A Process Model for Goal-Based Information Retrieval

Harvey Hyman
Advanced Technology Program, Florida Polytechnic University
Lakeland, Florida 33801, USA
hhyman@floridapolytechnic.org

ABSTRACT

In this paper we examine the domain of *information search* and propose a “goal-based” approach to study search strategy. We describe “goal-based information search” using a framework of *Knowledge Discovery*. We identify two Information Retrieval (IR) goals using the constructs of *Knowledge Acquisition* (KA) and *Knowledge Explanation* (KE). We classify these constructs into two specific information problems: An *exploration-exploitation* problem and an *implicit-explicit* problem. Our proposed framework is an extension of prior work in this domain, applying an IR Process Model originally developed for Legal-IR and adapted to Medical-IR. The approach in this paper is guided by the recent ACM-SIG Medical Information Retrieval (MedIR) Workshop definition: “methodologies and technologies that seek to improve access to medical information archives via a process of information retrieval.”

Keywords: Information retrieval, Medical information retrieval, Process model, Exploration, Exploitation, Implicit, Explicit, Knowledge discovery, Knowledge acquisition, Knowledge explanation.

1. INTRODUCTION

A recent review of literatures in this area [1], [2], [3], [4] suggests that there are three common research themes being pursued in the domain of Legal-IR and Medical-IR: (1) Visual display of information, (2) Incorporating external amplifying information and (3) Integrating internal explanatory information. The first theme of visual display centers on how to design a visual interactive system, to foster better user understanding of terminologies and vocabularies contained in an electronic document. The second theme of incorporating external information is concerned with how a system can be used to support a user’s information retrieval need by accessing external sources. The third theme of integrating explanatory information describes the problem of interpreting text based content and assimilating domain dictionaries and lookups. In this paper we explore a specific example of this in the domain of Medical-IR where the user has access to taxonomies and libraries such as SNOMED and UMLS.

The framework described herein proposes a goal-based approach to IR using *acquisition* and *explanation* as descriptive goals serving varying user information needs and varying user knowledge levels. We propose this framework of Knowledge Discovery to design and develop IR oriented information systems, and to evaluate how well a system supports the goals of amplification of user knowledge and increased user understanding using internal and external sources.

2. DEFINITIONS

Traditional Knowledge Discovery comes from the domain of data analytics. It is concerned with the discovery of new patterns emerging from implicit information [6], with few or no pre-defined goals. An interesting characteristic of IR is that it is concretely defined by two overriding main goals: improved understanding and increased knowledge in a user specified topic scope. The “discovery” is one of *understanding* and *amplification* of knowledge using pre-defined terminologies [1]. Using a Knowledge Discovery Framework as our guide, we seek to classify information search as two distinct goals of *knowledge explanation* and *knowledge acquisition*. In this paper, we apply this classification method to the domain of Medical-IR, to leverage the predefined taxonomies and libraries available.

We define Knowledge *Explanation* (KE) as an explicit-implicit problem. This is a problem of definable, but complex terms needing to be reduced to a layman’s level of understanding. We believe this definition best describes research problems centered about the goal of explaining terms that need to be better understood. The problem considers that explicit knowledge represents information that is common knowledge or readily accessible to the layman, easily codified in written form and often found in manuals, documents, and various web media outlets (links, pages, etc.); and that implicit knowledge, represents information that is not commonly known, but its meaning is often based upon specialized knowledge of a narrowly focused community of experts in the area. This type of knowledge is sometimes called tacit knowledge [8]. The problem of translating the tacit (implicit) to the explicit has been previously defined by Nonaka in this work on social transfer theory (SECI) [9].

We apply the Knowledge Explanation construct to the instance of Medical-IR where a specific research question focuses on how to design a visual interactive system to support patient understanding of terminologies in discharge summaries using internal sources of local vocabulary dictionaries and thesauri such as UMLS and SNOMED. We categorize the terminologies in documents such as discharge summaries as implicit, insofar as their usage is operationalized as common parlance of the experts and thereby outside the knowledge base of the patient (layman). The system objective here is to convert the implicit to the explicit, to achieve the stated goal of better patient understanding [10].

We define Knowledge *Acquisition* (KA) as an exploration-exploitation problem insofar as it is an information search with the purpose of acquiring additional external documents to amplify a user's knowledge on a subject matter, topic or terminology. Knowledge Acquisition is used as a supporting construct for addressing how to provision a user's information retrieval need by accessing external sources. We define this acquisition need as a situation whereby a user desires to amplify information about a specific topic.

