Thesis

A User-Driven Adaptive Interaction Strategy To Support Exploratory Searching

562

Colofon

Document:	Master Thesis
Version:	1.0
Date:	09/08/2007

Name:	Daniël Rutten
Studentnr:	0424625
Thesisnr:	562
E-mail:	DanielRutten@student.ru.nl

Institution:	Radboud Universiteit Nijmegen
Programme:	Informatica
Department:	Information Retrieval & Information Systems (IRIS)
Supervisor:	Prof. dr. ir. Theo van der Weide
2nd Teacher:	Prof. dr. Mario van Vliet
Address:	Toernooiveld 1, 6525 ED Nijmegen

Abstract

Most Information Retrieval models offer only limited support to searchers who possess only a vague and inexpressible understanding of their actual information need. A search process that begins with this inexplicit understanding is referred to as exploratory searching. This research aims to improve the current state of research surrounding exploratory searching, by providing the searcher with additional support throughout the search process.

Prior to presenting our vision to address this dilemma, some fundamental concepts within Information Retrieval are revisited and expanded upon, based on insights derived from established research within this field.

The presented model for Information Retrieval aims to facilitate exploratory searching by supporting the searcher through adaptive interaction, precipitating the searcher's increasingly accurate understanding of his information need.

To augment this model, its various stages are expanded upon and discussed in greater detail with respect to a possible implementation of the model. Furthermore, a technical analysis and design is made of the model, in order to outline the internal organisation and structure that such an implementation would require.

Keywords: Awareness Development Paths, Consolidated Intermediation Model, Exploratory Search, Human Computation, Information Behaviour, Information Need Awareness, Information Need Levels, Information Retrieval Intermediation, Intermediation Strategies, Search Behaviour, System Awareness, System Awareness Levels

Preface

First and foremost, I would like to thank my research partner, Jeroen Bakker, with whom I've worked on this thesis for the past few months. I've enjoyed working with you and I appreciate all your assistance and support.

Furthermore, the both of us would like to thank Theo van der Weide and Mario van Vliet of the Information Retrieval & Information Systems (IRIS) group for their supervision and support. Additionally, there are several people that have helped us in many different ways, ranging from discussing the research to providing moral support. We would like to thank everyone that helped us in any way.

Table of Contents

1. Introduction

"The mere existence of information does not ensure access to it" - Thomas Mann

"A User-Driven Adaptive Interaction Strategy To Support Exploratory Searching" is the result of a master thesis project executed by Daniël Rutten and Jeroen Bakker, two computer science graduate students at the Radboud University of Nijmegen, under the supervision of prof. Theo van der Weide and prof. Mario van Vliet. This research is keen on improving the state of current research on information retrieval and (in particular) laying a foundation for effective support of exploratory searching. The structure of this document is as follows:

- Chapter 2 covers the *Problem Statement* and includes the research questions.
- An analysis of *Information Behaviour* is provided in chapter 3, which defines the terminology and related concepts used in this research, in order to establish a common vocabulary and eliminate ambiguity.
- Chapter 4 deals with *Search Behaviour*, exhibited by the searcher during his interaction with Information Systems.
- In chapter 5, *Practical Intermediation Models* are considered as a source of inspiration and best-practices of everyday intermediation scenario's, in which a searcher is guided in his quest for information.
- Chapter 6 offers a *Consolidated Intermediation Model* based upon the aforementioned intermediation models, and is designed to assist and guide a searcher in fulfilling his need for information during the exploratory search process.
- Chapter 7, *Application Of Consolidated Model* expands upon the consolidated intermediation model and details several techniques to implement the various stages of the model.
- A *System Design* is offered in chapter 8, in the form of Data Flow Diagrams (DFD) and Data Dictionaries (DD), in an attempt to analyse the internal organisation of a potential implementation of the consolidated model.
- The *Conclusion* of our research and the contribution it offers to the field Information Retrieval is reflected upon in chapter 9.
- Additional *Recommendations,* to further expand upon this research, are offered in chapter 10.

This research builds on and attempts to incorporate multi-disciplinary research done in the fields of Information Retrieval (IR), Information Seeking and Human-Computer Interaction (HCI). As such, it encompasses various perspectives on IR, including the searcher's cognitive-behavioural aspects as well as interface design and user-adaptation.

2. Problem Statement

"The best way to escape a problem, is to solve it."
- Alan Saporta

This chapter discusses the research questions. Some background information is also provided to help establish the context of the research.

2.1. Background

In everyday life, people are faced with problems and situations that need to be overcome by making a concious decision. Typically, the person's initial approach would be to resolve the situation by using his personal knowledge to oversee the problem. However, when this knowledge fails to provide the person with a reasonable approach to the situation, he will have to obtain more knowledge, to get from his current state of knowledge to the desired state of knowledge. The difference between these two states is known as the *knowledge gap*. Bridging this gap is accomplished by having the person assume the role of a searcher, searching for and assimilating information. In today's society, however, there is an abundance of information available to anyone who is interested in it. This virtually unlimited availability of large quantities of information describing any number of topics in varying depths, introduces a serious dilemma: how does the searcher manage to locate the information most relevant to him?

The goal of the searcher's searching behaviour, is to acquire the appropriate knowledge for bridging his knowledge gap. Typically, there are several aspects to consider when 'building' this bridge. For instance, the searcher can ask someone else for information, or try to find this information in one of many information objects (books, papers, articles, etc.), made available by an information repository, known as an Information System.

However, the bigger the knowledge gap, the more effort it takes to bridge this gap. Furthermore, when the problem space is unclear and ill-defined, the exact specifications for the bridge are vague. The searcher's ability to express his need for information depends on the searcher's perception of his knowledge gap and his intrinsic degree of uncertainty. As such, searchers do not always succeed in accurately formulating their information need. Taylor has done some pioneering work, identifying several stages of a searcher's *information need awareness*, referring to them as *information need levels* [TAYLOR1968]; the ability to express one's need depends on the information need awareness they are subjected to.

2.2. Research Questions

Many search-engines and -techniques currently exist; most of them, however, offer only a single form of interaction and do not adapt their strategies to the searcher's inherent information need awareness. Although most existing Information Systems are quite effective when searchers are in a high state of information need awareness, they offer little to no support to searchers with a low information need awareness. As such, the guidance offered to the searcher in the IR process is not necessarily optimal and is subject to improvement. Therefore, this study will focus on *exploratory searching* (see *paragraph 4.1: Exploratory Searching*), in which the searcher is initially unable to formulate a precise query. This suggests that some form of guidance and support during the search process would act as a catalyst for further refining the searcher's information need.

The main research question that can be derived from the aforementioned problem is as follows:

***Main Research Question**: "How can a searcher best be assisted in refining his information need throughout the process of exploratory searching, aimed towards guiding the searcher to the most relevant resources available?"*

In order to ensure effective communication throughout this paper, a common foundation of IR (including vocabulary and semantics) is required to remove any ambiguity of the concepts in our research. This leads to the first research question:

***Research Question 1**: "Which models are used to represent various aspects of Information Retrieval and how do the concepts defined therein relate to one another?"*

To assess how best to assist people during their search process, it is important to understand the nature of exploratory searching and how people search for information in the first place. The following research question can therefore be formulated:

***Research Question 2**: "Which search strategies do people employ to individually fulfil their information need, with respect to their intrinsic information need awareness?"*

To this end we will analyse and compare several key models of the search process in the field of cognitive science and information retrieval, in order to establish the current state of affairs of the information searching process.

In addition to the (commonly unassisted) search strategies derived from research question 2, searchers can also be assisted in their search process by an *intermediary*. To define the strategies that an intermediary can employ, inspiration is acquired from everyday situations in which people are assisted via human-to-human intermediation. This results in research question 3:

***Research Question 3**: "Which practical intermediation models are employed in everyday life, to support searchers in their quest to fulfil their information needs?"*

By using the insights derived from the practical intermediation models, devising a new (consolidated) intermediation model for Information Retrieval may be possible. Therefore, the following research question can be formulated:

***Research Question 4**: "Can the insights derived from the practical intermediation models be translated to the field of digital information retrieval and, if so, can a (consolidated) intermediation model be created?"*

The following set of research questions will continue to build upon the consolidated intermediation model (see *chapter 6: Consolidated Intermediation Model*) and will serve to describe the various aspects of the model in greater detail. First of all, the human computation component of the consolidated model details the acquisition of the human perspective information, such that it can be used during the intermediation stages. This human perspective information will encompass the external indices that apply to individual documents. In order to successfully support a searcher through intermediation, this information will need to be accurate (with respect to the document they are meant to describe) and meaningful (in the sense that they match the perspective of human searchers on the given document). This leads to research question 5:

***Research Question 5**: "How can accurate and meaningful (external) indices be acquired through human computation, that apply to their respective documents?"*

Once a set of relevant external indices is known (through the requirements acquisition stage of the consolidated model), this set needs to be translated into the corresponding internal indices, such that these internal indices can be used to describe and execute a query on the documents within the IS's repository. However, since the relationship between internal and external indices is unknown, how can this translation be accomplished? This leads to the following research question:

***Research Question 6**: "Based upon the document and Human Perception Information of a given document, how can a searcher's external indices be converted into their respective internal indices?"*

With all of the aspects of the consolidated intermediation model described in detail, it becomes possible to create a design for a possible implementation of the consolidated model. Using Data Flow Diagrams (DFD) and Data Dictionaries (DD), the various data-flows and data-processes within such an implementation can be identified and described. This is addressed by research question 6:

***Research Question 7**: "What data-flows and data-processes can be identified from the interaction, to and from the various components of an IS built according to the consolidated model?"*

3. Information Behaviour

"Information is not knowledge." - Albert Einstein

In order to establish an unambiguous understanding of IR, some concepts (and inherent relationships) will first be defined, to allow for a unified context throughout this document.

3.1. Information Representation

Information Retrieval pertains to the process of a searcher requesting for information, followed by the searcher's consumption of that information. Once consumed (i.e. read) and interpreted by the searcher, the information becomes knowledge. The various relations between these concepts are as follows:

- **Data**: The raw material used as a vessel to store information in a given storage format (data structure).
- **Information**: The presentation of knowledge-concepts (usually through natural language) in a way that should enable knowledge gain. The latter can be achieved by the consumption of the information.
- **Knowledge**: A searcher's actual awareness and understanding of information, such that it can be applied when appropriate. Knowledge representation can be typed as either declarative (facts) or procedural (routines or sequences to perform a specific task).
- **Wisdom**: Some would argue that knowledge is superseded by wisdom, which has the added effect of being able to extrapolate and infer new knowledge from existing knowledge, through logical reasoning.

For example, a spreadsheet can contain various financial or other numeric data. Once (part of) the data is represented in a way such as to be useful (e.g. the sum of assets or the operating results), it becomes information. Reading and interpreting the information allows a searcher to become aware of it (knowledge), and act according to its implications (wisdom).

This classification strengthens the discerning of the following forms of retrieval [BAEZA-YATES1999]:

- **Information Retrieval**: In Information Retrieval, the searcher's information need is expressed as an *information request.* This request is often specified in the form of natural language, which is less structured and possibly ambiguous. As such, a document may not exactly be a complete match for the (vague) request, but may still be relevant. In Information Retrieval (IR) such a document would still be selected; based upon the request and the IS's interpretation of the documents, the relevance of each document is determined, which dictates whether or not a document is returned.
- **Data Retrieval**: In the case of data retrieval, the information need is expressed in the form of a *query*. The results are only deemed relevant when they exactly match the query's conditions. Obtaining a high precision is made possible due to the highly structured nature of the data, allowing for a specific instruction to only return the data containing information that's relevant for the information need.

Obviously, the difference in forms of retrieval leads to a significant dissimilarity between information requests and queries; formulating a query has a higher threshold than the formulation of a request, because the searcher needs to have extensive knowledge of (the internal structure of) the IS. Due to the fact that this research is not confined to dealing only with expert searchers (capable of formulating precise queries to reflect their information need), the focal point is on Information Retrieval.

3.2. Context of the Problem Space

With respect to the typical Information Retrieval (IR) process, the previously defined concepts relate to each other, the problem space, and the IS, in a certain way. We introduce the following schematic to illustrate their inherent relationships:

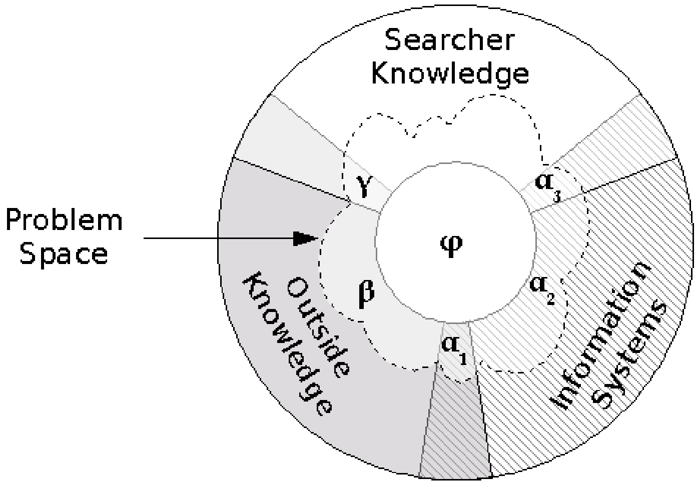


Illustration 1: Problem Space Context

- **Searcher Knowledge**: The total sum of knowledge that the searcher already possessed prior to issuing an information request on a given Information System (IS).
- **Information Systems**: The total collection of information within the reach of an IS. When executing a searcher's query, the result-set it yields will be a subset of this collection. Note that an Information System is defined as an agent, capable of providing information based on an information request. Consequently, the semantic scope of an IS in this research is more general than that of a *digital* Information Retrieval system.
- **Outside Knowledge**: The combined knowledge of all human beings, except the searcher. This could be interpreted as the 'society' the searcher is surrounded by, with respect to understanding and knowledge concepts.
- **Problem Space**: The information pertaining to the given problem space, that is required to overcome the related problem. The part of the problem space that falls outside of the searcher's knowledge is known as the *knowledge gap*. Note that some of the aspects of this area may still be unknown and/or unexplored. As such, part of this category may fall outside of the boundaries of all available information (as indicated by the problem space extending beyond all available information in the illustration).

