What is Information Discovery About?

H.A. Proper
ID Research
Groningenweg 6
2803 PV Gouda
The Netherlands
E.Proper@acm.org

P.D. Bruza[1]
Faculty of Information Technology
Queensland University of Technology
GPO Box 2434
Brisbane 4001, Australia
Bruza@icis.qut.edu.au

Version of June 23, 2004 at 10:32

If you know
what you are looking for
why are you looking
and if you do not know
what you are looking for
how can you find it?

Old Russian proverb

PUBLISHED AS:

H.A. Proper and P.D. Bruza. What is Information Discovery About? *Journal of the American Society for Information Science*, 50(9):737–750, July 1999.

Abstract

The Internet has led to an increase in the quantity and diversity of information available for searching. Furthermore, users are bombarded by a constant barrage of electronic messages in the form of e-mail, faxes, etc. This has led to a plethora of search engines, "intelligent" agents, etc. that aim to help users in their quest for relevant information, or shield them against irrelevant information. All these systems aim to identify the potentially relevant information in amongst a large pool of available information.

No unifying underlying theory for information discovery systems exists as yet. The aim of this article is to provide a logic-based framework for information discovery, and relate this to the traditional field of information retrieval. Furthermore, the often ignored user receives special emphasis. In information discovery, a good understanding of a user's (sometimes hidden) needs and beliefs is essential.

We will develop a logic-based approach to express the mechanics of information discovery, while the pragmatics are based on an analysis of the underlying informational semantics of information carriers and information needs of users.

KEYWORDS: Logic-based information retrieval, information discovery, information retrieval.

1 Introduction

With the increased use of the Internet (the net) comes an increase in quantity and diversity of information carriers offered on the net. Most visible is the increased use of the World Wide Web. Information carriers

[1] Also: Research Discovery Unit, Research Data Network Cooperative Research Centre, Level 7, Gehrmann Laboratories, The University of Queensland, Brisbane, 4072 Australia

accessible through the net include web pages, newsgroups, mailing-list archives, networked databases, applications, business services, as well as indexing services. For users of the net these carriers are at their disposal for doing business, search other information, educational purposes, or relaxation. The net can therefore be seen as a large market place where information demand meets information supply. Since the net literally spans the world, the number of accessible information carriers is astronomical. This makes life rather difficult for the average user who shops around *to discover* information carriers that fulfill his or her given information need. Existing Internet search tools return many information carriers. Users are still required to manually wade through large result sets in search of relevant information carriers.

On top of this, most users are bombarded by a (mostly unsolicited) stream of messages in the form of e-mail, notifications of new WWW pages, news-feeds, faxes, and phone messages. This constant, and still increasing, bombardment of information has led to a feeling of information overload. Users need mechanisms *to shield* themselves from irrelevant information.

On the eve of what is sometimes called the information-age already two serious long-term problems can be identified: discovering the relevant information in a huge ocean of information, and simultaneously shielding ourselves from irrelevant information coming at us. No unifying underlying theory for information discovery systems exists as yet. The aim of this article is to provide a framework of understanding for information discovery, and relate this to the traditional field of information retrieval.

1.1 Information Discovery

The problem of discovering information carriers on the net is related to the classical field of information retrieval [Rij79]. However, there are a number of clear differences as well. The *information retrieval* field has traditionally focussed on searching relevant documents in fixed document collections; usually textual documents. Users are presumed to have a very clear understanding of their information need. Although it is acknowledged in e.g. the Cranfield tests [Cle91] that users have difficulty in expressing this need in a formal language, the fact that searching for information is more of an interactive process of learning and discovery is not taken into account. This latter limitation of the information retrieval field is most apparent in the way systems are evaluated. The effectiveness of an information retrieval system is measured in terms of *precision* and *recall*[1] for a fixed set of queries on a standardised document collection.

Information retrieval can clearly been distinguished from *information discovery*. For example, *information discovery* is performed in an open networked environment. As a consequence, the document collection is not fixed. Moreover, the documents, or rather information carriers, are not necessarily textual but may be of a heterogeneous or aggregated nature. Aggregation makes the problem of discovering the right information carriers to fulfill a user's needs even harder. We agree with Lynch [Lyn95] that information discovery is

> a complex collection of activities that can range from simply locating a well-specified digital object on the network through lengthy iterative research activities which involve the identification of a set of potentially relevant networked information resources, the organization and ranking resources in this candidate set, and the repeated expansion or restriction of this set based on characteristics of the identified resources and exploration of specific resources.

There has been much recent work on web-based information discovery, for example, Chen *et al.*[CHSS98] and Desai [Des97] are recent expositions in this area. Information discovery is sometimes equated with the term *resource discovery*. The latter term is prevelant in digital library circles. We will adhere to the term *information discovery* in this article.

This brings us to the *information discovery paradigm*. Figure 1 portrays the essential aspects of the information discovery problem. On one side (the right hand side), there are information carriers as provided by the collections of information carriers that are at our disposal. These information carriers, which may

[1] Recall is the ratio of relevant retrieved objects to retrieved objects, whereas precision is the ratio of relevant retrieved objects to retrieved objects. These effectiveness criteria are generally applied in a controlled experimental environment.

