.

The work in (BYSTRÖM; JÄRVELIN, 1995) presents a more detailed definition of the types of information sources involved in the exploration processes and also investigates the relationship between the use of information sources and the task complexity, which tends to have high degree for exploration tasks. The work presents three types of information sources: problem information, domain information, problem solving information. They address respectively

information concerning the problem context, structure, and expected outcomes; the formally recorded information that will be used in the solution process; the solution methodology or problem solving strategy employed. The work demonstrates through user studies that the higher the degree of complexity and lack of structure of the task, the greater is the necessity of problem and problem solving information, which usually is not available in the environments. This work presents a proposal to capture, represent, and reuse problem solving information.

With regards to the domain information of exploration tasks, it tends to be composed of multiple and orthogonal concepts or dimensions since the tasks are often multi-faceted. For example, consider a student with the task of writing a survey covering some research field. The student has to figure out what are the most authoritative publications, the relevant authors, the main publication venues, and the intersections of the research field with other fields. Therefore, there are multiple facets that should be addressed in order to accomplish the task.

Other characteristics of exploration tasks that are interrelated are the generality and the ill-structuredness, which requires exploratory actions to build an understanding of the problem itself before solving it. For example, planning a safe trekking to high-altitude places (KINLEY *et al.*, 2012) would require an investigation of illness related to high altitudes and preventions before the planning. This way, the more abstract or vague is the problem definition, the higher is the need for exploratory behavior.

Besides the cognitive characteristics, there are also behavioral aspects that are commonly present in exploration tasks. First, it is hard to figure out when some tasks finish. For example, the task of tracing the patenting behavior of a set of competing companies may not have a well-determined end since the outcome is relevant competitive intelligence information for the life of a company. This type of task is called open-ended. Moreover, this task depends on multiple items (patents) and can also be iterative on the set of competing companies to be analyzed.

2.2. Exploration Process

Since Exploration is a specialization of general Information-Seeking tasks, models in information seeking are also valuable to explain exploratory behaviors during the task resolution process. The majority of the models are devised through

qualitative research, which is considered more adequate for explaining the levels of abstractions, the dimensions, the strategies and tactics, and the user-system interactions in exploration tasks (WILSON, T.D., 1999). IS models can be understood in terms of three facets: Abstractions, Behaviors, and Strategies. They respectively describe the abstract concepts involved, such as, Users, Tasks, Information Sources, etc.; the procedural characteristics, such as, iterations, directness, or serendipity, as well as states and state transitions; and goals, atomic actions, activities, and strategies.

Wilson's model (WILSON, T.D., 1999) describes the IS task as composed by the originating context of the information need, which can be from the *environment*, such as, the work place or the school, the *social role*, such as, the position at work, parent, voter, etc., and the *personal context*, such as, psychological, affective, and cognitive contexts. Moreover, (WILSON, T.D., 1999) recognizes the existence of barriers that should be overcome before the information seeking behavior starts. For example, the cost and availability of the information sources can prevent information seeking behavior. The collaborative aspect is also described in Wilson's model, captured in the "Information Exchange" stage, where the user can "transfer" information found to other users. From the behavioral aspect, the task is considered an iterative process that can culminate in success or failure, as shown in Figure 2. From the strategic dimension, the main component is the "Information Seeking Behavior", which is defined as a set of IS activities and is left as a "hotspot" which can be instantiated, for example, with Ellis' activities model (ELLIS; COX; HALL, 1993) (WILSON, T.D., 1999).

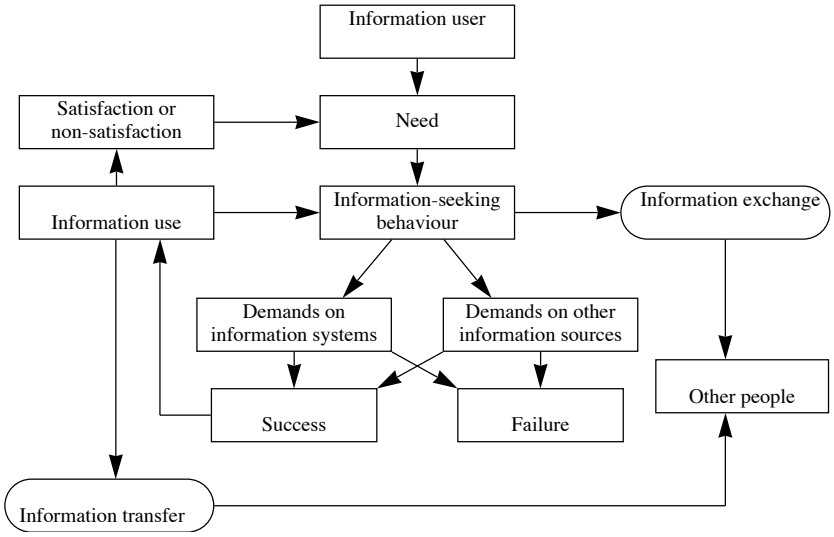

Figure 2 - Wilson's Model of Information Seeking (WILSON, T.D., 1999)