As seen in the illustration, all of the above areas have some degree of overlap. The following overlapping areas are discernible:

- **Common Knowledge (γ)**:
  This area is defined as the intersection of searcher knowledge and outside knowledge. In contrast, the union of these areas defines all knowledge constructs that are present in the world.
- **Published Personal Knowledge (α_3)**:
  This area is defined as the intersection of searcher knowledge and information from the Information Systems. In contrast, the union of these areas represents the knowledge state a searcher can achieve by interacting with the Information System.

- **Published Outside Knowledge (α_1)**:
  This area is defined as the intersection between outside knowledge and (the knowledge published as information in) Information Systems. In contrast, the union of these areas represents all possible expansions of the searcher's knowledge, although outside knowledge constructs first need to be translated to information (and consequently, ported to an Information System), before the searcher can absorb them.

Although not apparent from the illustration, the intersection of searcher knowledge, outside knowledge and Information Systems represents general concepts, which are common knowledge *and* published as information in Information Systems. Due to the trivial nature of this category, it is omitted from the schematic.

The center of the problem space which does not intersect any of the 3 outer segments (φ), is the area of knowledge that is (as yet) unexplored and thus unavailable. An example of this type of knowledge is found in absolute facts, such as the relativity theory: although this theory was once unknown to mankind, the underlying effects have always been apparent in the universe. Once the theory describing this phenomenon was conceived, this absolute fact was moved from the area of unexplored area (φ) to the outside- or common knowledge. In relation to this illustration, wisdom can thus be perceived as a function to use existing knowledge to create more (inferred) knowledge, thereby exploring this area of unknown knowledge, and confining it even more. However, one could argue that new insights can also possibly lead to more questions, thereby even deepening the area of unexplored knowledge.

It is important to note that the above illustration is to be considered disproportionate and unspecific, because the ratio between the areas is ill-defined and subject to several parameters. For example, some situations might offer a relatively small and well-explored problem space, with all possible solutions to that problem being known to man. In this case, there is no unexplored knowledge (φ) regarding the Problem Space. Similarly, the contribution of an IS in terms of conveying information about concepts present in outside knowledge, could likely be higher than that of conveying information not present in outside knowledge (i.e. knowledge concepts being dormant in society, yet preserved in overlooked information).

The typical course of events leading to search behaviour, involves a person acknowledging his knowledge gap. The knowledge gap is the part of the problem space that is not fully present in the searcher's combined knowledge of the problem space. The person might experience contradicting thoughts and beliefs (a need for confirmation), or incomplete thoughts and beliefs (a need for clarification). In order to address this knowledge gap, the person defines the related problem space to the best of his knowledge: this problem space is described in terms familiar to the person. Then, the person establishes an approach to resolve his knowledge gap in the problem space and exhibits search behaviour (see *chapter 4: Search Behaviour*). Resulting from the search behaviour, information will be presented to the searcher, who then analyses and incorporates the relevant information into his personal knowledge. Depending on both the significance of the knowledge gap for the problem space and the complexity of the problem space, the searcher will at some point determine he's obtained sufficient knowledge to resolve his problem. If the searcher was unable to acquire the knowledge he needs, he might choose to continue the search, or abandon it altogether.

3.3. Information Relevance

There are various perspectives on which to judge the relevance of given information. The following paragraphs will explore two of those aspects:

3.3.1. Personal and Semantic Relevance

Because searchers by definition can not assimilate knowledge directly, they need to obtain information instead. Therefore, the searcher derives an information need from his knowledge gap, which he then needs to specify to an IS as a request. The resulting documents of this request will then fall in one of the following categories:

- **Semantic Relevance***:* When issuing a request to a given IS, the resulting set of relevant documents will fall within area α (α_1 U α_2 U α_3). Documents that pertain to this area are known as the *semantically relevant documents* [STOJANOVIC2005], that the IS is capable of offering with respect to the given problem space. These are the documents that are relevant with respect to the searcher's request. However, part of this area (denoted as α_3) falls outside of the searcher's sphere of interest, since the concepts from this information are already known to him (to some degree).
- **Personal Relevance***:* Documents pertaining to area (α_1 U α_2) are known as *personally relevant documents* [STOJANOVIC2005], that the IS is capable of offering and that the searcher still needs, in order to better comprehend the given problem space. This area defines the searcher's actual information need that can be fulfilled by the IS: the knowledge constructs from within the outside knowledge are unavailable to the IS, and the exclusion of area α_3 corresponds with the fact that the searcher already possesses the knowledge constructs contained therein.

Most commonly, Information Systems are designed to yield the semantically relevant documents; documents that match the query in terms of the system's repository and domain model, rather than only those those documents that are personally relevant to the searcher's information need. As such, the searcher might be disappointed, as those results may have a big overlap with his present knowledge (α_3).

3.3.2. Relativity And Satisfaction Relevance

Following the issuing of the formulated information need, the IS will typically evaluate the input, and present the searcher with a number of (presumed) relevant resources. Consider a collection of documents A, B, C, D, E and F, given a specific problem space. The picture below represents the overlap and relation between documents and problem space:

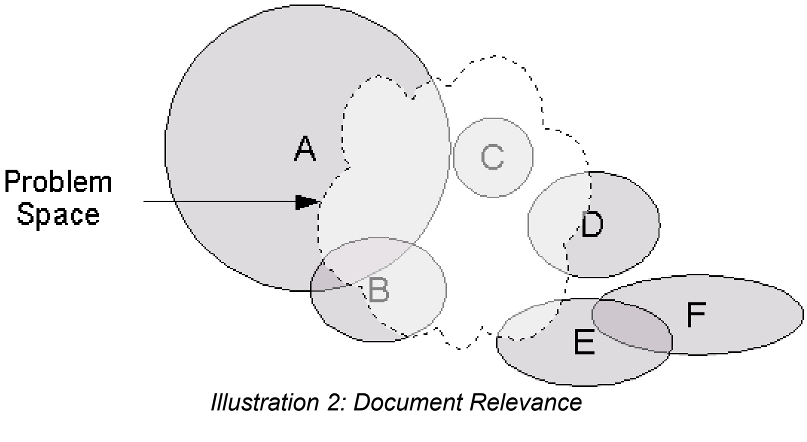


Illustration 2: Document Relevance

In light of this schematic, we can propose that there are two different measures of document relevance:

- **Satisfaction**: Documents are ranked with respect to the part of the problem space they affect. In this case, those documents that satisfy the largest part of the searcher's knowledge gap, are considered to be most relevant. This perspective on relevance focusses on the effectiveness of a given document in fulfilling the searcher's information need. In the above illustration, the ranking would be: A, B, C, D, E.
- **Relativity**: Documents are ranked with respect to the ratio of the part of the document that intersects with the problem space, to the part that doesn't. The documents of which most of their content falls within the problem space, are considered to be most relevant. In other words, this measure of relevance focusses on the efficiency of the document's ability to fulfil the searcher's information need. In the above illustration, the ranking would be: C, B, D, A, E.

In relation to *Illustration 1: Problem Space Context*, obtaining new knowledge from documents causes the concepts therein to shift from one part of the classification (area $\alpha_1 \cup \alpha_2$) to the published personal knowledge area (α_3). This expansion of the latter area entails the searcher gaining knowledge, at the expense of the IS' population of searcher-relevant information: information already conceptualised in the searcher's knowledge is henceforth deemed irrelevant.

Thus, exhaustive knowledge acquisition in terms of a given IS, relates to expanding the personal knowledge to the extent of the IS being void of relevant resources to expand this area. At this point the searcher would have extensive knowledge of this IS's repository and may choose to consult another IS if the information need is not yet satisfied.

3.4. Information Retrieval Process

Generally speaking, Information Retrieval is performed by a searcher, interacting with an IS, in the searcher's quest to find information that will fill his knowledge gap. When further dissecting the various elements of an IS, the following schematic representation of the IR process can be obtained:

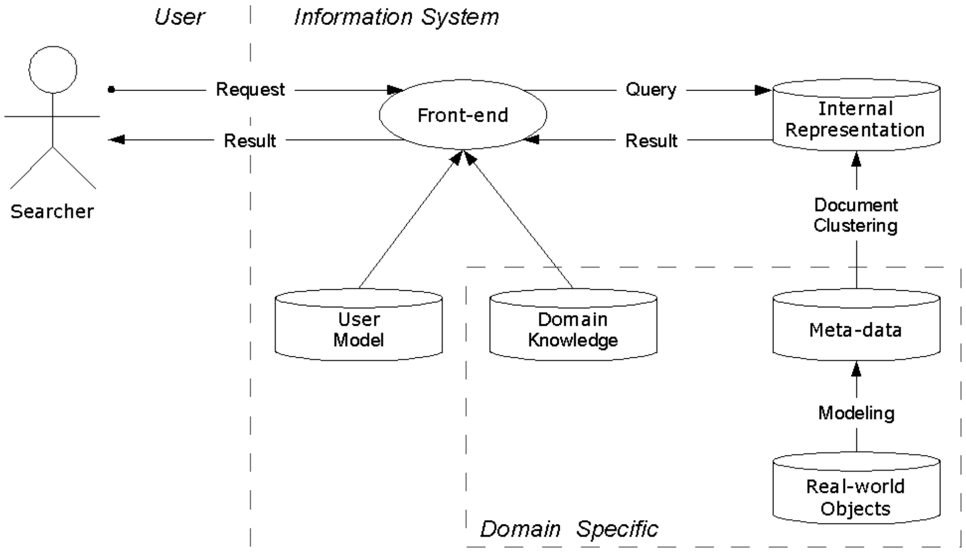


Illustration 3: The Information Retrieval Process

In the illustration above, interaction occurs between the searcher and the IS' front-end (a form of information retrieval), which interprets the searcher's information request and executes a query (a form of data retrieval) on the internal representation of the meta-data. The query's results can then be analysed and used to help satisfy the searcher's information need. Also, in accordance with developments in the field of User Modelling, the front-end may be able to make use of a user-model (involving various details of the searcher's background) and other domain knowledge (i.e. business rules applicable to the given domain) to further narrow the set of results that is offered to the searcher.

The internal representation can be arrived at through document clustering, by making use of techniques such as Latent Semantic Indexing, Self-Organizing Maps or other techniques. This however falls outside the scope of this research.

4. Search Behaviour

"Seek, and ye shall find." - Bible, Matthew 7:7

This chapter serves to make an analysis of the aspects that affect the searcher's state during the various stages of the exploratory search process, and how this expresses itself in actual search behaviour.

4.1. Exploratory Searching

Exploratory searching [WHITE2005] is a search process, in which the searcher is (initially) unaware of the exact nature of his information need and does not have a defined target. This implies that the searcher is then also unable to express or formulate this need, due to a lack of knowledge of the problem space. More explicitly, during such a process the searcher's state is not constant, but changes continuously, as the searcher gains a increasingly better understanding of his knowledge gap. Several of these states, all of which refer to a searcher's *information need level* have been named by [TAYLOR1968] which are defined as follows:

- **Visceral Need**: A vague sense of dissatisfaction, due to an actual, but unexpressed need for information.
- **Conscious Need**: A conscious mental description of the need can be made, in the form of a (possibly ambiguous) narrative or as examples/analogies.
- **Formalized Need**: The need can be formalized as a rational question-statement or a topic with clear and concise boundaries.
- **Compromised Need**: The representation of the need within the constraints of the system and its information repository.

When looking at it from the perspective of fuzzy logic, the information need awareness can be considered a linguistic variable, which can take the value of any of the information need level denominations mentioned above.

As the searcher consumes information, his knowledge of the problem space increases. This also causes a shift from the searcher's anomalous state of knowledge ([BELKIN2005], cf. knowledge gap). With each consecutive need level, the searcher's ability to express his information need (and the corresponding accuracy) is improved. Note that a search process need not necessarily traverse each of these stages; a searcher might either not be sufficiently motivated to continue searching or might have fulfilled his information need in one of the earlier information need levels through serendipitous discovery (i.e. stumbling upon the required relevant information by chance).

Several of the *information seeking strategies* (ISS) from [BELKIN1993] are able to represent the various information need levels. Each strategy is represented as a set of values within the following set of dimensions:

- **Goal of Interaction**: The goal of the interaction might be to *learn* about (characteristics of) an object, or to *select* useful objects for retrieval.
- **Method of Interaction**: This dimension pertains to either *searching* (looking to find a known object) or *scanning* (looking for something interesting in a collection).
- **Mode of Retrieval**: The mode of Retrieval can either be by *recognition* (identifying relevant items while stimulated by the possible options), or by *specification* (looking for identified objects).
- **Resource Considered**: Finally, this dimension describes the resource focus, being either the actual object (*information*), or information about this object (*meta-information*).

In [BELKIN1994] the set of dimensional values [Learn, Scan, Recognize, Meta-information] is described as "*a situation in which a person needs to learn about characteristics of the knowledge resource before the information search can begin. This can also be understood as the ISS associated with an unformulated and unspecified information problem.*"

This accurately represents the visceral need level. Also, the set [Select, Search, Specify, Information] is described as "*a prototypical example of a well understood information problem, in which the goal of the interaction is not to learn about the system, but to select items which can be specified by the user.*", which depicts a formalized/compromised need level.

Similarly, we can use the information seeking strategies to differentiate the various aspects of each of the other information need levels:

Need Level	***Goal***	***Interaction***	***Retrieval***	***Resource***
Visceral	Learn	Scan	Recognize	Meta-information
Conscious	Learn/Select	Scan/Search	Recognize	Information
Formalized	Select	Search	Specify	Information
Compromised	Select	Search	Specify	Information

4.2. System Awareness

Note that the aspects of the formalized and compromised levels only differ in the form in which the need for information is communicated. In the compromised need level, the information need is expressed in terms of the IS; the searcher understands the language used in interacting with the IS, as well as the structure of the documents within the IS. The compromised need level, therefore, implies a greater familiarity with the IS' interface and the available collection of documents.