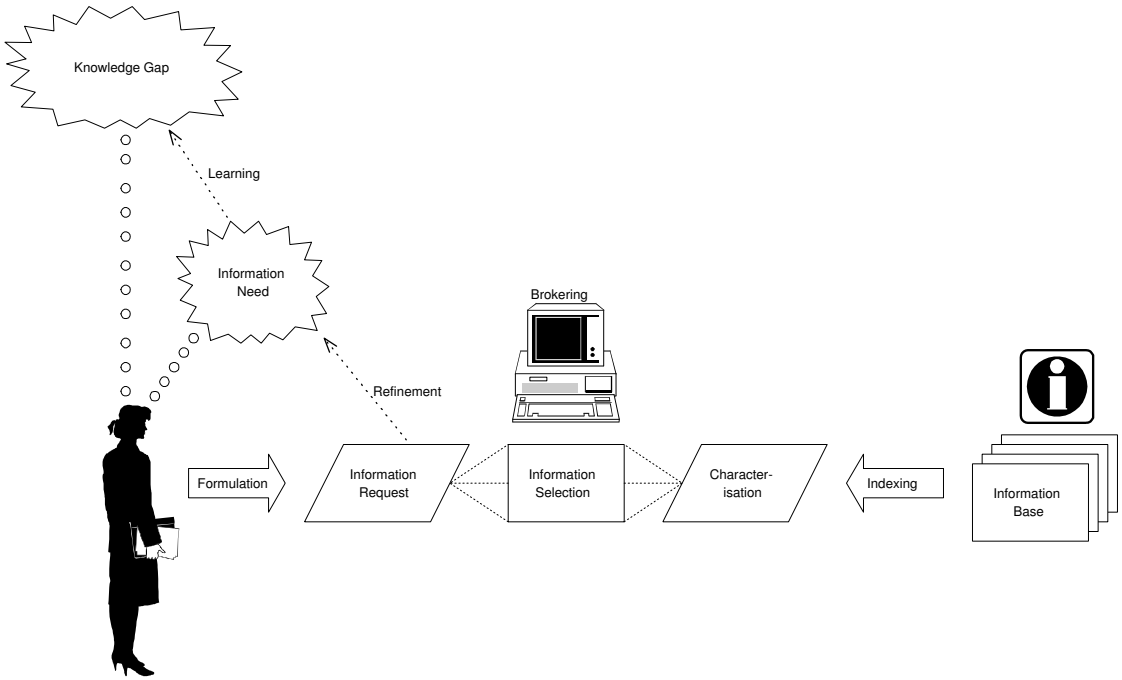

Figure 1: Information discovery paradigm

be aggregated, are characterised in some way to facilitate their discovery. Note that even though we shall use the term *information carrier*, the carriers actually only carry *data*. The data carried does not become information until a user interprets the data. Nevertheless we will adhere to the term *information carrier*.

Facing the information carriers is the user with an information need. The user expresses this need in terms of an information request; a query. The query will usually only be a crude description of the actual carrier(s) needed to fulfill the given information need. Therefore, it is also useful to allow further refinements of this need as the search proceeds. This refinement process is usually referred to as *relevance feedback*.

The need for information can be caused by a number of reasons. The focus in this paper is when the information need arises from a gap in the user's knowledge. For example, the user needs to know something in order to complete a task. Relevant information is discovered and then absorbed by the user to fill the knowledge gap. A knowledge gap may also arise out of idle curiosity. For example, some users of the Internet begin surfing the internet with no specific goal and then encounter some topic that engages their curiosity in the sense they want to learn more about it.

The knowledge gap can range from being fairly specific such as learning *the latest price of 19 micron wool*, to the very broad, such as learning about the *theory of relativity*. A specific need can usually be satisfied by a small collection of facts, while a broad need usually requires a wider variety of facts. Observe that during the search process users may learn more and more about their knowledge gap, and may thus discover aspects of this gap they were initially not aware of. This means that the actual information need of a user may evolve as they are exposed to new information.

Given a query, a selection of information carriers that are considered relevant can be made. This selection mechanism can be compared to an automatic brokering service, matching demand to supply. Initially, only a limited number of the selected carriers can be shown to the user to obtain *relevance feedback* from the user to further refine the query.

The information discovery problem boils down to finding the right information carriers that will fill the user's given knowledge gap. Three issues play a central role in the information discovery problem:

1. formulation of information requests

2. characterisation of information carriers
3. selection of information carriers

The formulation of information requests involves two important issues. First of all, it requires some formal language in which to express the query. Secondly, a precise formulation of the *true* information need is required. Obtaining such a formulation has proven to be a non trivial task [Cle91].

Good characterisation of information carriers is imperative for effective information discovery, as poor characterisations inevitably lead to the retrieval of irrelevant information, or the missing of relevant information. An important question is of course which properties to include in a characterisation. A useful property to include seems to be what an information carrier is *about*. In addition, properties like authorship, price, medium, etc. may be included. In the literature standard attribute sets to characterise information carriers can be found in the context of metadata standardisation efforts [BL94, SM94, WGMD95].

The selection of relevant information carriers for a given query q is a well understood problem. For finding unstructured information carriers, the field of information retrieval has developed a number of retrieval models. However, this field is still very much at the stage of simply returning information carriers which the user must then peruse in order to glean the information that fills their knowledge gap.

A next step would be to support *knowledge discovery*. In knowledge discovery one would try to derive the exact fragments of knowledge the user is after from the relevant information. So users would not have to read entire documents, but the system would give an exact and concise answer to the user answering their queries. This would require the system to sometimes interpret the information found and autonomously infer new information.

1.2 Information routing

Besides actively searching for information, users and organisations are confronted with a constant stream of electronic messages. These messages range from simple notifications, via e-mail messages and notifications of new WWW pages, to voice mail. For this, a more passive form of information discovery is required. Incoming messages need to be filtered in order to partition the potentially relevant messages from the irrelevant ones.

Conceptually, messages can be seen as a pointer to a freshly created information carrier (the actual body of the message). This view concurs with the view that modern software for messaging seems to take with, for example, a *universal inbox* for all incoming messages be it e-mail, fax, or voice messages. These messages need to be routed to the appropriate message-box(es) of the right person(s), and should then be prioritised within the message-boxes. This means that information filtering, discovering relevant messages in the incoming stream, involves two activities: routing and ranking.

In figure 2, it is illustrated how we view this process of routing and ranking. Each incoming message passes through a layer of routing modules that select the appropriate message-box(es), which could be from a multitude of users. Each message-box has an associated ranking module that ranks the messages currently in the message-box using user specified criteria.

In the remainder we shall use the term *information discovery* for the process of actively searching information, as discussed in subsection 1.1, and the term *information filtering* for passively discovering relevant information in an incoming message stream. The theory that will be developed in this article is focussed on a reasoning mechanism for relevance of information carriers. This theory will then be applicable to the selection process of information discovery, as well as the routing and ranking of messages for information filtering.