As demonstrated in (WILSON, T.D., 1999), Ellis' model (ELLIS; COX; HALL, 1993) can instantiate the "Information-Seeking Behavior" component in Figure 2, which aims at describing single IS activities in the library context. Briefly, the activities are:

- Starting: comprises any action the user can take to start an IS task, such as, identifying a key paper to start the task (ELLIS; COX; HALL, 1993);
- Chaining: following footnotes and citations in the current material. Chaining actions are classified in two types: forward and backward;
- Browsing: considered "semi-structured searching" is the act of scanning a wide range of journals to select those more relevant;
- Differentiating: filtering the list of materials by comparisons of known characteristics;
- Monitoring: keeping up-to-date on a particular topic. It can involve successive checks for updates on the sources of information;
- Extracting: select relevant materials in the information sources;
- Verifying: assess the accuracy and relevance of the extracted information;
- Ending: task closing actions.

One worthwhile fact to note is that (ELLIS; COX; HALL, 1993) identified traces of collaborative information seeking in the starting stage, where users seek out help from specialized people in order to identify relevant papers and sources.

Although (ELLIS; COX; HALL, 1993) describes key activities, there is no information about such process, e.g., whether it is iterative or the activities may be concurrent. Moreover, subjective aspects, such as, degree of uncertainty is not precisely accounted.

The “berrypicking” model, presented by (BATES, 1989) and illustrated in Figure 3, advances the traditional IR query-response paradigm by modeling the interaction as a complex search process composed of multiple and connected query-response interactions.

The “berrypicking” is based on the metaphor of picking berries in a forest, where the explorer picks berries that are distributed on many bushes. In berrypicking, the information needed is fragmented among many documents. The explorer navigates along the information space gathering pieces of information until she/he feels satisfied. Each piece of information gathered gives clues about where to go next. The places in the information space are subsets of documents retrieved by a query. Each information fragment gathered elicits a query reformulation action, which takes the explorer to a different place in the information space. Therefore, exploration tasks “are as much about the journey through the information space as the destination” (WHITE; ROTH, 2009).

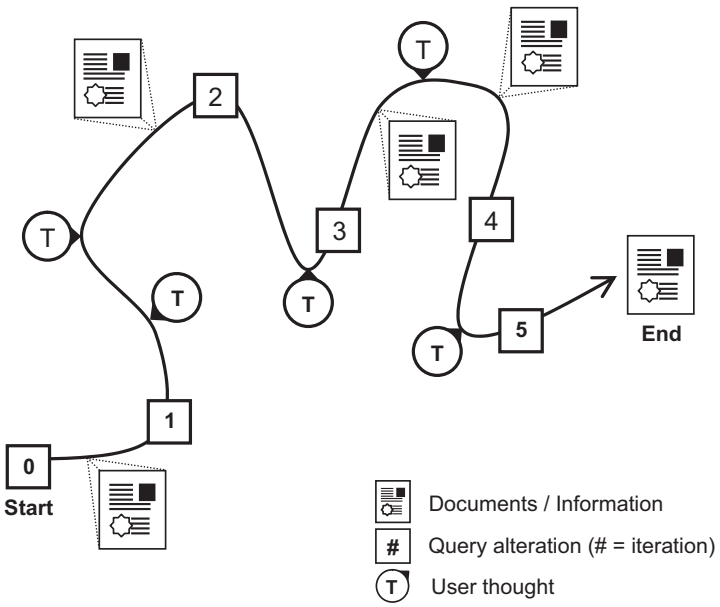

Figure 3 - Bate's Berrypicking Model (BATES, 1989)

It is worth noting that the “query alteration“ activity in the berrypicking model is not restricted to information retrieval matches but it stands for any kind of search tactic employed. (BATES, 1979, 1989) present several activities

organized in a four-level hierarchy: moves, tactics, stratagems, and strategies. Moves are atomic actions, which can be either physical or cognitive. Tactics can be composed of many moves. For example, the tactics CHECK and SPECIFY (BATES, 1979) can be used respectively to assess the relevance of the current status of the search with the original request and to restrict the search terms to the ones that are as specific as the information desired. Many moves can be used to accomplish these two tactics, such as filtering, querying a thesaurus, and browsing. Stratagems are compositions of tactics, such as using a citation index to achieve and identify relevant works that have cited a particular paper. Strategies, in turn, are compositions of stratagems, tactics, and moves, where the activities are directly connected to the general problem, such as "Finding central papers in a given topic." A complete list of tactics organized by category can be found in (BATES, 1979, 1980).