Because of this difference, we propose that the compromised need is not an awareness level, but rather a state of being able to translate the high information need awareness (formalized) to an IS.

Due to this clear distinction between a searcher's information need awareness and his awareness of the system, we introduce an additional dimension, called *system awareness*. This dimension positions the aforementioned information need levels as follows:

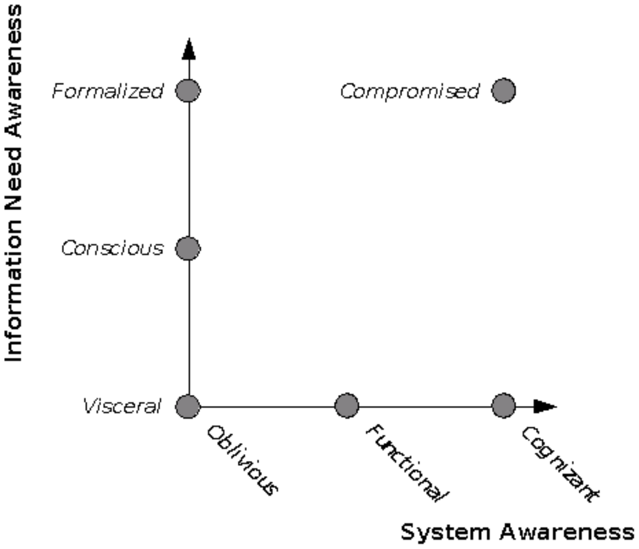


Illustration 4: Information Need Awareness vs. System Awareness

Much like the various levels of information need awareness, system awareness can also be considered as a linguistic variable, the value of which can range through the following denominations:

- **Oblivious**: Initially, the searcher lacks any knowledge of both the (structure of the) repository that the IS can deliver and the language used to interact with the IS.
- **Functional**: Once the searcher has gained some understanding of the language used by the IS, the searcher is increasingly proficient at expressing his perceived information need to the IS.
- **Cognizant**: In this level of system awareness, the searcher is 'fully informed': this implies that the searcher has a sufficient system awareness to formulate his perceived information need as a IS-specific query.

Note that by this definition, it is prerequisite to have a thorough insight into the language used by the IS, prior to gaining sufficient understanding of the IS's repository.

From this we can infer that the search process of a given searcher can be described as a path between these two awareness dimensions; such a development path can be formed with respect to the progression of the searcher's information need awareness and system awareness. These *Awareness Development Paths* (ADP's) will vary from person to person. We will illustrate this concept through several exemplary ADP's, numbered 1 through 5 in the following illustration, showing the development of the searcher's awareness dimensions during the interaction with a given IS:

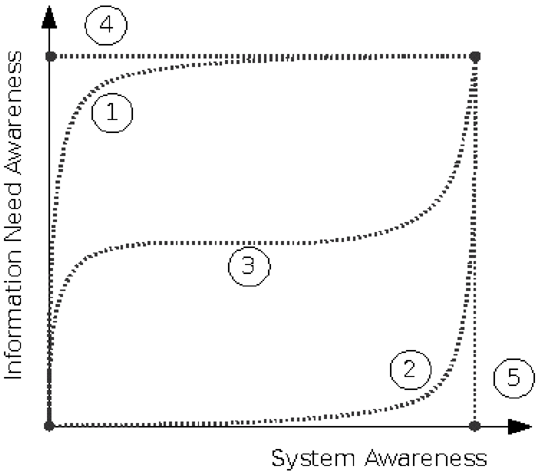


Illustration 5: Exemplary Awareness Development Paths

In the above illustration, the following ADP's are identified:

- **ADP1 - Taylor-esque**: This pattern adheres to Taylor's classic progression of need levels: the initial part of the interaction between the searcher and the IS is focussed on increasing the searcher's awareness of his actual information need, followed by the increasingly accurate communication of this need in terms of the IS. This path is typical for people who are unable to express their information need (and thus have a low information need awareness) and strive to use the information system to gather more information in general areas of interest. Once they become more aware of the exact nature of their information need, they will formalize a specific request to the system. By doing so, they will learn more about the system (and increase their system awareness) and consequently be more proficient in expressing their information need in terms of the IS.
- **ADP2 – Experimental Search**: Once the decision for a particular IS has been made, some people will aim to gain system knowledge prior to increasing their information need awareness. In effect, they are postponing the resolution of their information need, in favor of experiment with the IS, to increase their understanding thereof.
- **ADP3 – Arrested Development**: Quite often, people will fail to attain a sufficiently high degree of information need awareness and are therefore unable to express their need satisfactorily. As a result, the progression of their information need awareness will stagnate. When this occurs, people tend to attempt to formulate *tentative queries* [WHITE2005] and present these to the system and, by doing so, gain increased system awareness. This will continue until the searcher experiences a breakthrough in his information need awareness, which is enabled by one of the results from the tentative query.
- **ADP4 – Formal Searcher**: For a given information need, people will always traverse the lower stages of information need awareness (e.g. visceral and conscious). However, this need not necessarily be done in unison with an IS; a searcher may have been able to acquire an accurate perception of his actual information need, even prior to using an IS. As such, with respect to the IS, the searcher is already in the formalized need level and will use the IS to (learn to) express his need in terms of the IS.

- **ADP5 – Skilled Searcher**: Conversely, people may have a high degree of system awareness, yet fail to understand the exact nature of their information need. This can occur when the searcher is in fact an expert of the domain in question, but is unable to make a choice in what would best suit his needs. This can occur when the searcher is confronted with too many options (as a form of information paralysis) and requires additional help to make a decision.

The set of awareness development paths is practically endless, as a given searcher can start from any arbitrary point in the spectrum of information need awareness and system awareness, and proceed from there. The path between these dimensions indicates a progression of the searcher's awareness, as well as time.

4.3. Search Strategies

The searcher's degree of information need awareness calls for a specific search strategy, that the searcher will employ to further increase his information need awareness. Various models offer different types of search strategies (such as [WEICHOO2000], [SONNENWALD2001] and [KUHLTHAU1993]), but fail to connect these to either the searcher's information need awareness or related information need level. For each of these models, the various search strategies relate to the information need levels as follows:

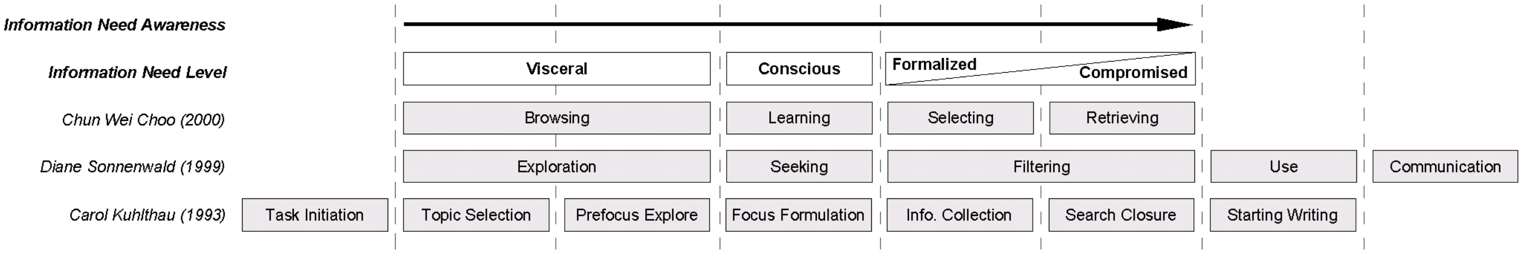


Illustration 6: Search Strategies Related To The Information Need Levels

Note that some of these information strategies aren't part of the user's search behaviour, as they precede (*Task Initiation*) or succeed (*Use, Starting Writing* and *Communication*) the search for information. As such, they are not considered part of search behaviour, but rather as information behaviour.

From all of these behaviour models, one unified model emerges, offering 4 general strategies that each confers with one of Taylor's need levels. Although these strategies vary in name for each of the aforementioned models, they each describe the same aspect of the user's search behaviour. For the unified strategies, we will reuse the terminology used by Choo, for the close resemblance with the terminology as used in this research:

- **Browsing**: The searcher has general areas of interest, but is unable to describe them. In this strategy the searcher attempts to improve his perception of his information need, through negotiation with the IS. This is accomplished via recognition of relevant information and its underlying characteristics; more specifically, by scanning a broad range of sources and judging them based on recognizing terms and concepts from the easily accessible (meta-)information.
- **Learning**: The searcher aims to further increase his knowledge of the problem space, by actively searching for topics of interest, within the general areas of interest that resulted from the previous strategy (Browsing). His understanding of the extent of the problem space has somewhat increased, but the information need awareness is still somewhat blurred. This results in a broad and unrefined specification of his information need.

- **Selecting**: Having increased knowledge of the problem space, the searcher is able to compose an accurate description of his information need. As such, the searcher will formalize a plan in the form of a formal description of the searcher's information request.
- **Retrieving**: The searcher's precise description of his information need is adapted to the information system, generally aiming for high precision, at the expense of recall. In this scenario, the personal relevance of the results will closely relate to the semantic relevance thereof.

4.4. Search Flow

Based on models from [MARCHIONINI1995], [TAYLOR1968], [WILSON1997], [BRUCE2005], [NIEDZWIEDZKA2003], a unified flow chart modelling the search process is derived:

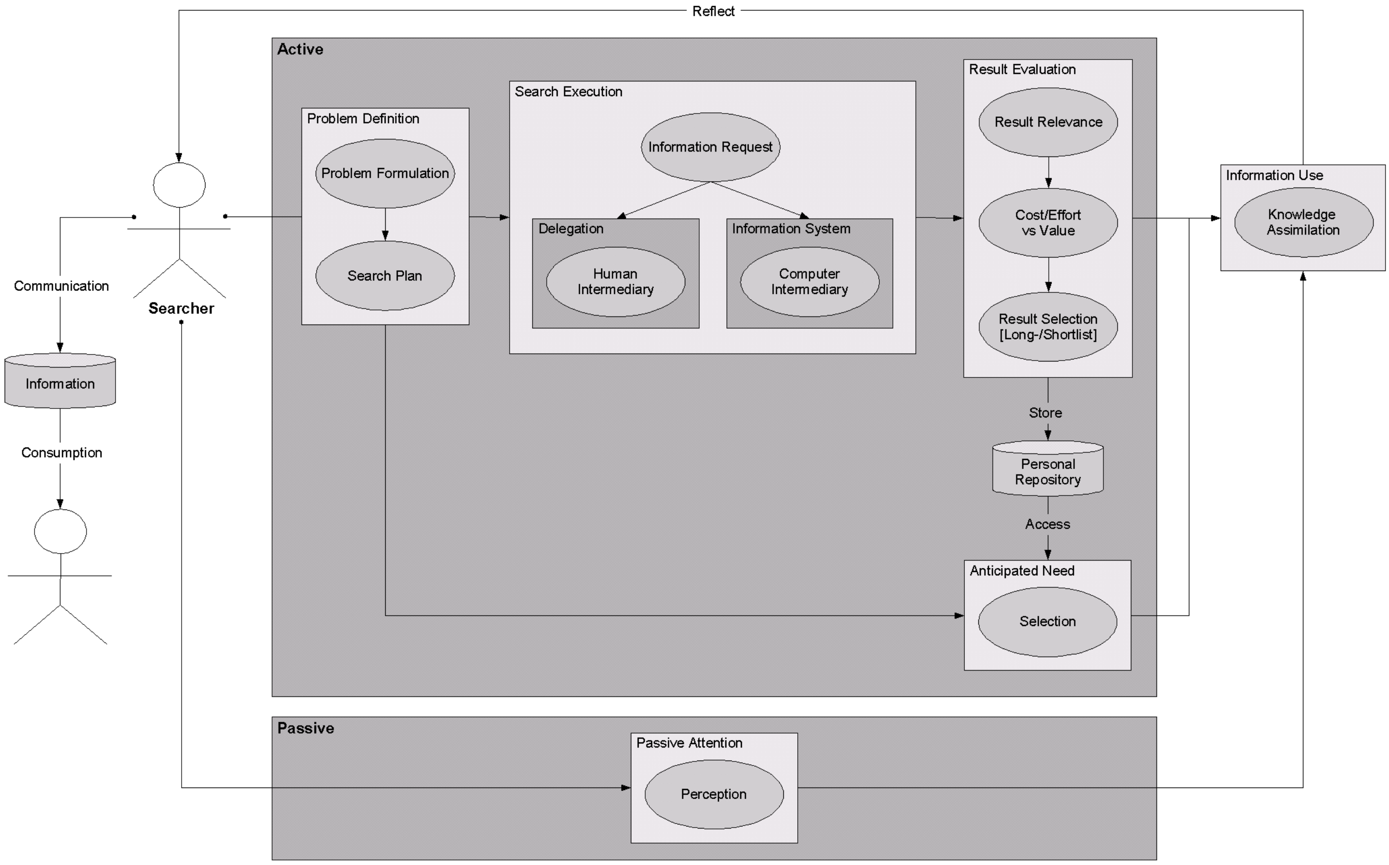


Illustration 7: The Search Process, As A Flow Of Operations

The unified search flow chart encompasses suggestions made by the individual contributors, the most significant segments of which are as follows:

- **Active Search**: A searcher who experiences a need for information will initially try to define the problem he faces. This is done by composing some type of plan to fulfil his knowledge gap and consequently trying to express or formulate his problem, to be used for the information request. The actual search execution can take two forms ([NIEDZWIEDZKA2003] and [TAYLOR1968]:

  - **Information System**: By conveying his information request to a digital IS, the searcher initiates the search process (as represented in illustration 3: *The Information Retrieval Process*) and acquires the results of his request.
  - **Delegation**: Rather than engaging in the actual search himself, the searcher may delegate this task to someone else, by conveying the information request to a human intermediary. This intermediary will then act based upon this information request and (essentially) perform the entire search process recursively instead of the actual searcher. In this recursion, the intermediary will become a searcher himself, formulating his own search plan and executing it. The results thereof are then returned to the original searcher. Further insight into this process is provided in *chapter 5: Practical Intermediation Models*.