1.3 Structure of the paper

In section 2 the philosophical preliminaries (way of thinking) are discussed, and special attention is paid to the user in the discovery process. A generic reasoning mechanism for the relevance of information

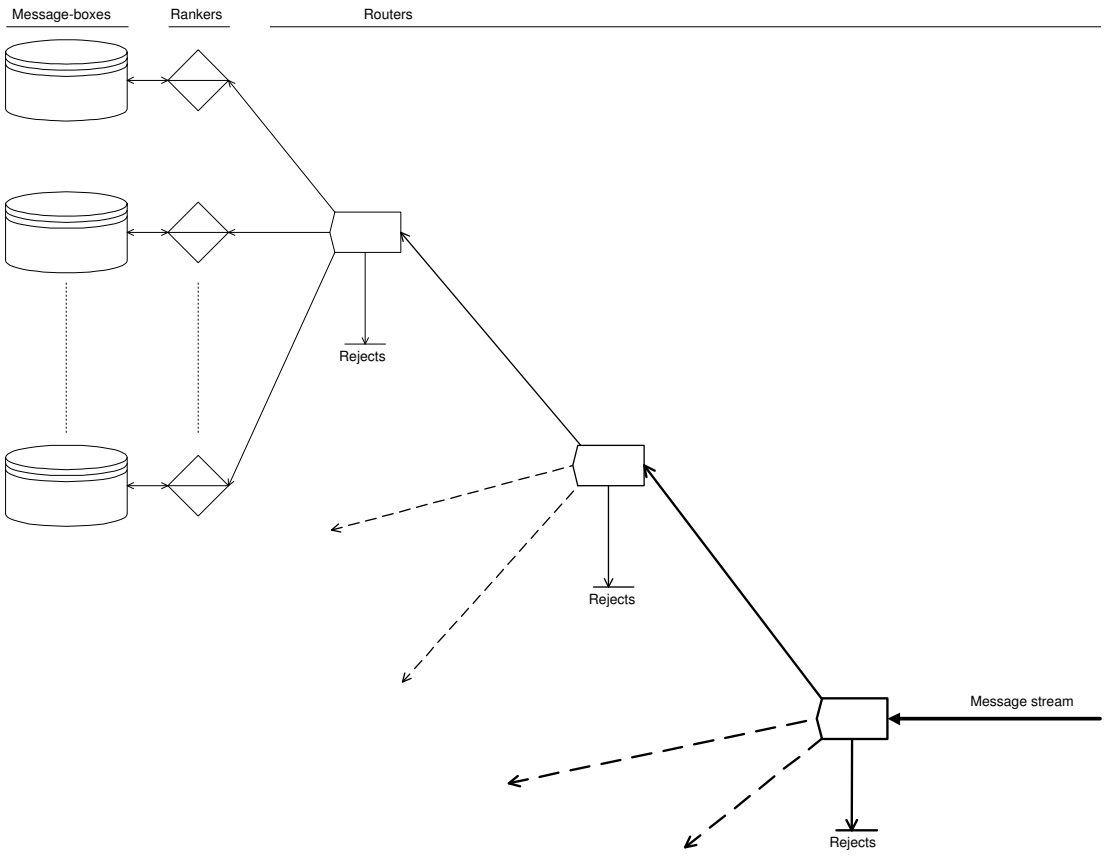

Figure 2: Information routing

carriers is provided in section 3. A discussion on how user and information need specific requirements can be introduced in the discovery process is put forward in section 4. As we will argue, *aboutness* plays an important role in determining the relevance of an information carrier. In section 4.3 we therefore take a closer look at this property before concluding the article.

2 Towards a Theory for Information Discovery

In developing a theory for information discovery, one must first resolve two fundamental questions. What is an information carrier, and what is the information carried by it. The latter question of course raises the issue of *what is information*? This section aims to provide our view on these questions.

2.1 What is an information carrier?

Thus far the term information carrier has been used without actually providing a definition. In the context of the net, an information carrier can be defined as:

> any entity that is accessible on the net, and which can provide information to other entities connected to the net.

A definition that truly supports the open character of the net. Examples of information carriers included are:

- Web pages (including free text, sound, images, and video fragments)
- Free text databases
- Traditional (relational, object-oriented, ...) databases. Both the databases as such, as well as their *instances*
- People's e-mail addresses
- Information about the location of non-electronic information carriers
- Aggregations/groupings of information carriers

A very special class of information carriers are aggregated information carriers. An obvious example of an aggregation is a database. A database in itself is an information carrier. However, it can also be seen as a collection of information carriers since each of its instances in itself carries information. Besides database based aggregations, one can imagine creating general collections of information carriers that are strictly based on some thematic commonality, or some common purpose. Information carriers can obviously be present in multiple aggregations.

2.2 Infomantics

What information exactly is has been studied intensively before. Different authors have provided alternative theories of information [Sha48, Lan86, Bar89, Los90, Los97]. The goals of this article do not include a definition of what information exactly is. We take a very modest approach to information theory. It is only assumed that information can be conceptualised as consisting of *information particles* called *infons* as suggested by Barwise [Bar89], and applied to the field of information retrieval by [RL96, Hui96b, Hui96a, HB96]. The set of all such infons is denoted by $\mathcal{IF}$. Some information particles will contain more information than others, therefore it is reasonable to presume the existence of an information containment relation $\succcurlyeq$ with as intuition:

If $f \succcurlyeq g$ then f contains at least the same information as g.

Two special elements in $\mathcal{IF}$ are presumed to exist. One element that represents *the most* information: $\top_i$, and one that contains *the least* information: $\bot_i$. The subscript i is used to denote that these elements belong to the infon space.