The exploration-exploitation problem describes the dilemma of a user's decision to focus attention and commit resources to the current selection versus abandoning it in favor of searching for a new selection. This method is operationalized by the IR process model (Figure 2) adapted to Legal-IR and Medical-IR. The process model provides a mechanism to design an interface to support the behaviors of exploration and exploitation.

Exploration as a construct explains human search behavior [11]. Operational examples include electronic search and methods that support browsing. The exploration construct itself, can be further classified into extrinsic and intrinsic. Extrinsic exploration typically has a specific task purpose, whereas intrinsic exploration is motivated by learning [12].

3. APPROACH

Our approach begins with defining a framework for *knowledge discovery* (KD). In this framework we classify two distinct objectives to support the goal of improved patient knowledge and understanding. We describe these two objectives using the constructs of *Knowledge Acquisition* and *Knowledge Explanation*. Next, we apply a previously validated IR process model from the domain of eDiscovery [10], [11], adapted to Medical-IR as a mechanism to implement the KD framework and support the exploration-exploitation balance. We use the process model as a benchmarking tool for evaluating prototype systems built to support IR goals and objectives.

The original process model has been designed to support context learning and knowledge discovery. The adapted model is based on an *explicit-implicit* scheme (defined previously) to support knowledge explanation and an *exploration-exploitation* scheme to support knowledge acquisition. When we apply this approach to our Medical-IR instance, the objective of *explanation* is supported by direct injection of query expansion, using terms from internal documentation (discharge summaries) indexed against predefined taxonomies (SNOMED and UMLS). The objective of acquisition is supported using external information search with relevancy rankings based on context and content similarities.

4. FRAMEWORK

Figure 1 depicts the Knowledge Discovery Framework. Within the framework we classify knowledge discovery into the constructs of *Acquisition* and *Explanation*. The acquisition construct represents externally acquired information for the purpose of supporting patient amplification of knowledge on a topic, condition, or terminology. We apply the explanation construct in this case to the Medical-IR problem of supporting better patient understanding of a diagnosis, condition, or terminology contained within a discharge summary and indexed against a predefined taxonomy such as SNOMED or UMLS.

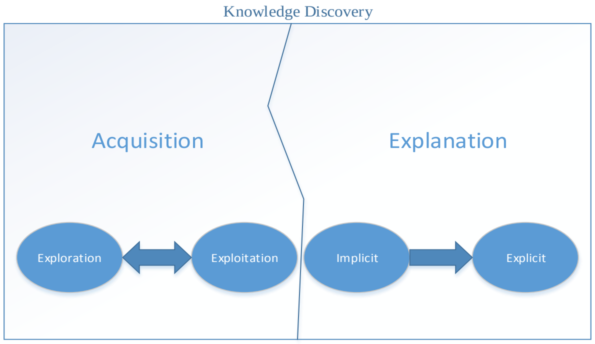

Figure 1: Knowledge Discovery Framework for IR

5. APPLYING A PROCESS MODEL

We assume the two goals of acquisition and explanation. Figure 2 depicts the iterative and cyclic approach to the exploration-exploitation behavior in the search process. The IR process begins with an initial search structure. This structure can be based on a specific mental model of the user, or in the case of Medical-IR, we begin with specified terminologies in the discharge summary. The system (whatever system is used) then will begin a retrieval action to suggest documents or links containing possibly relevant content. The user is presented with a "snapshot" of documents to facilitate the exploration process. The model describes three levels of exploration and simultaneous exploitation: scanning, skimming and scrutinizing. Scanning describes the superficial review of a list of titles, skimming describes a superficial review of a document selected from the list of titles, and scrutinizing describes a deeper, more critical review of the selected document.

In the goal of knowledge *acquisition*, we are focused on searching for documents that will amplify or expand the user's knowledge on a specified topic. In the example of Medical-IR, we can think of a term in a discharge summary or a condition described in a medical diagnosis. In this case we are dealing with bulk information in the form of web pages, links and other documents, similar to a conventional information search. The goal here is to *acquire* additional information on a topic or term to broaden or deepen knowledge. This is a relevancy problem (comparing potential documents, determining the most relevant and displaying a ranking method).