  The result of the search execution is then evaluated; firstly on the relevance of the information and secondly on the cost or effort to consume the information. Ultimately, based upon these evaluations a selection of the information resources is made and consequently consumed.
- **Passive Attention**: In contrast to active search, *passive attention* [WILSON1997] is not initiated by the searcher. Although the searcher experiences (but not necessarily perceives) an actual knowledge gap, he is not actively searcher for information, but rather absorbed it passively (e.g. listening to the radio or from an advertisement).
- **Anticipated Need**: [BRUCE2005] describes the concept of *anticipated need*, in which a searcher encounters information resources that fulfils an information need he does not yet have. Instead of consuming them, the searcher can choose to *store* the resources (or references thereto) in a personal repository, based on the assumption that they will prove relevant later on. When the need (that the searcher anticipated he would have) emerges, the searcher is able to formulate a search plan involving *accessing* and selecting resources from his personal repository. An example of such a personal repository are bookmarks or favourites, in an internet browser.
- **Communication**: Once given information has been consumed (and has become knowledge) it can be communicated to someone else. To do so, it first has to be reformulated and expressed as information (e.g. writing, speech, etc.). This information can then be consumed by others.

5. Practical Intermediation Models

"Those that won't be counselled can't be helped."
- Benjamin Franklin

Besides individual (unassisted) search behaviour, people in need for information can be assisted during their search process. In everyday life, such scenario's are quite common, yet (digital) Information Retrieval often ignores this aspect of the search process. This chapter explores the use of intermediation strategies to support each of the searcher's search strategies for digital Information Retrieval, and seeks inspiration from (human-to-human) practical intermediaries that exist in everyday life.

5.1. Intermediated Information Systems

Most existing Information Systems only employ one form of interaction to interact with the searcher, regardless of the searcher's information need awareness. Given that the searcher's degree of information need awareness will result in different search strategies (see *paragraph 4.3: Search Strategies*), this would imply that the support that the IS offers to the searcher is not always optimal and therefore subject to improvement. Improvement can be made by employing different intermediation strategies for the different stages of a searcher's information need awareness and system awareness, allowing the intermediary to accurately address the searcher's current state in his Awareness Development Path. A more detailed conceptualisation of the various intermediation strategies between the searcher and the intermediary can thus be illustrated as follows:

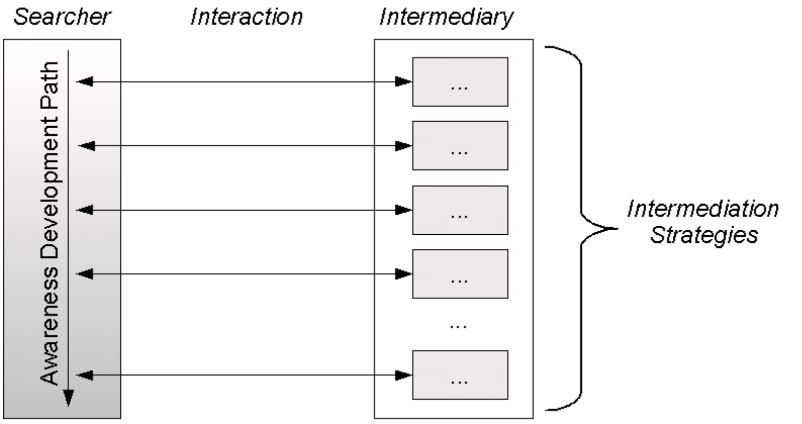


Illustration 8: Consecutive Intermediation Strategies During Development Of Awareness

Note that there can be more intermediation strategies than the information need levels as defined by Taylor, since these levels are not formal categories, but actually shade into each other. In the above scenario, each intermediation strategy continues until the searcher has reached a higher information need- or system-awareness. For example, a searcher who's operating in a visceral level might benefit from an intermediation strategy aimed at brainstorming [VANVLIET2007], whereas the same strategy is considered too cumbersome for a formalized need level, which in turn would benefit more from intermediation aimed at providing structure and abstraction of the concepts in the problem domain.

This implies that the system must have a means to detect what information need awareness the searcher is currently affected by, in order to know when to apply a different intermediation strategy. Rather than asking and interrupting the searcher (who might not even know the answer himself), this information should be derived from the searcher's actions and behaviour. Because the searcher's need level is dynamic and "*these four levels of question formulation shade into one another along the question spectrum*" [TAYLOR1968], detecting the searcher's information need level is not (or even vaguely) formally described.

Depending on the information need awareness, people may employ different strategies to search for information (see *paragraph 4.3: Search Strategies*). In real-life, there are several scenario's in which people find themselves in need for information and, due to the uncertain nature of their need, will consult an intermediary to aid in their quest for information. By studying such scenario's, in which people engage in exploratory searching, we are able to extract successful intermediation models for assisting and guiding a searching to obtain a higher information need. In the following paragraphs, several intermediation models employed in practice, by real-life intermediaries, are described. Each of those models consists of various intermediation strategies, that apply to the varying stages of the searcher's information need awareness.

5.2. Physicians

Patients that have a physical ailment, will most often be unaware to the exact condition that affects them and will only be able to describe its symptoms. Much like exploratory searching in IR, patients may seek the help of a physician to aid them in discovering what is affecting them (their knowledge gap) and apply the appropriate treatment (the applicable relevant documents). From the model of [LARSEN1997] the following intermediation strategies that are employed by physicians to treat patients are derived:

- **Observation and Interpretation**: In this stage, the patient is communicating his awareness of the symptoms to the physician. The physician in turn listens to and observes the patient, to start establishing a mental model of the patient's condition. Finally, the physician summarizes the patient's story, to make sure his message was communicated successfully. This stage results in the patient feeling understood (reducing anxiety), while providing the doctor with a general sense of the problem context.
- **Interview and Examination**: The physician tries to clarify the problem context, by linking this general view to his intrinsic medical knowledge. For example, the patient could only be aware of a subset of all his symptoms, so the physician will interview and examine the patient, to find and/or rule out possible other symptoms. Once the physician has gathered enough information pertaining the condition, he will match the problem context to medical knowledge (either already available, or to be obtained by sessions of information seeking), and come up with a diagnosis.
- **Devise Treatment**: The diagnose links the found problem context to comparable problem situation(s), which have been established through medical research. This research might have devised several relevant treatments for the problem condition, some of which may be inappropriate in the patient's personalized context (due to allergies, medical history, etc.). The physician comes up with a treatment, compatible with the patient's situation.
- **Apply treatment**: The treatment is conveyed to the patient, and applied to remedy the patient's ailment. The application can occur under the physician's supervision, or at a scheduled time by the patient himself.
- **Treatment Monitoring**: A follow-up appointment will be scheduled, at which time the physician can determine the effect of the treatment on the problem, and adjust the treatment accordingly.

5.3. Salesmen

It is the task of the salesman to aid customers, that are in need of a certain product to fulfil a certain task, in discovering what product best suits their needs. Applying the analogy to Information Retrieval, the salesman will interact with the customer to discover the exact nature of their need (their information need) and consequently recommend applicable products (relevant documents) to satisfy the need. The models of [BERGMANN2002] and [AGMRC2007] enable the translation of the analogy of the salesman to the field of Information Retrieval, by introducing a virtual sales assistant that uses his knowledge of products *and* of clients to recommend suitable products. From this model we can derive the following consecutive intermediation strategies that a (virtual) salesman would employ:

- **Experience**: The salesman needs a comprehensive understanding of the products in his assortment and the features that each offers. Also, the salesman needs to be able to relate a customer's need to the applicable features. This understanding can come from studying the assortment, as well as experience with customer interaction.
- **Initial Contact**: The salesman needs to initiate contact with the customer and offer to help him achieve his goal (i.e. fulfil his need).
- **Requirements Acquisition**: The salesman interviews the customer and attempts to ascertain the nature of their need. Obviously, the customer will not wish to answer an endless list of questions. This implies that the questions must be as discriminative as possible, in order to effectively narrow down the set of applicable products. In order to do this, the salesman will need to use his knowledge of the store's assortment to generate relevant questions. Also, the salesman will need to be able to translate his knowledge of the assortment to terms and concepts that the customer will understand, as the customer need not necessarily be an expert in the domain of such products. Generally, there are two tactics for capturing and formalising the customers information need [SHIMAZU2002]:

    - **Navigation-by-Asking**: The salesman generates a dialog, based upon the features of the products in the available assortment. The customer is asked to choose between features that interest him, thereby navigating through and reducing the set of features of the assortment.
    - **Navigation-by-Proposing**: Based upon some initial information derived from the customer, the salesman is able to propose products to the customer. Based upon this proposal the customer is able to navigate through the assortment by evaluation the proposing product with respect to his need. The customer can, for example, explain that he wants something cheaper or more powerful. Based upon these explanations, the salesman is then able redefine his perception of the customers information need and make additional proposals.

- **Product Search**: Based upon the salesman's perception of the customer's need, he can evaluate the alternatives from the assortment and create a consideration set. Generally, there are two tactics to evaluate the applicability of given products:

  - **Content-based Approach**: Based on the extracted features from the requirements acquisition, the sales man is able to determine the similarity between these extracted features and the products from the assortment. The products that exhibit the greatest similarity can be included in the consideration set.
  - **Collaborative Approach**: This tactic emerges from "*the potential for a collaborative approach to product retrieval where users are matched to products based on past behaviour or consumption history*" [BERGMANN2002]. As such, the salesman's experience with other customers, who exhibited a similar need, can be used to elect products into the consideration set. This similarity can be determined by looking for correlations between customers in terms of their ratings assigned to items in a user-model.

- **Product Presentation**: At this point the products from the consideration set are presented to the customer. The type of product presentation can vary upon the customers individual information need as well as the type of the product (text, sound, video, etc.). Also, "*the presentation form of an information entity should be chosen so that the respective message is communicated with a minimum effort of the user*" [BERGMANN2002]. As such, the customer should be enabled to easily browse through the products, review them, and compare two or more different products.

  Note that the requirements acquisition and the product presentation are repeated in a cycle, which may be repeated several times before a satisfactory product is found. As such, "*requirements acquisition and product presentation often cannot be completely separated*" [BERGMANN2002]. This implies that, even during the product presentation, the customer is able to further refine the set of products and provide additional feedback on a given product, resulting in a new product search and consideration set.
- **Sale**: When one of the products from the consideration set is accepted by the buyer, the sale is confirmed. The customer supplies the required (economic) cost and the seller delivers the product in question.
- **Evaluation** (optional): The salesman can check to verify the customer's satisfaction with the delivered product. If needed, the product can be revised, based on the customer's feedback. Additionally, this step provides reflection (for future deals) for both salesman and customer.

5.4. Librarians

In [TAYLOR1968], Taylor describes the '*communication between inquirer and librarian during negotiation process*'. Five stages (know as "*filters*") are identified, that the librarian uses in his intermediation task:

- **Subject Definition**: By defining the subject, the librarian can more accurately delineate his search space, in order to optimize his scope for the retrieval of (possibly) relevant answers. Once the subject's scope is relatively clear, the librarian might be inclined to '*make a brief search to determine the extent of the subject*', to determine the extent of the subject, based on the information available to him.
- **Objective and motivation**: Understanding the motivation behind the searcher's request for information also helps the librarian in verifying and adjusting the established subject definition. Additionally, the insight in the searcher's point of view influences the '*shape, size and form of possible answers*'. Finally, searchers often '*cannot define <u>what</u> they want, but they can discuss <u>why</u> they need it*', making the motivation behind the search the only concrete source of direction for the librarian.

- **Personal Background of inquirer**: Knowledge about the searcher's context in terms of intrinsic knowledge, awareness and familiarity with the library has relevance to the negotiation process. It for example determines the urgency, dialogue level and critical acceptance of results. Having interacted with the searcher before provides some inside knowledge that will ease the dialogue.
- **Relationship of inquiry to file organization**: The librarian "*becomes a translator, interpreting and restructuring the inquiry so that it fits the files as they are organized in his library*". It is up to him to transform the information request into a system query to allow for an efficient system search, yielding the right answers for the searcher. Aside from using the library system, the librarian might be inclined to use (refer to) additional sources of information, e.g. other librarians, or recently used information.
- **Provide acceptable answers**: The librarian must address the searchers' information need by providing information that is correct (relevant) for the searcher. The searcher has a certain expectation of acceptable answers in terms of format, size, completeness, soundness, etc. The librarian must provide the right amount of information in the right format to address the searcher's information need.

5.5. Evaluation

It is evident that the salesman model and the physician model have striking similarities. Even though the salesman model explicitly mentions 'Experience' as a strategy in which information is gathered about both products and customer interaction, the same principle holds true for the physician, who has spent time studying medicine and gained practical experience during internships. Aside from this dissimilarity, both models are evidently equivalent.

However, the librarian model would (at first glance) not seem to exactly correspond to the previous models. The reason behind this is the fact that the librarian model is not composed of consecutive steps, but rather a set of 'filters' as arbitrarily applied by librarians in the intermediation process. When examined closer, these filters are translatable to both physician and salesman model, yet not necessarily in a direct 1:1 relationship. The following table relates each of the strategies from the various models and offers the underlying description for all of the individual models, to form a series of generic strategies:

Physician	***Salesman***	***Librarian***	***Strategies***
	● Experience		Acquire Intermediation Information
● Observation and Interpretation ● Interview and Examination	● Initial Contact ● Requirements Acquisition	● Subject Definition ● Objective and Motivation ● Personal Background of Inquirer	Requirements Acquisition
● Devise Treatment	● Product Search	● Relationship of inquiry to file organization	Convert Request to Query
● Apply Treatment ● Treatment Monitoring	● Product Presentation ● Sale ● Evaluation	● Provide Acceptable Answers	Result Presentation

With respect to the field of (digital) Information Retrieval, these generic strategies can be described as follows:

- **Acquire Intermediation Information**: During this strategy, that precedes the actual intermediation process, the intermediary is trained to gain the information necessary to perform the required intermediation tasks in the following strategies.
- **Requirements Acquisition**: In order to acquire an accurate perspective of the searcher's information need, the searcher is enabled to express his request in his own terms (navigation-by-asking). The intermediary then attempts to help the searcher increase his information need awareness, by offering related terms and keywords that may also be relevant to his information need. This helps the searcher narrow down the exact nature of his information need.
- **Convert Request to Query**: Once the intermediary understands the searcher's information need, the intermediary is able to translate the searcher's request to the indexes of the documents. This translation is aimed at increasing the searcher's system awareness. Using this translation of the searcher's request, the intermediary is able to execute a query on the set of documents in the IS's repository.
- **Result Presentation**: The results of this query are then presented to the searcher. By using the searcher's feedback on given documents (navigation-by-proposition) the query can be further refined and recast to the IS's repository, in order to offer an increasingly relevant set of documents.