Khulthau's model (KUHLTHAU, 1991) advances the comprehension of IS behavior by dividing the IS process into a sequence of stages and describing cognitive and affective aspects of users while they move forward in their tasks. Khulthau's divides the IS task execution in six stages. The task starts by recognizing the information need in the *Initiation* stage. Next, the user moves to the *Selection*, *Exploration*, and *Formulation* stages, where they select a general topic, engage in exploratory actions to discover new information on the selected topic, and break up the topic into specialized sub-topics in order to determine a focus. Thereafter, the user enters in the *Collection* stage, where s/he collects relevant information on the formulated focus and organizes it. The task is concluded in the *Presentation* step. (KUHLTHAU, 1991) found that the user experiences high levels of uncertainty, doubt, and anxiety in the first steps. These feelings tend to reduce as the user moves through the stages and gains better understanding of the information space and the task. Khulthau's main concern is in the emotional aspect, hence, no claim is made with regards to the range of actions within each stage. The stages, though, provide relevant information on the possible contexts in which single actions can be carried out.

The more recent work by Marchionini's (MARCHIONINI, 2006) was the first to introduce the term "Exploratory Search", which also contrasts with traditional keyword searches, in terms of the goals, the process, and required

system support. Marchionini describes the process in terms of actions within three major goals: Lookup, Learn, and Investigate, as Figure 4 shows.

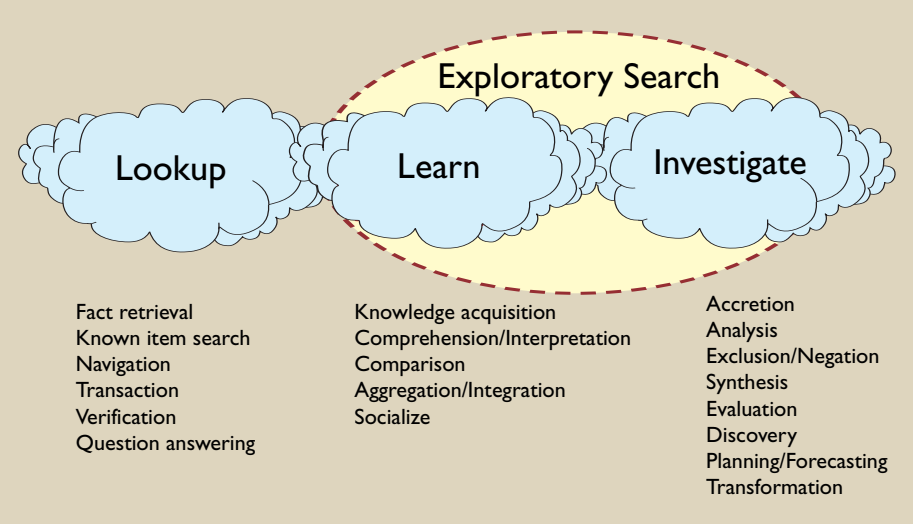

Figure 4 - Marchionini's Exploratory Search Model (MARCHIONINI, 2006)

Lookup tasks have a precise and well-defined query, and return a discrete set of matching items. The main goal of traditional search engines and database systems is to provide precise lookups, however, they fail to give support to further comparisons and examinations when the goal is to learn and investigate, which are the motivators for Exploratory Search. Learning is defined as the task of developing new knowledge of any kind. Based on Bloom's taxonomy of educational objects (KRATHWOHL, 2002), Marchionini defines the activities within the *Learn* goal, such as knowledge acquisition, aggregation, and interpretation. *Learn* and *Investigate* can be complementary since the latter is composed of a set of activities that aims at discovering new information or knowledge gaps in some area. It is carried out through iterations of critical analysis, evaluations, and synthesis. Marchionini does not define a precise boundary between lookup, learning, and investigate - the user, for example, can be executing many lookups in order to investigate the absence of materials in some area. Although Marchionini's work presents an expressive set of activities, it brings no information with regards to the user behavior or problem solving strategies that can be adopted.