Infons express information about objects. Therefore, we also introduce a set of objects $\mathcal{IO}$. In [Bar89] a concrete notation for infons is proposed. For example, consider the infon $i = \langle\langle R, o_1, \ldots, \emptyset_n; 1\rangle\rangle$. This denotes that the objects $o_1 \ldots o_n$ stand in relation R to each other. (In predicate logic this would be denoted as $R(o_1, \ldots, o_n)$). In this article we are only interested in the objects that play a role in a given infon. To this end a function $\mathsf{Involve} : \mathcal{IF} \rightarrow \wp(\mathcal{IO})$ is presumed to exist. By way of illustration $\mathsf{Involve}(i) = \{o_1, \ldots, o_n\}$.

The conceptualisation of information discussed above can be captured formally by what will be termed *infon space*:

$$\mathcal{IS} \triangleq \langle \mathcal{IF}, \mathcal{IO}, \mathsf{Involve}, \succcurlyeq, \top_i, \bot_i \rangle$$

In infon space three axioms are presumed to hold. Firstly, the information containment relationship is assumed to be transitive. This is in line with Dretske's Xerox principle[BE90]. Intuitively, this principle states that information is nested, so if A contains information about B and B contains information about C, then A contains information about C.

[IS1] (Transitivity) $f \succcurlyeq g \succcurlyeq h \Rightarrow f \succcurlyeq h$

Furthermore, the special constants behave as they are expected to do:

[IS2] (Extremes) $\forall_{f \in \mathcal{IF}} [\top_i \succcurlyeq f \succcurlyeq \bot_i]$

Finally, information containment leads to containment on the objects involved:

[IS3] (Containment) $f \succcurlyeq g \Rightarrow \mathsf{Involve}(f) \supseteq \mathsf{Involve}(g)$

Note that the reverse implication does not generally apply.

The broad view of what information is, as taken in this article, is in line with the approaches taken in [Lan86] and [Bar89].

2.3 Infomantics of information carriers

The information need a user U may have corresponds to the need for infons. This is modeled by the function $\mathsf{Demand}_U : \mathcal{IN} \rightarrow \wp(\mathcal{IF})$, where $\mathcal{IN}$ represents the set of possible information needs. This latter set is an imaginary set in the sense the elements cannot be denoted or named. An information need is an abstract and subjective concept that only exists in a user's mind. However, for the purpose of our discussions it is convenient to presume such a set to exist.

Depending on the media used for an information carrier, users can read the carrier, listen to it, or view it. As a generalisation, the term *experience* shall be used. By stating that a user *experiences* an information carrier, it is meant that the user reads, or views, or listens to, the information carrier; in other words, a user using their sense organs to take up the information provided by the information carrier.

The information needed by users is provided on information carriers. Formally, information carriers are introduced as the set $\mathcal{IC}$, which is presumed to be closed under carrier composition (so any combination of given information carriers is another information carrier). As stated before, a carrier really only carries data. The information carried by a carrier depends on the user. In other words, information is transferred to the user via the carrier. The information transferred can be expressed in terms of infons:

$$\mathsf{Supply}_U : \mathcal{IC} \rightarrow \wp(\mathcal{IF})$$

The supply function is highly user dependent. In [Bar89], Barwise discusses an illustrative example of a person encountering a tree stump. Inspired by this example, now consider the following:

When some person encounters a tree stump, they may simply conclude from this situation that there used to be a tree here. Another person may come along and see from the rings on the stump that the tree was in fact 20 years old when it was felled. Yet another person may see from the colourations of the rings that in its tenth year the tree survived a forest fire.

This example goes to illustrate the subjectivity of the Supply_U function as each person is extracting different infons from the situation. Additionally, it is reasonable to assume that for a given user U, the supply function is not constant. Depending on the mood, fatigue, etc., a user may absorb differing amounts of information.

The semantics of an information carrier c, its infomantics as it will be termed, in the context of a user U is defined as the set of infons it provides to the user:

$$\mathsf{Supply}_U(c)^*$$

where for any $x \in \mathcal{IF}$ and $X \subseteq \mathcal{IF}$, we use the following definition of *infomantic closure* of X:

$$\begin{aligned} x^* &\triangleq \{ f \mid f \succcurlyeq x \} \\ X^* &\triangleq \bigcup_{x \in X} x^* \end{aligned}$$

The intuition behind infomantic closure is the following. Given some information X (represented by a set of infons), the closure is all information contained in these infons. The infomantic closure captures implicit information. In this way, we can model, for example, that in the infomantic closure of "salmon" we have the information "fish".

The information need N will be satisfied by a set of infons which relieves the need. The infon set is referred to as the demand of the need, denoted $\mathsf{Demand}_U(N)$. The subscript U reinforces the intuition that the demand is user dependent.

A notion of *relevance* between an information need N and a carrier c can be modeled as a supply and demand of infons:

$$c \,\mathsf{RelevantTo}_U\, N \triangleq \mathsf{Demand}_U(N)^* = \mathsf{Supply}_U(c)^*$$

Observe that carriers are considered relevant if and only if they meet the whole demand, and nothing but the whole demand. In practice, however, this is too strong a requirement. Carriers will only provide parts of the information need, or may provide too much information. Therefore, an order based on preference, a so-called preferential ordering, on information carriers needs to be introduced. This order needs to be such that the closer an information carrier matches the actual need the more preferred the carrier is.

In an ideal situation, for a given information need N, information discovery now involves searching for the proper information carriers such that they are considered relevant:

$$\mathsf{Search}_U(N) \triangleq \{ c \in \mathcal{IC} \mid c \,\mathsf{RelevantTo}_U\, N \}$$

This paints an idealistic situation. In practice, these Demand_U and Supply_U functions will not be available as concrete and well-defined functions.

The above discussion on the nature of information discovery allows us to highlight a key difference between information filtering and information discovery. In the case of information filtering the information needs (the information interests) involved have a more static and persistent nature. Information need in the context of information discovery tends to have a more temporary and ad hoc nature.