The primary difference between an intermediary and a traditional IS, lies in the intermediary's knowledge of searchers and their perception of documents. A traditional IS knows a great deal of the documents it offers, but is less proficient of linking the searcher's expression of his information need to (the indexes of) the documents. As a result, traditional IS's require searchers to express their need (i.e. their request for information) in the indexes that the IS uses to index its documents. These indexes need not correspond to the searcher's perception of those documents, causing problems with the interpretations of the searcher's information need. Intermediaries, however, use their knowledge of the human perspective on documents to allow searchers to express their need in their own terms, and (transparently) convert this to the system's indexes.

6. Consolidated Intermediation Model

"Quidquid latine dictum sit, altum sonatur."
- Origin unknown

From the practical intermediation models and the observations made from them, we can now derive a new model that consolidates the practical strategies, for the purpose of supporting (digital) exploratory search. Generally speaking, this involves gaining knowledge about both documents, the searcher's request and a way of fulfilling this request with the aforementioned documents. The following paragraphs in this chapter will each describe a part of the resulting consolidated intermediation model:

6.1. Acquire Intermediation Information

This part of the consolidated model focuses on acquiring the information necessary to offer intermediated support for the searcher.

6.1.1. Intermediation Information

Generally, there are two ways for a searcher to express his information need; in terms that a consumer would express his need and in terms that a producer would describe his product. For example, a searcher in need for a safe car, might express his need in terms of "safety" or "security", whereas an applicable product might be described using the related (but non-equivalent) terms "airbag" or "brake-assist". As such, the indices that apply to the documents in the IS's repository (internal indices) don't always match the indices that a searcher might use to express his need (external indices).

From the practical intermediation models, we can see that the intermediaries use several sources of information during the intermediation process. The first is derived throughout the intermediation process, whereas the others are obtained through training and/or experience:

1. **User Information**: The information that the intermediary derived from the searcher during the interview/negotiation process, which can be used to shape the intermediary's perception of the searcher's actual information need. The user information can be determined through requirements acquisition, which will be detailed in the next paragraph.
2. **Meta-Information**: The intermediary's knowledge of the intrinsic characteristics (internal indices) of the documents in the IS's repository. Knowledge of this type of information by the searcher, results in a higher system awareness (see *paragraph 4.3: Search Strategies*).
3. **Human Perception Information**: Using the intermediary's knowledge of searchers, and (specifically) the way they view the applicable documents, he is able to translate the searcher's need (external indices) to relevant documents' characteristics (internal indices).

Typically, search-engines require searchers to translate their information need to internal indices, even though this translation may be biased or incomplete. Even though in some cases (and trivial, non-ambiguous terminology) the internal indexes seem to match the specified keywords due to both parties sharing natural language, the translation still needs to be made. This implies that such search-engines assume a higher system awareness of the searcher, even though this awareness may be too low. An IS that uses human perception information to facilitate the translation from external to internal indices, would most likely yield results that are a better match to the searcher's expressed information need, because the external indices would better correspond to his vocabulary.

6.1.2. Human Computation

Meta-information can be acquired in the same ways as existing search-engines do, through document clustering. Human perception information, however, cannot be acquired in the same manner, as it implies a human perception on the (external) indices that are applicable to documents. For computers, this is obviously an open AI problem. A possible solution to this is offered as *distributed knowledge acquisition* [VONAHN2004]. Applying this concept onto information retrieval would call for groups of people to (indirectly) aid other searchers; people could aid searchers by applying indices which they perceive to be relevant to given documents. People could be convinced to index and categorize objects from a given domain through the use of *games with a purpose* [VONAHN2006]; whilst playing such a game, people will (implicitly) acquire meaningful data over specific objects. This data could then be used to relate the searcher's external indices to actual documents. Employing humans in this fashion, in order to solve open AI problems, is called *human computation* [VONAHN2006].

In such a *game-with-a-purpose*, users (in a multiplayer game) could be confronted with an document and be asked to guess what the other players might be thinking. When the players have nothing in common, except from the document in question, the players will ultimately attempt to provide terms or keywords that best describe the object. Using techniques such as taboo words (words that are disallowed in the game), it can be ensured that the set of keywords for a given document are as broad and as complete as possible. The more people that play the game (and thus provide input for a given document), the more accurate and error-free the set of keywords would be; a given incorrect keyword is less likely to be submitted that a correct keyword, meaning that incorrect input can be eliminated through statistical analysis. For more measures to assure the quality of the keywords, see [VONAHN2004]. The acquisition of external and internal indices, through human computation and document clustering respectively, can be illustrated as follows:

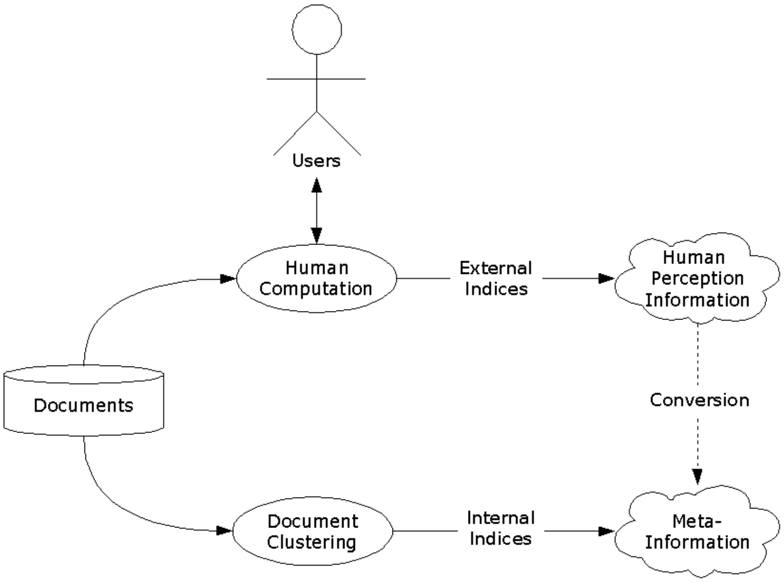


Illustration 9: Acquisition Of Intermediation Information

Human computation can thus be seen as a parallel process, that can run alongside the intermediation process. Although a certain minimum of information must have been acquired in order to perform intermediation in meaningful way, both human computation and intermediation can run independently of one and other.

6.1.3. Relationship-map

Once external indices have been acquired on a given document (through human computation), these indices need to be structured in some form of hierarchy. Such a hierarchy of indices should not separate the set of characteristics into a strict categorical directory listing, as it may well result in "*arbitrary hierarchical arrangements*" and would suffer from "*subjectivity of rating and annotating resou*rces" [WEBLIMINAL2006].

In order to perform the requirements acquisition (as described in the next paragraph), the IS will need to have an understanding of the various types of linguistic relationships that might exist between the external indices. A relationship-map should keep track of the various types of relations that might exist between the indices, such as holonyms / meronyms (*x part of y*) and hypernym / hyponym (*x is type of y*). Such a relationship-map could automatically be derived by applying the total set of indices to a lexical database, such as Princeton University's *WordNet* [WORDNET2007].

For example, in the human computation stage, users playing the *game-with-a-purpose* could be confronted with an "Airbus 360" (either in graphical form or textual description). The information acquired from the users would take the form of the external indices that most users believe to apply to the Airbus. These could include "wing", "airplane" or "vehicle". Using a lexical database, these (and other) indices could be structured as follows:

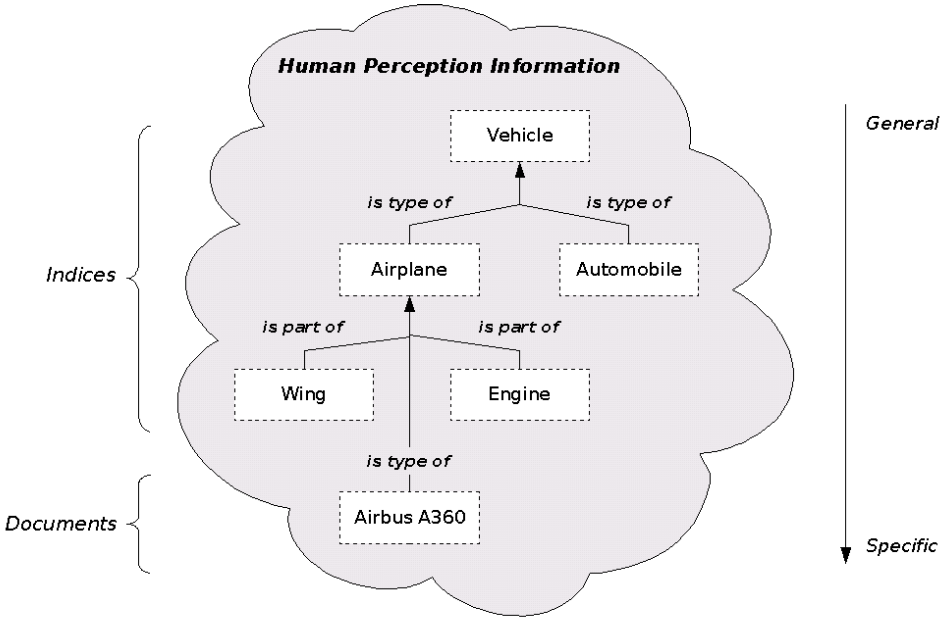


Illustration 10: Identifying Relationships In Human Perception Information

The structure derived from the lexical database can be used to help define the contribution (grain, focus, relation) of the external index to the actual object indexed.

6.2. Requirements Acquisition

The requirements acquisition stage serves to improve the searcher information need awareness, to the point where he is able to express his need (in external indices), as well as enabling the intermediary to create a perspective on the searcher's information need. From the practical intermediation models, we can derive that during the requirements acquisition stage the intermediary applies two separate strategies to improve his perspective of the searcher's information need. Note that both strategies revolves around the external indices, as the searcher may well be unaware of the applicable internal indices:

- **Passive Acquisition**: Initially, the intermediary enables the searcher to narrow down the area of his information need, by using the most discriminative indices as possible. By using the most discriminative indices, the intermediary needs fewer indices to be able to proceed to the following strategy. This matches Taylor's intuitive approach; "*We work from the general to the specific*" [TAYLOR1968].
- **Suggestive Acquisition**: Based upon the intermediary's perspective on the searcher's information need, additional suggestions for indices are offered that the intermediary assumes to be relevant. This perspective is based upon the indices that the searcher previously indicated to be relevant to his need. The suggested index will thus have some type of relationship with the original indices, that emerged from the relationship-map. The suggestions will serve to further narrow down the area of the searcher's information need by following the path through the relationship-map, from the initially relevant indices (general) to the more specific indices.

Passive acquisition can be accomplished by presenting the searcher with a set of (otherwise random) external indices. These indices would come from (or near) the top of the relationship-map, as those indices are most discriminative. The searcher would then be able to cast a *judgement* on those indices relevant to his information need. In case none of the indices from the set are relevant to the searcher, the set should automatically update to include other (equally random) indices.

The suggestions are based upon the intermediary's perspective on the searcher's information need, which in turn is based upon the searcher's judgements. Once it becomes clear that the suggestions receive positive judgements from the searcher, it can be assumed that the intermediary's perspective on the searcher's information need is correct. As discussed in [VANVLIET2007], the suggestions stop "*when the reaction of the searcher seems to be reasonably predictable, and suggestions for extensions are exhausted*". At which point, the interaction can proceed to the next strategy, as discussed in the following paragraph.

6.3. Convert Request To Query

To improve the searcher's system awareness, the intermediary's perception of the searcher's information need, expressed in (judgements on) external indices, needs to be translated into internal indices. This matches Taylor's description of a librarian, who *"becomes a translator, interpreting and restructuring the inquiry so it fits the files as they are organized in his library"* [TAYLOR1968]. In effect, this step interprets the searcher's request for information (expressed in judgements on proposed indices) into a query on the total set of documents.

This conversion process is accomplished by determining the correlation of the external indices to the internal indices on a given document, as defined in the human perception information and meta-information respectively. Schematically, this can be viewed as follows:

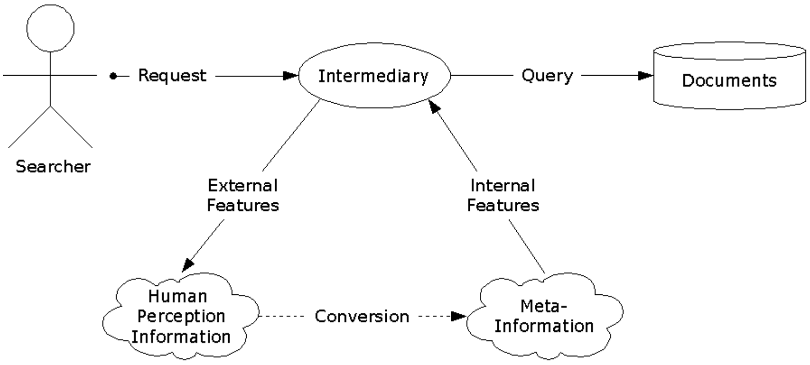


Illustration 11: Conversion Of Request To Query

6.4. Result Presentation

The result presentation stage serves to improve the searcher's system awareness, to the point where he is able to express his information need in internal indices, and consequently return the resulting documents that match those internal indices. Additionally, relevance feedback is employed to refine the query.