(WHITE; ROTH, 2009) defines the exploration process as a range of actions that varies from exploratory browsing to focused search, where exploratory browsing is a sequence of movements within some connected space in

order to better define the problem and the solution space while focused search involves following a well-defined trail for reaching the solution, which can involve item lookups, such as known-item searches and a certain degree of navigation along previously known paths.

The Exploratory Browsing step in (WHITE; ROTH, 2009) can be motivated by curiosity, involves creative and lateral thinking, and serendipity. Although serendipity has been recognized as a valuable tool for "accidental" discoveries, it is considered at risk in digital information seeking due to the excessive attention to best-match and ranking paradigms, which aims at high precision and may reduce the browsing possibilities (FOSTER; FORD, 2003). In exploration, there must be a shift in the focus from precision to recall (MARCHIONINI, 2006; WHITE *et al.*, 2007; WILSON, MAX L.; SCHRAEFEL; WHITE, 2009). Focusing on recall would increase the number of exploration possibilities and, hence, discoveries.

The concept of creativity in the context of information seeking can be approached as divergent thinking, which consists in recognizing similarity relationships of otherwise semantically distant concepts (FORD, 1999). The more dissimilar are the concepts, the higher the creativity thinking degree. For example, applying evolutionary principles of one animal species to another is less creative than applying the same principles to computers due to the degree of dissimilarity of the concepts (FORD, 1999). Therefore, formulating a focus on two dissimilar concepts and exploring their correlations is an interesting feature for (CHOI, 2010) an IS system that has been leveraged by Semantic Web and NoSQL technologies. In this work, this concept leveraged the discussion of branching possibilities, where multiple and alternative sets of items can be explored simultaneously. This is also related to the design issue of choosing between unifocal and multifocal interfaces, described in chapter 7.

Another relevant concept in IS is "Teleportation", which consists in jumping directly to relevant information previously found (O'DAY; JEFFRIES, 1993). This behavior is common in the use of search engines. Instead of browsing along a web site in order to find the correct page, the search engine results page offers links that "teleports" us directly to a specific web page containing the keywords within a web site. Therefore, once relevant information is found, it is desirable to jump directly to it in future tasks (O'DAY; JEFFRIES, 1993).

From the point of view of the process, Exploratory Search is similar to Tukey's Exploratory Data Analysis (EDA) (WHITE; ROTH, 2009). However, when engaged in an EDA process, the user follows a more pragmatic learning approach and assumes the role of a detective, where she/he manipulates the data to formulate hypothesis and test them. In EDA, the user carries out a series of data transformations, summarizations, and representation changes on batches of data in order to derive hypothesis for explaining some phenomena (TUKEY, 1977).

2.3. Summary

The theoretical models presented in this chapter give us a rich background to understand the characteristics of exploration tasks and stages, the users' behaviors and activities, and solution strategies. These models can be used to answer questions of why an exploration starts, what activities may be involved, and which behaviors must be supported at each phase of the process. Now, we are in position to establish the features an exploration environment must offer. (WHITE; ROTH, 2009) list the following features resulting from discussions between experts and independent workshops on exploratory search systems:

1. "Support querying and rapid refinement"
2. "Offer facets and meta-data based result filtering"
3. "Leverage search context"
4. "Offer visualization to support insight and decision making"
5. "Support learning and understanding"
6. "Facilitate collaboration"
7. "Offer stories, workspaces, and progress updates"
8. "Support task management"

Although this list gives good directions of what exploration systems should provide, it is too abstract to describe the physical processing operations that must be offered for a given exploration scenario. For example, the majority of faceted search tools offer querying and refinement possibilities (1) and meta-data based filtering (2). However, some faceted tools are more adequate for some tasks than the others, as our evaluations in chapter 6 shows. Another example is the requirement 7: "Offer stories, workspaces, and progress updates". As we demonstrate in chapter 7, the majority of the tools offer some kind of history and

task progress management but the question is: how to know if they efficiently communicate task information? What are the types of task information that should be communicated? Therefore, there is a gap between the semantics of these models and their physical implementation in real exploration systems that must be bridged.

In order to leverage the design of efficient exploration systems, there are more specific questions that must be answered. For example, the list of features presented in (WHITE; ROTH, 2009), or taxonomies of activities, as the ones presented in (MARCHIONINI, 2006), says nothing about which types of query can be issued or which filtering criteria can be applied. The main contribution of this work is a framework of data processing operations that covers all these aspects and presents a rich semantics for both designing and evaluating real exploration systems.