With regards to information filtering the view can be taken that a user (or work-group) may have a number of different interests they would like to be kept informed about. These *information interests* can be expressed as a set of (dormant) information needs $N_1, \ldots, N_n$. A set of incoming messages X can then simply be viewed as a subset of all (at that moment) known information carriers $\mathcal{IC}$. If X is a set of such incoming messages, then for each interest N_i the relevant messages are given by:

$$\mathsf{Filter}_U(N_i, X) \triangleq \{ c \in X \mid c \,\mathsf{RelevantTo}_U\, N_i \}$$

If $c \in \mathsf{Filter}_U(N_i, X)$, then carrier c is deemed relevant for interest N_i. This definition illustrates that in the ideal situation, the filtering mechanism must also have an understanding of the supply function of user U.

2.4 Preferential ordering in information discovery

Each relevant information carrier is preferred over each irrelevant information carrier. This can be generalized into a preferential ordering of information carriers: $\langle \mathcal{IC}, \sqsubset_N \rangle$. If $c \sqsubset_N d$, then c is preferred over d to fulfill the user's information need N.

There are two requirements on a preferential ordering that express its semantics in terms of the infon space. These requirements state that if the supply of infons from an information carrier c more closely matches the demand of an information need N, than the supply of a carrier d, then c is preferred over d (in the light of this information need). In other words c "fits" N better than d. The first requirement states that c is preferred over d as it over supplies the demand of the information more closely than d does:

[PO1] $\mathsf{Demand}_U(N)^* \subseteq \mathsf{Supply}_U(c)^* \subset \mathsf{Supply}_U(d)^* \Rightarrow c \sqsubset_N d$

The second requirement states that c is preferred over d as it under supplies the information need less than d does:

[PO2] $\mathsf{Supply}_U(d)^* \subset \mathsf{Supply}_U(c)^* \subseteq \mathsf{Demand}_U(N)^* \Rightarrow c \sqsubset_N d$

The above requirements express how the preferential ordering on information carrier is a consequence of the supply of information offered by the carriers in relation to the demand of the information imposed by the need. This sheds light on the nature of the preferential ordering assumed by several authors[BH95a, Won96, AG96, BH95b]. The requirements do not (and cannot) lead to an operational definition of the preferential ordering. In practice, this ordering must be approximated. For example, Amati et al[AG96] employ a standard relevance feedback mechanism to construct a preferential ordering on terms (primitive information carriers). Berger & Huibers[BH95a] use a navigation path through a thesaurus as a means of gleaning a preferential ordering on situations (sets of infons). Bruza & Linder[BL97] and Wondergem[Won96] propose using a query by navigation mechanism to distill positive and negative user preferences for information. A correspondence theorem shows how these preferences identify a preferential ordering on the underlying set of documents.

In short, preferential orderings are an emerging semantic framework for information retrieval theory. Later in this article we will show how they can be used to underpin a theory of information discovery.

2.5 Preferential ordering in information filtering

In the case of information filtering, preferential ordering can be used to provide an ordering on the contents of message-boxes; i.e. the ranking modules from figure 2. Each message-box has an associated information interest N_i. This information interest will be used by the routing modules to do the actual routing, but it can also be used to provide a ranking on the messages within a message-box.

2.6 Summary

This section has presented a formal framework for defining some essential concepts in information discovery. The user's information need has been conceptualized as a supply and demand situation involving information particles. More specifically, the information need is a demand for information particles. Information carriers supply information particles. Relevance is defined as supply meeting demand. Additionally, it is proposed that preferential orderings on the information carriers is a consequence of the fact that some information carriers meet the demand of the information need better than others.

3 Logical foundations of Information Discovery

Information discovery has its roots in the field of information retrieval. Over the last thirty years a number of information retrieval models have been developed. These have mostly been numeric models conceived solely for driving the information retrieval process. Such models have advanced the field of information retrieval from a practical point of view, but have not proven to be instructive in answering the more fundamental questions about information retrieval itself. This has led some researchers to turn to logic as a means to find the answers to these questions.

In recent years the logic-based approach to information retrieval has clearly come to the fore as a framework for investigating such questions [LR92, Rij93, Bru93, Lal96, CL96, RL96, Hui96a, Nie90]. Recent surveys of the area have been prepared by Lalmas[Lal98], Lalmas & Bruza[LB] and Sebastiani[Seb98]. These investigations appeal as they place information retrieval in a neutral framework (independent of any given retrieval model) and allow it to be described at a level of semantic detail hitherto not possible. Revealing insights have thus been gained, and as a by product, an underlying theory for information retrieval is beginning to take shape.

For the above reasons, as well as clarity of exposition, we propose a logic-based approach to information discovery. This logic will be based on the preferential ordering introduced in the previous section.

3.1 Carrier logic

When judging whether a given information carrier is more preferred than another carrier, a user first needs to determines the relevance of the carriers involved.

When humans judge the relevance or irrelevance of information carriers, they tend to do so in terms of properties they observe the carriers to have. These properties are collectively referred to as metadata, and each of the individual properties as a metadata attribute [WGMD95]. Metadata attributes may range from fairly simple such as: *authorship*, *medium*, *pricing*, *quality*, and *location*, to extremely complicated such as: *the information provided (infomantics)*. No explicit choice on the set of metadata attributes for information carriers will be made in this article; a more general approach is adopted. The following signature format is used as a basis for for the syntax of the *carrier logic*:

$$\Sigma \triangleq \langle \mathcal{MD}_1, \ldots, \mathcal{MD}_l; f_1, \ldots, f_m; R_1, \ldots, R_n \rangle$$

In this signature, $\mathcal{MD}_1, \ldots, \mathcal{MD}_l$ represent sets of $\mathcal{M}$eta$\mathcal{D}$ata values, such as `Price` and `Author`. Functions on these values are provided by the function symbols $f_1, \ldots, f_m$, for instance $+$ and $-$ on `Prices`. The set of relations symbols $R_1, \ldots, R_n$ provide relations over the metadata attributes. Example relation symbols would be Author and About (more about this relation shortly).