Once the query is executed and a result-set is obtained, the results can be presented to the searcher for evaluation and feedback. Since *"requirements acquisition and product presentation often cannot be completely separated"* [BERGMANN2002], we need to enable some type of relevance feedback to further refine the query to match the searcher's information need. In order to quantify the searcher's interest in a given document, one must attempt to ascertain the searcher's focus of interest.

White ([WHITE2003]) accomplishes this by separating a document into various stages of representation, in which each stage increases the level of detail. The searcher is then enabled to drill-down, through the path of representations and acquire more information about the document's content. One could, for instance, create a consecutive path of representations for a document, by using Ingwersen's *"modes of presentation of objects"* [INGWERSEN2000]:

- **Bibliographic Data**: The most abstract representation offers easily accessible information (author, title, keywords) to facilitate recognition of relevant topics and concepts.
- **Summary**: The summary offers a slightly more elaborate synopsis of the document. Note that it is even possible to automatically generate an (indicative) summary from a given document [MINYENKAN2003].
- **Object**: Once the path of representations has been completely traversed, the searcher is presented with the document itself.

Obviously, more and different steps of document representation can be devised. The searcher is ultimately in control as to what representations he wishes to see and if he wishes to drill further down its path. When the searcher chooses to abort the path, it can be assumed that the information in the current representation was not personally relevant. The distance travelled down the path indicates the searcher's commitment to that document, and the personal relevance thereof.

When the searcher chooses to drill-down, the current representation can be assumed to be personally relevant. This mean that the terms in that representation can be used to extend or refine the query, allowing for a greater precision of the query's results. In this way, the separation of a document into various representations enables implicit relevance feedback, without interrupting the searcher during his search process. Based on this feedback, the set of documents offered to the searcher can be changed, to more accurately reflect his information need.

Intuitively, the most recently viewed path of representations can be assumed to be more relevant than previously viewed paths, since the searcher's concept drift (a drift in the nature of his interest) or due to the fact that the searcher may well be experimenting with the search-engines interaction.

7. Application Of Consolidated Model

"The essence of knowledge is having it, and applying it." - Confucius

When creating an IS according to the consolidated model, the various intermediation strategies need to be expanded upon in greater detail. This chapter describes the various components of the consolidated model, along with the required techniques needed to implement such an IS in practice.

7.1. Quality Of Human Computation

The human computation component of the consolidated model is employed to acquire accurate and meaningful indices for each document. These indices then form the human perception information, that can be used to support a searcher in the expression of his information need. The degree to which this support is actually helpful, depends a great deal on the quality of the assigned indices. The following aspects are important when it comes to the quality of the indices within the human perspective information:

- **Quantity**: As more indices are assigned to documents, the certainty with which one such index applies to the corresponding document, increases. This is due to the fact that when many people assign the same index to a given document, there would seem to be an general consensus that this index is in fact applicable. As such, the more indices that are assigned, the more the human perspective information becomes reliable and meaningful.
- **Diversity**: The indices that have been assigned to a document will need to be as diverse as possible. Generally speaking, when people are asked to assign a term or concept to a presented document, their initial response will be usually be similar. Meaning that their response will typically be one of a very small set of indices. Yet, the diversity and exhaustiveness of the assigned indices are crucial to the IS's ability to use the human perception information to support the searcher with respect to his expressed information need.
- **Reliability**: The reliability of the assigned indices is crucial to effective intermediation, since a polluted set of indices would fail to relate relevant indices to the corresponding documents. Care must be taken to ensure the accuracy and reliability of assigned indices.

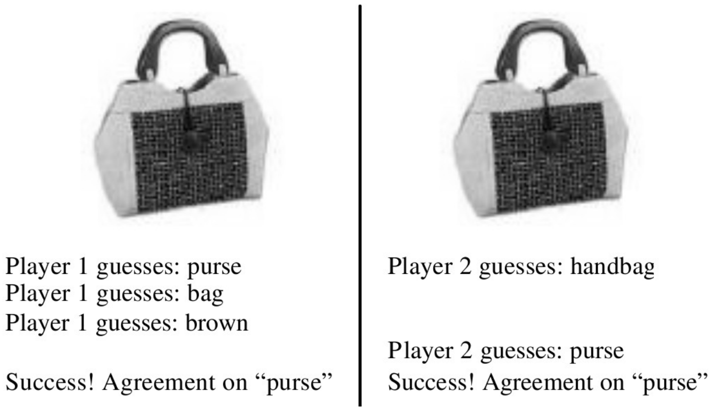
This chapter will discuss several techniques that can be employed to ensure the quality of the indices (as defined by the aspects above) that are acquired via human computation.

7.1.1. Game-With-A-Purpose

Human computation, as described by [VONAHN2006], attempts to solve the aforementioned quantity aspect by employing a so-called *game-with-a-purpose*. By making that game entertaining enough, it can be assumed that people will want to play the game voluntarily, rather than having to encourage them to do so. This degree of entertainment has proven to be enough to ensure that sufficient quantities of people will play such games. As an illustration of the success of this concept, Von Ahn's ESP Game [VONAHN2004] (used to provide labels for given images) acquired more than 10 million image labels from thousands of players within the first few weeks of operation: *"We don't expect volunteers to label all images on the Web for us: we expect all images to be labelled because people want to play our game."*

In order to make the game appealing, the game should support multiplayer modes, in which people co-operate in assigning indices. In the ESP Game, this aspect of the game was found to be most appealing to players. Two players could be presented with a (representation of) a certain document and asked for a index. In order not to confuse the searcher with terms like "indices" or "document", players could simply be asked to guess what the other person is *thinking*. In order to prevent co-operation of players outside of the game itself, it is important that players are anonymous within the game. Otherwise, players could use their own lines of communication to confer with one and other, in order to assign erroneous indices to documents. In the absence of any other means of communication with the other player, and without being able to see the other player's guesses, the only way to venture a reasonable guess is based upon the only thing that both players have in common: the presented document.

A single round continues until both players agree on a given index. An added benefit of this multiplayer scenario, is that such guesses are instantly validated; a round doesn't end until both players agree, meaning that a certain amount of reliability can be assigned to the resulting index. Consider the following illustration of Von Ahn's ESP Game:



Player 1 guesses: purse
Player 1 guesses: bag
Player 1 guesses: brown

Success! Agreement on "purse"

Player 2 guesses: handbag

Player 2 guesses: purse
Success! Agreement on "purse"

Illustration 12: Two Players Co-operate In Assigning Indices

This illustration represents a single round within such a game-with-a-purpose, which can similarly be applied to the consolidated model. Initially, both players guess something different, but at his second guess, the second player guessed the index "purse", which the first player had already proposed. In this case, an agreement was reached upon the index "purse", thereby ending that round of the game. In the game, both players could then be assigned certain points (for example, based upon the speed with which this agreement was made) and allowed to proceed to another round.

7.1.2. Taboo Words

In order to ensure the diversity of assigned indices, it is necessary to convince players to assign indices not previously assigned by others. Taking a reference from the word-guessing party-game *Taboo*, players could be presented with a list of words that the players cannot use to describe the document. This list of *taboo words* could be composed of words that have previously been assigned to the document in question. This would ensure that the player is forced to use other (less obvious) words to describe the document, therefore ensuring a broad diversity of indices. Furthermore, as a positive side-effect, this additional obstacle serves to make the game more challenging and (hopefully) more entertaining.

Ultimately, however, the set of taboo words would become too large, resulting in the fact that people are no longer able to conceive additional indices to describe the document. For such scenario's, people should be able to skip a given round. When several rounds of the same document have been skipped a certain number of times, it can be assumed that the taboo list has grown too large with respect to that particular document. At which point, the taboo list could be reset and allow players to start assigning all indices all over again, thereby repeating the cycle.

7.1.3. Test Cases

Human computation could possibly suffer from malicious intent, in the form of players assigning unrelated indices to documents. Within specific (online) communities, people might even be encouraged to do so. Even though there are no lines of communication between individual players, and they cannot agree on indices for individual documents, it may well be possible that people with malicious intent will call upon players of this game to assign one specific index to any and all documents. Obviously, this will pollute the data.

Therefore, in order to ensure the reliability of assigned indices, it can become necessary to 'test' individual players for their accuracy. In order to do so, a set of predetermined test cases can be formulated, in the form of a document and an exhaustive list of applicable indices. At irregular intervals, players of the game could be confronted with one of the test cases, rather than one of the documents from the repository. The index they then attempt to assign, will not be used to aid in intermediation for that document, but rather as a means of determining the accuracy of those particular players. If the player assigns an incorrect index, that player can be assumed to be inaccurate and unreliable. Any indices that that players would assign from then on, would then need to be ignored.

7.2. Conversion Of External To Internal Indices

Once the requirements acquisition stage is completed, and the searcher's request (in the form of a set of relevant external indices) has been acquired, this request must be translated into a query (in the form of internal indices) that can be executed on the set of documents within the IS. This paragraph deals with the correlation between external indices and their respective internal indices.

As described in the consolidated model, the internal indices are acquired via document clustering (for example, in the form of Latent Semantic Indexing) based upon the documents in the IS's repository. Conversely, the external indices are acquired through human computation, based upon those same documents. However, these steps fail to relate internal indices to their respective external indices (and vice versa). For example, a searcher might specify a need for a "car" and "environmentally friendly", in order to specify a need for a automobile with a low fuel consumption and/or low CO2 emissions. Such an information need may need to be translated to "Hybrid" or "Biodiesel", prior to executing such a query on the IS's repository. The correlation between external and internal indices can be established in the following ways:

7.2.1. Statistical Analysis Of The WWW

There is a vast amount of information available upon the World Wide Web (WWW). Although this information is inherently unstructured, the sheer quantity has made it an interesting research subject. Assuming that the information available upon the WWW is structured in a way that is representative of the target domain in question, then analysing the WWW would yield information that could then also apply to the domain of the IS. By analysing the WWW's corpus, there are various ways of establishing the strength of the relationship between two distinct concepts:

- **Googleshare**: Using the statistical results that Google.com offers, *Googleshare* (or Mindshare) [GOOGLESHARE2002] attempts to quantify the strength of the relationship between two concepts: "*Search for a term, then search within those results for another term. Divide the number of results for the second term by the number of results for the first term, and you have the "mindshare" of the second term in the domain of the first.*" The results of Google page count can thus be construed as the semantic mindshare that the second term has, with respect to the first. Consider the following exemplary Googleshares of terms related to the concept "Environmentally Friendly":

Domain		***Term Within Domain***		***Googleshare***
Term	***Results***	***Term***	***Results***	***Percent***
Environmentally Friendly	6,110,000	Hybrid	1,070,000	17,5 %
Environmentally Friendly	6,110,000	Hydrogen	949,000	15,5 %
Environmentally Friendly	6,110,000	Biodiesel	546,000	8,9 %

From this table, we can derive that the relationship between "hybrid" and "environmentally friendly" is strongest. When the query is to be formulated according to the Boolean model, the term "Hybrid" would therefore be the closest match to "environmentally friendly". Ideally, however, using the vector model it would be possible to construe a query in which all (or most) of the relevant terms are included, with respect to the strength of their individual relationships.

- **Normalised Google Distance**: Similar to the above concept of Googleshare, the Normalised Google Distance (NGD) [CILIBRASI2007] is another measure of semantic similarity between two terms that uses the probability of co-occurrences in documents within Google's repository. Although NGD was originally based on the Google search engine, this formula may be used in combination with other text corpora just as well.

  Applying the formula (as denoted in Cilibrasi's article) to all of the internal indices in the Meta-Information and the external indices in the Human Perception Information (see *paragraph 6.1: Acquire Intermediation Information*) would yield a multi-connection map of relationship strengths. Once this map has been established, it could be used during the intermediation stage to convert external into internal indices.

7.2.2. Human Computation

Within the consolidated model, human computation plays a role in acquiring the Human Perception Information; e.g. the external indices that represent how people view the documents within the consolidated model. As a second application of human computation within the consolidated model, a game-with-a-purpose could be employed in order to establish which external indices correspond to certain internal indices, and how strong this correlation is.

For example, players could be presented with a given external index, as well as several options for possibly related internal indices. By asking users to choose which of these options is most closely related to the given external index, statistical data emerges on the correlation of these indices. Although with human computation one always has to consider the entertaining value of such a game, the data collected from this game (after sufficient repetition) would yield the strength of the relationship between given external and internal indices. By applying the same strategy to all external and internal indices, a map of all connections (and their inherent strengths) would emerge, that can be used to translate a given external index (in the searcher's request) to the corresponding internal index (for the query on the IS's repository).

Furthermore, this simple example of the application of human computation, in order to solve a specific problem that computers cannot solve, illustrate how effective the concept of human computation is, in solving specific open-AI problems.

8. System Design

> *"In theory, there is no difference between theory and practice. But, in practice, there is." - Jan L. A. van de Snepscheut*

Data Flow Diagrams [GANE1979] show the flow of data from external entities into the system, and show how the data moves from one process to another, as well as to and from logical storages. The Data Flow Diagrams (DFD's) enable a structured analysis and design of each component of an IS built as an implementation of the consolidated model. Through DFD's, the following entities within a system can be identified:

- **External Entities**: These entities are the sources or destinations of data that exist outside of the system. These are represented as squares.
- **Processes**: The (data)processes perform some type of operation on incoming data flows and output the resulting data. Processes are represented as circles.
- **Datastores**: Datastores involve electronic (e.g. database, XML files) or physical (e.g. filing cabinet) forms of data storage, which may or may not be persistent. Datastores are represented as open-ended rectangles.
- **Data Flow**: The flow of data from one entity to another is represented as an arrow. The direction of the arrow indicates the direction of the flow of data.