In the goal of knowledge *explanation*, we can apply the IR process model to support information search via an iterative cycle of balanced exploration and exploitation activities to convert the implicit (complex descriptions) to the explicit (simplified or amplified descriptions). In traditional IR problems, we are dealing with a bulk volume of documents to classify and extract. In the applied instance of Medical-IR, we are dealing with a single patient and the two information goals based on a specific document (discharge summary, medical diagnosis, or other example). In either instance, we apply the IR process model as an integration tool – In the Medical-IR application we use a text mining method to recognize known medical terminologies (appearing in the domain dictionary such as SNOMED or UMLS). These terminologies are assumed to be highly relevant, meaning that we have a strong assumption that the patient seeks amplification of knowledge in this targeted area.

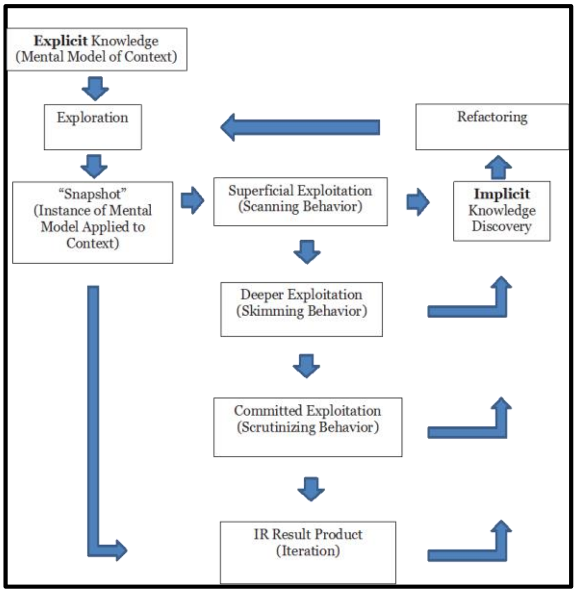

Figure 2: IR Process Model (Hyman et al.)[1]

[1] Cite as: Hyman, H. S., Sincich, T., Will, R., Agrawal, M., Padmanabhan, B., and Fridy, W., "A Process Model for Information Retrieval Context Learning and Knowledge Discovery," (Under Review).

6. CONCLUSION

The purpose of this paper has been to propose a framework for thinking about the convergent and divergent information goals in IR. We adapt the IR Process model from Legal-IR and apply this approach to an instance of Medical-IR to leverage the already existing taxonomies. We offer the methods described in this paper as a conversation starter and welcome comments and feedback from readers.

7. REFERENCES

[1] **Forner**, P., Müller, H., Paredes, R., Rosso, P., Stein, B., Editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization 4th International Conference of the CLEF Initiative*, CLEF 2013.

[2] **Tsikrika**, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E., *The Scholarly Impact of CLEF (2000–2009).*

[3] **Angelini**, M., Ferro, N., Santucci, G., Silvello, G., *Improving Ranking Evaluation Employing Visual Analytics,* CLEF 2013.

[4] **Kim**, J., Lee, J., *Subtopic Mining Based on Head-modifier Relation and Co-occurrence of Intents Using Web Documents*, CLEF 2013.

[5] **Hall**, M., M., Toms, M., *Building a Common Framework for IIR Evaluation*, CLEF 2013.

[6] **Frawley**, W., J., Piatetsky-Shapiro, G., Matheus, C., J., *Knowledge Discovery in Databases: An Overview*, CLEF 2013.

[8] **Polanyi**, M., Personal Knowledge. Towards a Post Critical Philosophy. London: Routledge (1958).

[9] **Nonaka**, I., Toyama, R., Konno, N, “SECI, Ba and Leadership: a Unified Model of Dynamic Knowledge Creation,” Long Range Planning, Vol. 33, Issue 1 (Feb. 2000).

[10] **Hyman**, H., S., Fridy III, W., “Using Exploration and Learning for Medical Records Search: An Experiment in Identifying Cohorts for Comparative Effectiveness Research,” NIST Special Publication, Proceedings: Text Retrieval Conference (TREC) 2012.

[11] **Hyman**, H., S., Sincich, T., Will, R., Agrawal, M., Padmanabhan, B., Fridy III, W., “A Process Model for Information Retrieval Context Learning and Knowledge Discovery,” (under review).

[12] **Berlyne**, D., E., Conflict, Arousal and Curiosity, New York: McGraw Hill (1960).