For example,

$$\langle \texttt{Author}, \texttt{String}, \texttt{RegExpr}; ; \mathsf{FirstName}, \mathsf{LastName}, \mathsf{Author}, \mathsf{Like} \rangle$$

where `Author` is a set of authors, `String` a set of strings, `RegExpr` a set of regular expressions, $\mathsf{FirstName}, \mathsf{LastName} \subseteq \texttt{Author} \times \texttt{Name}$ predicates matching first and last names to authors, and $\mathsf{Like} \subseteq \texttt{Name} \times \texttt{RegExpr}$ a regular expression checker that sees if the given name matches the regular expression. Note that this example signature does not include any function symbols. An example requirement for the relevance of an information carrier would be:

$$\exists_{a,f} \left[\mathsf{Author}(a) \wedge \mathsf{FirstName}(a, f) \wedge \mathsf{Like}(f, \text{"}\mathsf{E}\{\mathsf{a} - \mathsf{z}\}^*\text{"}) \right]$$

which requires the information carrier to be written by an author with a first name starting with an E.

One relevance criterion does deserve explicit attention. This is the *aboutness* of information carriers, i.e. a representation of the infomantics. Aboutness of information carriers is at the very heart of information

discovery. The underlying hypothesis is that if an information carrier is about the request from the user, then there is a high likelihood that the information carrier is indeed relevant to this need. For any information carrier it is relevant to discuss its infomantics in terms of what it is about. An aboutness specific metadata signature is defined as follows:

$$\langle \mathcal{KW}; \oplus; \mathsf{About} \rangle$$

where $\mathcal{KW}$ is a set of keywords and $\oplus$ is used to combine simple keywords into composed keywords. For example:

$$\mathsf{tiger\ prawn} \triangleq \mathsf{tiger} \oplus \mathsf{prawn}$$

Saying that a document is about tiger-prawns but not about tigers, can be expressed as:

$$\mathsf{About}(\mathsf{tiger} \oplus \mathsf{prawn}) \wedge \neg\, \mathsf{About}(\mathsf{tiger})$$

Now that we have defined what the signatures for a carrier logic look like, the actual logic itself can be discussed. Given a signature Σ, a language $\mathcal{CL}$ of well-formed logic formulae can be derived in the usual fashion. The resulting logic will be referred to as the *carrier logic*. A carrier c is deemed relevant to a formula ϕ via the satisfaction relationship $\models^{\mathsf{C}}$ over $\mathcal{IC} \times \mathcal{CL}$. It is not the aim of this article to go into detail on the definition of $\models^{\mathsf{C}}$. Different ways of 'implementing' this satisfaction relationship exist. For example, using a first order logic approach [MSST93], or an approach based on Kripke structures [Nie92]. Our aim is to study aboutness in more general terms, and rather define generic requirements *on* the definition of $\models^{\mathsf{C}}$ (and thus implicitly on the way it is "implemented") than limiting ourselves to one particular approach.

To summarise, for a given signature Σ of metadata we have the following carrier logic:

$$\mathbb{CL} \triangleq \left\langle \mathcal{CL}, \mathcal{IC}, \models^{\mathsf{C}} \right\rangle$$

3.2 Carrier reasoner

With a carrier logic we have a logic with which we can reason about the relevance of information carriers. However, the logic is not complete without a set of formulae, a theory, which defines the semantics of the different metadata attributes and operations. For example, for the metadata attribute `Price`, and operations $+$ and $-$, we would expect to hold:

$$(a + b) - b = (a - b) + b$$

Formally, a *carrier reasoning system* (carrier reasoner for short) can now be defined as a tuple:

$$\mathbb{CR} \triangleq \langle \mathbb{CL}, \Phi \rangle$$

where Φ is the theory defining the semantics of the operations and relations of the metadata attributes. The satisfaction relationship should honour the theory for the given carrier reasoner, so we should have:

$$\models^{\mathsf{C}} \Phi$$

3.3 Measuring the quality of the carrier reasoner

Inspired by the recall and precision measures found in the field of information retrieval, quality measures for carrier reasoners can be formulated. The satisfaction relationship $\models^{\mathsf{C}}$ expresses what the carrier reasoner 'believes' to be relevant information carriers. This is the 'computer' perspective. We should also take the user's perspective into consideration. To do this, we need to introduce an alternative semantics for the carrier logic:

$$c \models^{\mathsf{U}}_{\mathsf{N}} \phi \text{ iff user } U \text{ observes } c \text{ to support } \phi \text{ in the context of information need } N$$

In other words: would the user find carrier c to be relevant for query ϕ when trying to satisfy information need N?

When ϕ only deals with simple metadata attributes like prices, authors, etc. the user based semantics will generally be clear and most likely be an exact match to the semantics of the carrier reasoner. However, in the case of aboutness these semantics become less obvious due to the subjectivity of aboutness. Therefore, this is not yet a satisfactory definition of the user's view. We can go even one step further. When formulating their information need, users will express this in terms of some formulae taken from $\mathcal{CL}$. These formulae are referred to as *clues* as they provide the carrier reasoner with clues on the information need of the user.

3.4 Formulating information needs

In the remainder of this section we will look at how realistic the assumption that users can formulate clear clues about their information need really is. We will also highlight how a system can help users with the task of providing these clues.

The *clues* which a user U is able to give us about their information need can be captured by a predicate $\mathsf{HasClue}_U \subseteq \mathcal{IN} \times \mathcal{CL}$, where $\mathcal{IN}$ is the set of possible information needs and $\mathcal{CL}$ is the carrier logic language.

- Suppose a student is writing a report on river-pollution in The Netherlands. However, the student isn't familiar with pollution at all. The student does know that Greenpeace has, on numerous occasions, shown their concern about the pollution, and therefore assumes this is the case.

  To find more information, the student turns to an information discovery system to learn more about pollution. All this student knows about the needed information at this stage is that it must deal about pollution of rivers in The Netherlands. So if the student's information need related to the task of writing the report is N, then we have:

  $$\mathsf{HasClue}_U(N, \text{polution of rivers in The Netherlands})$$

  To the information discovery system, this is the first clue about the student's real information need.