The following sections offer an analysis of the various components of the IS, in terms of the above DFD-entities. Note that the DFD of the creation of the relationship-map has been omitted due to the trivial nature that this DFD would have. This is due to the fact that this element is more focussed on the internal data structure that the relationship-map would have, rather than the data flows and processes that it would involve. Therefore, the following components will be analysed:

- **Human Computation**: The human computation component is responsible for acquiring accurate and meaningful external indices on the documents that the IS offers the searchers. Human computation will take the form of a playable game, allowing the external indices to be acquired through interaction with the player(s). In order to make the game more appealing to the players, the game will be multiplayer and allow two players to cooperate in assigning keywords to a given document.
- **Requirements Acquisition**: This component serves to complete the acquisition of the searcher's requirements, that describe his information need. By presenting the searcher with external indices, the searcher is able to judge which indices are relevant to his need. Based upon these judgements, this component is then able to make predictions on related indices that might also be relevant to the searcher, and present these to the searcher. Based upon the searcher's judgement of those predictions, the prediction can be altered, or (if the prediction is accurate) the requirements acquisition can be halted and the system can proceed with the next stage of the consolidated model.

- **Convert Request To Query**: This stage of the consolidated model serves to translate the request (in the form of external indices acquired in the requirements acquisition component) to a query (in the form of internal indices denoted within the Meta-Information). Using additional data concerning the strength of relationships between external and internal data (see *paragraph 7.2: Conversion Of External To Internal Indices*), this translation can be accomplished. This data will be known as the *Semantic Conversion Map.*
- **Result Presentation**: Result Presentation is the component of the consolidated model that is responsible for presenting the results to the searcher. If the searcher is somehow interested in a particular result (i.e. based upon its title), he can request a higher level of the document's representation (see *paragraph 6.4: Result Presentation)*. The document is then parsed according to this level of representation (either showing only the bibliographic data, an indicative summary or the entire document itself) and presented to the searcher. Requesting an additional representation indicates an interest by the searcher in that document, therefore implying that it is (most likely to be) relevant. This request can then be used to update the relevant indices (stored within the User Information) such that it can be used in the "Convert Request To Query" component to recast a query.

The following paragraphs will each offer the Data Flow Diagram and Data Dictionary for their respective component:

8.1. Human Computation

8.1.1. Data Flow Diagram

The following schematic offers the DFD of the human computation component, with the two players (Player 1 and Player 2) as external entities. These players will be external entities within the resulting diagram and will offer the indices to fill the document information. The Human Perception Information (which stores the (external) indices that have been previously assigned) and the Documents database are included as datastores. Furthermore, two additional datastores are added: Test Cases (as part of a quality control measure to ensure the reliability of the players) and the Game Data datastore (the storage of all of the information gathered during an instance of the game).

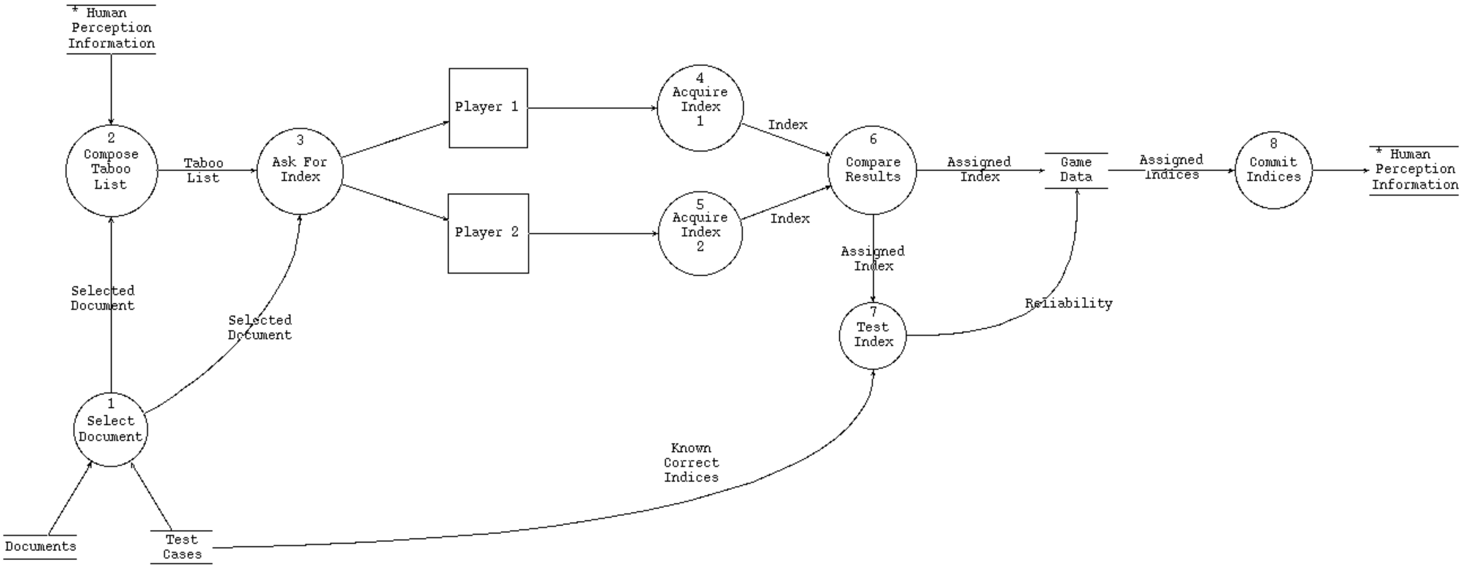


Illustration 13: Data Flow Diagram For The Human Computation Component

8.1.2. Data Dictionary

From this DFD, the following (data)processes can be identified:

- **Select Document**: For the purpose of the game, a single document must be selected from the total set of documents, that can be presented to both players. As part of the quality assurance of the results of human computation (See *paragraph 7.1.3: Test Cases*), a test case can be selected, rather than an existing document.
- **Compose Taboo Words**: In order to guarantee the diversity and completeness of the acquired indices, previously assigned indices of the given document are retrieved from the document information and used as taboo words (See *paragraph 7.1.2: Taboo Words*).
- **Ask For Index**: At this point, the document and the taboo words are presented to both players and they are asked to guess what the other player is thinking, based on the presented document, without making use of the taboo words.
- **Acquire Index**: The indices are acquired from each respective player.
- **Compare Results**: By comparing the two indices, the results of that particular round can be determined. If the indices are similar, the result will serve to increase both player's scores. Furthermore, the similarity implies that the resulting index is reliable and somehow applicable to the document. If the indices as dissimilar, the round continues and additional indices will be asked and required from the players.
- **Test Index**: If the document in question was in fact a test case, the result of the comparison can be used to determine the reliability of the players. If the result is similar to one of the indices that are known to be correct, the reliability of the other indices (that were assigned to non-test cases) is increased.
- **Commit Indices**: Once the game is over, the acquired indices and the reliability of the players that assigned them, are committed to the Human Perception Information datastore. Based upon this reliability, the applicability of a given index to the corresponding document can be ascertained.

8.2. Requirements Acquisition

8.2.1. Data Flow Diagram

The following schematic offers the DFD of the requirements acquisition component, in which the searcher (whose requirements are to be acquired) has been included as an external entity and the Relationship-Map (which includes the hierarchy of relationships between external indices) and User Information (storing all of the indices that the searcher judged to be relevant) have been included as datastores.

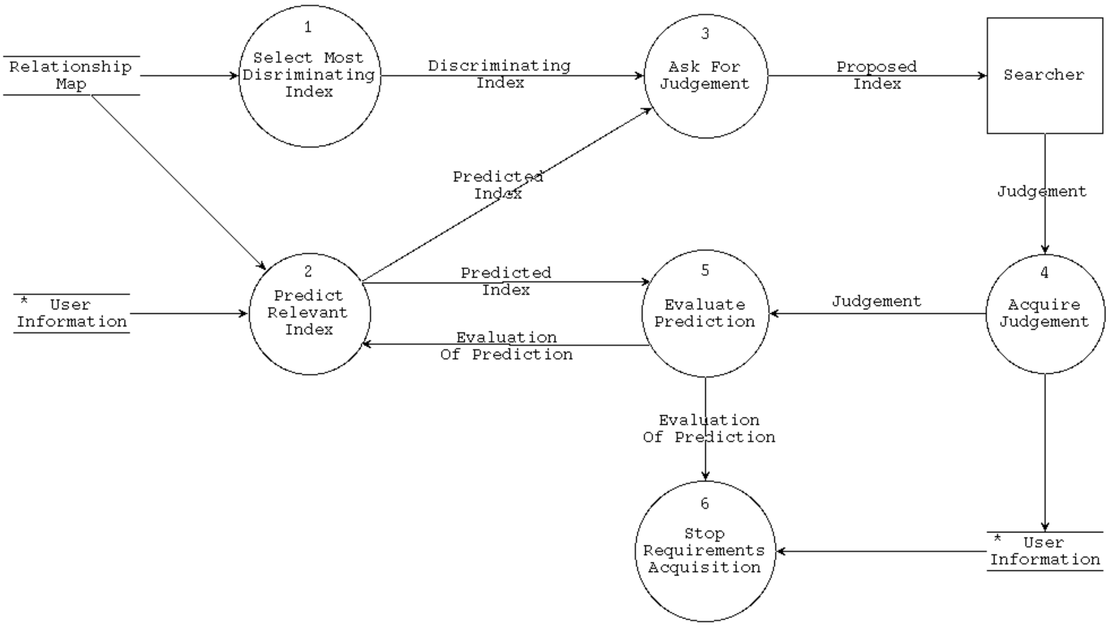


Illustration 14: Data Flow Diagram For The Requirements Acquisition Component

8.2.2. Data Dictionary

From this DFD, the following (data)processes can be identified:

- **Select Most Discriminative Index**: This data process is focussed on analysing the relationship-map in order to find the most discriminative indices within it. These are the indices that are at the top of the relationship-map and are the most generalized concepts within its hierarchy.
- **Predict Relevant Index**: Aside from selecting the most discriminative index, the index that is to be proposed to the searcher, can also be based upon a prediction of (presumed) relevant indices. This prediction is based upon the User Information (containing indices that the searcher has judged to be relevant) and the relationship-map. The latter contains indices that are related to the relevant indices (of the User Information datastore), which may thus also prove relevant to the searcher. Furthermore, this data process can use the evaluation of previous predictions to further refine its consecutive predictions.
- **Ask For Judgements**: The most discriminative index is then proposed to the searcher, in order to allow him to judge the relevancy of this index (as described in *paragraph 6.2: Requirements Acquisition*).
- **Acquire Judgement**: The resulting judgement (whether or not the index is deemed relevant) is then acquired from the external entity "Searcher".
- **Evaluate Prediction**: By analysing the predicted index, based upon the searcher's judgement, an evaluation can be made of the value of that prediction. This can then, in turn, be used by the "Predict Relevant Index" data process to further refine the next prediction.
- **Stop Requirements Acquisition**: Once the searcher's information need is sufficiently predictable (as described in *paragraph 6.2: Requirements Acquisition*) the requirements acquisition component can stop; leaving the relevant indices in the User Information datastore, which is to be used by the "Convert Request To Query" component.

8.3. Convert Request To Query

8.3.1. Data Flow Diagram

The following schematic offers the DFD of this stage of the consolidated model, in which the Documents (which host the documents within the IS's repository) and User Information (which stores the acquired relevant external indices) have been added as datastores. Furthermore, three additional datastores have been added: Semantic Conversion Map (the set of relationships between external and internal indices, as a result of *paragraph 7.2: Conversion Of External To Internal Indices*), Query (which stores all of the internal indices while the other external indices from User Information are yet to be translated) and Result-set (which stores all of the results of the executed query).

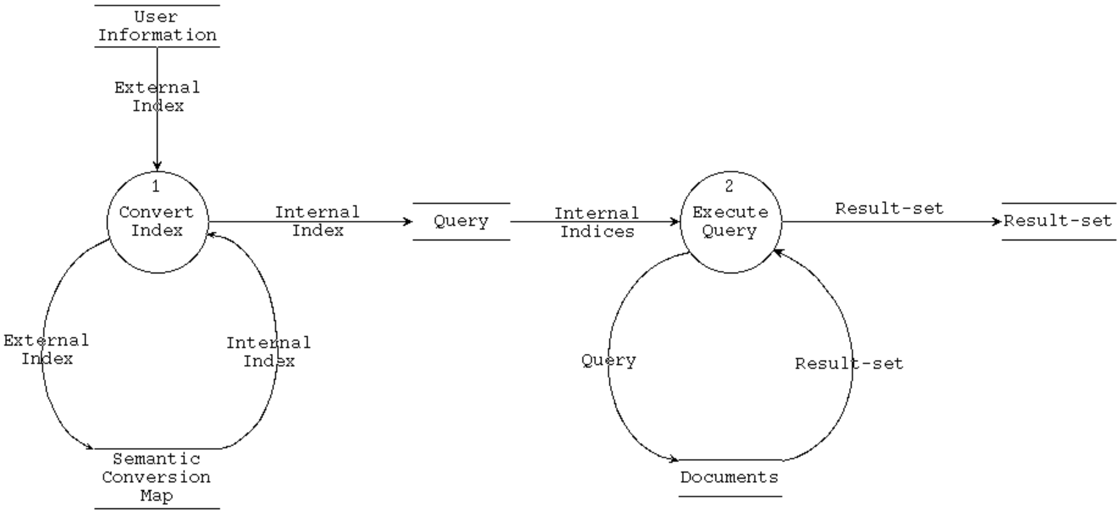


Illustration 15: Data Flow Diagram For Convert Request To Query

8.3.2. Data Dictionary

From this DFD, the following (data)processes can be identified:

- **Convert Index**: This data process focusses on translating a given external index (from the User Information datastore) that represents one of the indices that the searcher deemed as relevant, to an internal index. This translation is accomplished by requesting the internal counterpart(s) (one or more: this depends on the type of translation) for every external index. The result is then stored in the (temporary) Query datastore.
- **Execute Query**: Once all external indices are translated, and the Query datastore contains the internal indices need to execute the query, the query is parsed (using the internal indices) and executed on the set of documents. The resulting set of (presumed) relevant documents is then stored in the datastore Result-set, to be used by the "Result Presentation" component.

8.4. Result Presentation

8.4.1. Data Flow Diagram

The following schematic offers the DFD of the result presentation component, in which the Searcher is included as an external entity and the Documents datastore (which host the documents within the IS's repository) and the User Information (which stores the indices the searcher judged to be relevant) have been included. Furthermore, the Result-set datastore (which contains all of the documents that resulted from the query are stored) has been used, that was created in the Convert Request To Query stage.

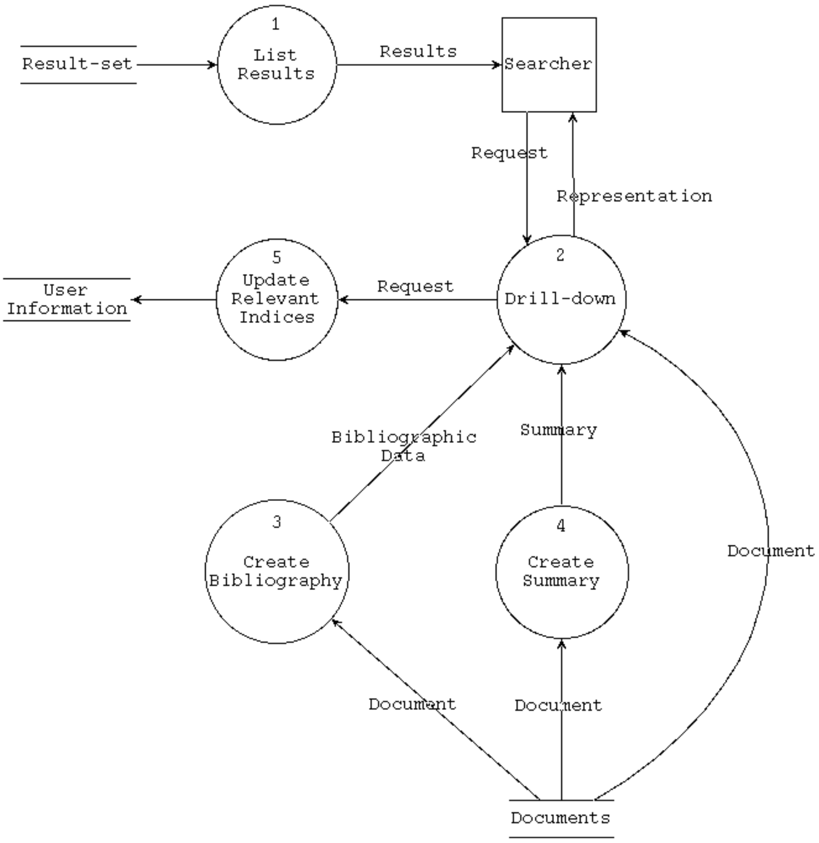


Illustration 16: Data Flow Diagram For The Result Presentation Component

8.4.2. Data Dictionary

From this DFD, the following (data)processes can be identified:

- **List Results**: Using the results listed in the Result-set datastore, the data process parses and presents the data to the searcher.
- **Drill-down**: If the Searcher wishes to see more of a given document (of the set of results), he can request to increase the level of representation. The result will be a new Representation will then be presented to the searcher.
- **Create Bibliography**: If the requested representation is, in fact, of the bibliographic level, this process will parse the document in question and return its bibliographic data to the "Drill-down" process.
- **Create Summary**: Conversely, if the requested representation is of the summary level, this process will parse the document in question (for example, using a means to automatically generate (indicative) summaries from a given document [MINYENKAN2003]) and return this summary to the "Drill-down" process.
- **Update Relevant Indices**: A request of a searcher for a more elaborate representation of a given document indicates a degree of interest in, and (presumably) relevance of, that document. This data process is able to use this request in order to update the set of relevant indices within the User Information, such that the query can be recast at a later time (as described in the "Convert Request To Query" component).

9. Conclusion

"Reasoning draws a conclusion, but does not make the conclusion certain, unless the mind discovers it by the path of experience." - Roger Bacon

By correlating the various concepts surrounding the searcher's problem space (e.g. searcher knowledge, Information Systems, outside knowledge), we were able to create a model in which the exact nature of their underlying relationships became apparent. Furthermore, the overlapping regions resulted in a comprehensive delineation of the various categories of information and knowledge, with respect to the searcher's knowledge. Using this model, we were able to more accurately express existing concepts (e.g. personal and semantic relevance) and the differences between them.

Taylor's definition of the various information need levels was not completely outlined and thus subject to interpretation. By expressing these need levels in terms of Belkin's information seeking dimensions, this definition was made more concrete. In doing so, we were able to connect various models, each describing a set of search strategies, to Taylor's information need levels. First of all, this enabled us to see how a certain state of information need awareness resulted in the corresponding search strategy, as employed during the search process. Second of all, this allowed us to see that each of the models described a similar set of search strategies, using different terminologies. Using this insight, a unified set of search strategies was proposed.

The addition of the system awareness dimension (aside from the information need awareness), expanded the set of dimensions in which Taylor's need levels could be described. Using this additional dimension, we were able to differentiate the compromised need level from the preceding need levels (i.e. visceral, conscious, formalized). This resulted us to conclude that the compromised level was in fact not a separate need level, but rather an expansion upon the formalized need level with additional system awareness. Furthermore, the system awareness dimension resulted in a Awareness Development Path (ADP) that describes searchers' progression between the two forms of awareness.

The unification of existing (overlapping) studies into the actions and operations performed by searchers resulted in a single search flow chart. This schematic representation of the possible sequence of search operations describes the many ways in which people are able to acquire, consume and filter information.

By acquiring inspiration from the practical (human-to-human) intermediation models and by translating our observations derived from those practical models, we have composed a consolidated intermediation model to enable digital IR intermediation. The introduction of human computation in order to acquire human perception information allows an IS (built according to the consolidated intermediation model) to interact with searchers and perform requirements acquisition in terms derived from the human perception on documents. This addition of external indices to supplement the internal indices allows for a more true to life interpretation of the searcher's information need. Combined with the relationship-map, which allows an IS to propose related additional keywords much like a human intermediary would, we believe the consolidated intermediation model to be more proficient (than existing search engines) at guiding searchers during exploratory search.

Although, in order for the external indices derived from human computation to be most effective at assisting searchers in conveying their information need, one first needs to obtain a sufficient degree of information (in the form of external indices) through human computation. Depending on the size of the repository and the willingness of people to participate in this type of implicit data-acquisition, this process will take some time. However, we feel that the added value that the external indices offers this type of intermediation, ultimately outweighs the (initial) effort it requires.

10. Recommendations

> *"Why give advice?... A wise man won't need it.*
> *A fool won't heed it." - Origin unknown*

This chapter details several recommendations for future studies and for further expansion upon this research. Most notably, it details those aspects of this research that were impossible to accomplish due to time constraints.

10.1. Analysis of Data Structures

An analysis has been made of the various data flows and data processes (see *chapter 8: System Design*). However, due to the fact that the description of the data structures has yet to be defined, this does not yet offer sufficient information for implementing the design. Prior to making an implementation or proof-of-concept, it would probably be best to analyse the data that is present in the various components, in order to gain insight into the best means of structuring the data.

This is of particular interest in those elements that appeared as datastores within the aforementioned DFD's. For example, for the relationship-map, meta-information and human perception information, the means of structuring the available data would dictate how that data would be stored and accessed from within that data structure.

10.2. Proof-Of-Concept

Although time did not permit to create a working proof-of-concept implementation of the consolidated model, some of the various steps in this model have been derived from existing research which offer their own implementation. Therefore, the results of this existing research (and the conclusions that it offered from experimental research) can still be assumed to be applicable when considering the feasibility of an implementation of the consolidated model. In this research, the following applicable experimental studies were considered:

- **Human Computation**: Von Ahn's ESP Game [VONAHN2004], is an experimental system that offers people to assign indexes to a total of 350.000 images. When all of the required quality measures are applied (see *paragraph 7.1: Quality Of Human Computation*) the quality of the resulting image labels is high, and accurately represents the people's perspective on those images and the current zeitgeist that the images would be subject to. In his paper describing the results of four months of experimentation, Von Ahn observes that "*A total of 13,630 people played the game during this time, generating 1,271,451 labels for 293,760 different images.*" Since, in the case of the consolidated model, the quality of the intermediation can be assumed to improve when there are many assigned indices, this suggests that human computation is an effective method to acquire these external indices.
- **Result Presentation**: Result Presentation: The separation of individual documents into various forms of representations (as discussed in *paragraph 6.4: Result Presentation*) has already been attempted in [WHITE2003]. By using a path of different forms of document representations and by seeing how far individual searchers traverse down this path, it becomes possible to measure the user's interest in a given document. By doing so, the terms in that representation (that the searcher wished to know more about) can be assumed to be relevant and (the terms therein) can be added to the query.

> White's study illustrated that inexperienced users found this assistance in their query formulation helpful, though even experienced users appreciated the adaptive system better than the baseline system (which did not automatically adapt at all). The result of this study shows that this type of assistance was deemed helpful by end-users: "*It seems the actions of the system adequately reflected the degree of change in a user's information need.*"

From this set of experiments, we can conclude that various parts of the consolidated model can be considered practically feasible. However, since several others aspects have yet to be proven, let alone in unison with the other components within the consolidated model, it would be advisable to create an experimental system that integrates the various components into a working environment. At that point, aside from the technical feasibility, it would also become possible to evaluate the system's added value by exposing it to various groups of users, in order to ascertain the added benefit this system offers during the exploratory search process.

10.3. Applications of Human Computation

In the consolidated model, human computation is employed to gather external indices that apply to the documents within the IS's repository. The idea being, that the IS is able to use these external indices to better understand the searcher's information need and assist in refining it further. However, other applications of human computation can be devised that may well be able to improve our model even further.

For example, users of a specific game-with-a-purpose could be presented with the expressed information need that former searchers have experienced. By, in some form or another, requiring users of this game to 'solve' the proposed information need, users could be encouraged to gather documents/webpages that they deem to be relevant to the given information need. By analysing the resulting documents, this game could offer invaluable information about the relationship of the expressed information need to the documents within the corpus of the World Wide Web. This relationship could then be used to further refine similar information needs, which other searchers might experience in the future.

Bibliography

AGMRC2007: "*Relationship Selling: The Path to Sales Success*". 2007. http://*www.agmrc.org/agmrc/business/operatingbusiness*

BAEZA-YATES1999: Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "*Modern Information Retrieval*". , 1999.

BAKKER2007: Jeroen Bakker, "*A user-driven adaptive interaction strategy to support exploratory search*". Radboud University Nijmegen, 2007.

BELKIN1993: Nicholas Belkin, "*Interaction with texts: information retrieval as information-seeking...*". Rutgers University, 1993.

BELKIN1994: Nicholas Belkin, "*Cases, Scripts, and Information-Seeking Strategies*". Rutgers University, 1994.

BELKIN2005: N. J. Belkin, "*Anomalous State of Knowledge*". , 2005.

BERGMANN2002: Ralph Bergmann, "*Acquiring Customers' Requirements in Electronic Commerce*". University of Hildesheim, 2002.

BRUCE2005: Harry Bruce, "*Personal, anticipated information need*". University of Washington, 2005.

CILIBRASI2007: Ŕudi Cili brasi, Paul Vitanyi , "*The Google Similarity Distance*". , 2007.

GANE1979: Christopher Gane, "*Structured systems analysis: tools and techniques*". , 1979.

GOOGLESHARE2002: "*Googleshare*". 2002. http://*atomiq.org/archives/2002/11/googleshare.html*

INGWERSEN2000: "*Users in context*". 2000. http://*ei.cs.vt.edu/~cs5604/f01/lectures/ingwersen.pdf*

KUHLTHAU1993: Carol Kuhlthau, "*Seeking meaning*". Rutgers University, 1993.

LARSEN1997: Jan-Helge Larsen, "*P-R-A-C-T-I-C-A-L, a step-by-step consultation model*". Department of General Practice, 1997.

MARCHIONINI1995: Marchionini, "*Information seeking in electronic environments*". Cambridge University, 1995.

MINYENKAN2003: Min-Yen Kan, "*Automatic Text Summarization As Applied To Information Retrieval*". Columbia University, 2003.

NIEDZWIEDZKA2003: Barbara Niedźwiedzka, "*A proposed general model of information behaviour*". Institute of Public Health, 2003.

SHIMAZU2002: Hideo Shimazu, "*ExpertClerk: A Conversational Case-Based Reasoning Tool...*". Internet Systems Research Laboratories, 2002.

SONNENWALD2001: Diane Sonnenwald, "*Evolving perspectives of human information behavior: contexts, situation...*". University of North Carolina, 2001.

STOJANOVIC2005: Nenad Stojanovic, "*On the role of a user's knowledge gap in an information retrieval process*". University of Karlsruhe, 2005.

TAYLOR1968: Robert Taylor, "*Question-negotiation and information seeking in libraries*". Lehigh University, 1968.

VANVLIET2007: Mario van Vliet, Theo van der Weide, "*The sales dialog as a model for information retrieval*". Radboud University Nijmegen, 2007.

VONAHN2004: Luis von Ahn, "*Labeling images with a computer game*". Carnegie Mellon University, 2004.

VONAHN2006: Luis von Ahn, "*Games with a purpose*". Carnegie Mellon University, 2006.

WEBLIMINAL2006: "*Directories and Virtual Libraries*". 2006. http://*www.webliminal.com/search/search-web04.html*

WEICHOO2000: Chun Wei Choo, et al, "*Information seeking on the web: a model of browsing and searching*". First Monday, 2000.

WHITE2003: Ryen White, Joemon Jose, Ian Ruthven, "*Adapting To Evolving Needs: Evaluating A Behaviour-Based Search Interface*". University of Glasgow, 2003.

WHITE2005: Ryen White, et al., "*Exploratory search interfaces: categorization, clustering and beyond*". University of Maryland, 2005.

WILSON1997: Tom Wilson, "*Information Behaviour: An interdisciplanary Perspective*". University of Sheffield, 1997.

WORDNET2007: "*WordNet, a lexical database for the English language*". 2007. http:/*/wordnet.princeton.edu/*