With the above predicate, and the $\mathsf{RelevantTo}_U$ predicate as introduced in section 2.3, the following more exact definition of $\models^{\mathsf{U}}_{\mathsf{N}}$ can be provided:

$$c \models^{\mathsf{U}}_{\mathsf{N}} \phi \;\triangleq\; c\,\mathsf{RelevantTo}_U\, N \wedge N\,\mathsf{HasClue}_U\,\phi$$

In other words, the user would say that carrier c supports ϕ, iff carrier c is about an information need N with as clue ϕ. In figure 3 this definition is put in context. What should not be forgotten is that while a user is searching for information, they may already be learning more information that is relevant to their knowledge gap, possibly leading to a change in the actual information need.

Using the above user based semantics, two quality measurements for a carrier reasoner can be asserted. A carrier reasoner is called *precise* if:

$$\text{if } c \models^{\mathsf{C}} \phi \text{ then } c \models^{\mathsf{U}}_{\mathsf{N}} \phi$$

and *exhaustive* if:

$$\text{if } c \models^{\mathsf{U}}_{\mathsf{N}} \phi \text{ then } c \models^{\mathsf{C}} \phi$$

A precise carrier reasoner leads to a high degree of precision when used for retrieval, while an exhaustive reasoner would lead to a high recall.

The art of defining a carrier reasoner lies in finding a good balance between the resulting precision and recall. Building a carrier reasoner that is both exhaustive and precise is still an ideal, and may in fact be impossible. In practical situation a well founded trade off needs to be made between these two.

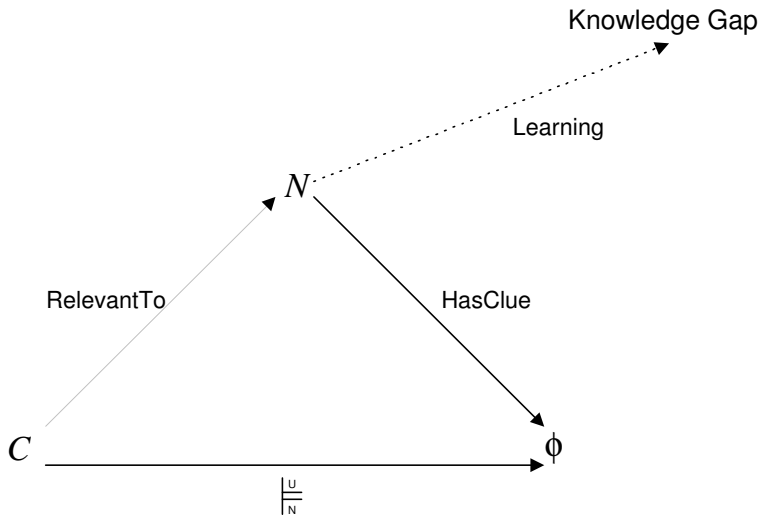

Figure 3: A user's view on relevance of information carriers

For example, information filtering systems may sometimes want to focus on exhaustivity. Missing important news items that are relevant will be considered more harmful than having to discard the 'odd' irrelevant carrier which evades the filtering mechanism. On the other hand, information discovery systems, which help users in discovering new information in 'unchartered waters', precision will be preferred to prevent users from drowning in new information. What can not be stressed enough though, is that carrier reasoning is a delicate balancing act between precision and exhaustivity.

The $\models^{U}_{N}$ relation depends on individual users and their specific information needs. This would imply that the quality of a carrier reasoner needs to be evaluated for each individual user and information need. In practice this is obviously very hard to cater for. An often used pragmatic way to circumvent this is to assume a definition of $\models^{U}_{N}$ which would satisfy the "average" user and information need; a consensus definition. In the context of the TREC (Text Retrieval Conference) conference series, a standardised database of information carriers, queries, and subsets of relevant carriers are used to evaluate information retrieval systems. This would lead to a pre-defined consensus definition of $\models^{U}_{N}$ that can be used to evaluate a carrier reasoner.

For a carrier reasoner, it means that some mechanism is required to make $\models^{C}$ more user and information need aware. The defaults harboured by a user should somehow be taken into consideration during the reasoning process. To this end we shall introduce a preference logic.

3.5 Summary

In this section, we have explored some of the logical foundations of information discovery. More specifically, we have introduced a carrier logic that enables the formal reasoning about the relevance of an information carrier for a user's information need.

It cannot be assumed that users will always be able to formulate exactly what their information needs are. Therefore specific attention was paid to a mechanism to aid users in expressing their information needs.

4 User preferences and the Carrier Logic

Whenever humans communicate with each other, a contextual background is often assumed. One way to view this background context is via a frame-based cognitive model [Bar92]. The frames are constructed

by attributes which may take on certain values. For example, the attribute surfing may take on the value wave, thus modeling the concept wave $\oplus$ surfing. It turns out that humans prime certain attributes with default values. A mismatch in defaults between two people communicating can therefore lead to mis-communication.

4.1 Preferences

In an information discovery setting a mis-communication between user and discovery system may occur; usually resulting in the selection of irrelevant information. When a users wants to learn something about surfing, while harbouring the default wave $\oplus$ surfing, the system should preferably not present information sources about wind surfing and certainly not about internet surfing. An advanced information discovery system will learn a user's preferences and anticipate further preferences based on those it has.

User preferences are intimately tied to the preferential ordering. For example:

$$\mathsf{About}(\mathsf{surfing}) \mid\!\sim \mathsf{About}(\mathsf{wave} \oplus \mathsf{surfing})$$

states

All preferred information carriers about 'surfing' are also about 'wave surfing'.

$$\mathsf{About}(\mathsf{nuclear} \oplus \mathsf{physics}) \mid\!\sim \mathsf{Author}(\mathsf{A.\ Einstein})$$

expresses that

All preferred information carriers about 'nuclear physics' are authored by A. Einstein.

User preferences can be derived from relevance feedback. Bruza & van Linder [BL97] propose translating a user's navigation path through a hyperindex into preferences. A hyperindex is a partially ordered set of index terms that can be browsed. For example, a given user may browse from surfing to surfing in hawaii to surfing conditionsin hawaii. Such a path can be translated into the following preferences:

1. true $\mid\!\sim$About(surfing)
2. About(surfing) $\mid\!\sim$About(hawaii)
3. About(surfing $\oplus$ hawaii) $\mid\!\sim$About(conditions)

The first preference states that the preferred information carriers are about surfing. The second preference expresses that the preferred information carriers about surfing deal with Hawaii. The third preference states that the preferred information carriers about surfing and Hawaii are also about conditions. (This reflects the information need being about surfing conditions in Hawaii).

Amati and Georgatos [AG96] put forward a method whereby positive and negative preferences are gleaned via a relevance feedback on documents in the result set. The preferences are based on *priority relations* denoted by $\prec_p$ and $\prec_n$. Both of these relations are defined over a set of terms T. The positive priority relation $\prec_p$ on terms is defined in terms of positive relevance feedback:

$$t_1 \prec_p t_2 \text{ iff } \left|D_{t_1}^+\right| \leq \left|D_{t_2}^+\right|$$

In the above formula $\left|D_t^+\right|$ represents the number of documents containing term t that the user has identified as being relevant. Similarly, the negative priority relation $\prec_n$ can be defined. Preferences of the form:

$$\mathsf{About}(t_1) \mid\!\sim \mathsf{About}(t_2)$$

can be defined using these relations.

A preference logic can be built using the user preferences. This allows the possibility to reason with user preferences and deduce new preferences. More about this in the next section. The preference logic we use here is defined on top of the existing carrier logic. For a given carrier logic: $\mathbb{CL} \triangleq \left\langle \mathcal{CL}, \mathcal{IC}, \models^{\mathsf{C}} \right\rangle$, we can therefore define an associated preference logic:

$$\mathbb{PL} \triangleq \left\langle \mathcal{PL}, \mathcal{IC}, \models^{\mathsf{C}} \right\rangle$$

Let $f, g \in \mathcal{CL}$ be closed formulae, then the language $\mathcal{PL}$ itself is defined by the following two rules:

1. $f \mathrel{|\!\sim} g \in \mathcal{PL}$
2. $f \mathrel{|\!\not\sim} g \in \mathcal{PL}$

The semantics are expressed relative to a preferential order on information carriers. For the preference logic we have a satisfaction relationship:

$$\models^{\mathsf{C}} \subseteq \wp(\mathcal{IC} \times \mathcal{IC}) \times \mathcal{PL}$$

Let $P = \langle \mathcal{IC}, \sqsubset_C \rangle$ be a preferential ordering over information carriers. Ideally, $\sqsubset_C$ will be a close match to the one ($\sqsubset_N$) harboured by the user. For a given formula f from the carrier logic, it can be now expressed that a carrier c is *the most preferred carrier* for f according to P by making the satisfaction relationship from the carrier logic aware of the preferential ordering:

$$\langle c, P \rangle \models^{\mathsf{C}} f \triangleq c \models^{\mathsf{C}} f \text{ and for each } d \sqsubset c \text{ we have: } d \not\models^{\mathsf{C}} f$$

When stating:

$$\mathsf{About}(\mathsf{surfing}) \mathrel{|\!\sim} \mathsf{About}(\mathsf{wave} \oplus \mathsf{surfing})$$

the intention was that all preferred information carriers about surfing were about wave $\oplus$ surfing. So if $c \models^{\mathsf{C}}_P \mathsf{About}(\mathsf{surfing})$, it should also be the case that $c \models^{\mathsf{C}} \mathsf{About}(\mathsf{wave} \oplus \mathsf{surfing})$.

With such an ordering the semantics of the preference logic can be expressed as follows. Let $c \in \mathcal{IC}$, $f, g \in \mathcal{CL}$, and $p, q \in \mathcal{PL}$, then:

1. $P \models^{\mathsf{C}} f \mathrel{|\!\sim} g$ iff for any $c \in \mathcal{IC}$ with $\langle c, P \rangle \models^{\mathsf{C}} f$ we also have $c \models^{\mathsf{C}} g$
2. $P \models^{\mathsf{C}} f \mathrel{|\!\not\sim} g$ iff $P \not\models^{\mathsf{C}} f \mathrel{|\!\sim} g$

Note that there is an important difference between $P \models^{\mathsf{C}} \mathsf{About}(\mathsf{surfing}) \mathrel{|\!\not\sim} \mathsf{About}(\mathsf{web} \oplus \mathsf{surfing})$ and $P \models^{\mathsf{C}} \mathsf{About}(\mathsf{surfing}) \mathrel{|\!\sim} \neg\, \mathsf{About}(\mathsf{web} \oplus \mathsf{surfing})$. The former rule expresses that some preferred carriers on surfing are not about web $\oplus$ surfing. The latter rule expresses that none of the preferred carriers about surfing should be about web $\oplus$ surfing; i.e. surfing preferentially precludes web $\oplus$ surfing.

A carrier reasoner $\mathbb{CR}$, can now be extended to a preference reasoner:

$$\mathbb{PR} \triangleq \left\langle \mathbb{CR}, \mathcal{PL}, \models^{\mathsf{C}} \right\rangle$$

The preference reasoner is an extension of the carrier reasoner by explicitly taking user preferences into account via the preferential ordering. The quality measures introduced earlier, precision and exhaustivity, can be employed to judge the preference reasoner. *a*

4.2 Preference reasoning

Preference reasoning requires sound inference rules. The inference rules allow new preferences to be derived based on the preferences that the user has expressed. Soundness insures that derived preferences are consistent with the ones expressed by the user. The following are a non-exhaustive selection of inference rules that have been put forward by several authors [BL97, AG96, Won96]. They are intended to illustrate patterns of inference available within a preference reasoner. Let α, β, γ be formulae in the preference logic:

$$\alpha \mathrel{|\!\sim} \alpha \quad \text{reflexivity}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta \quad \alpha \mathrel{|\!\sim} \gamma}{\alpha \mathrel{|\!\sim} \beta \wedge \gamma} \quad \text{and}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta \quad \alpha \wedge \beta \mathrel{|\!\sim} \gamma}{\alpha \mathrel{|\!\sim} \gamma} \quad \text{cut}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta \quad \alpha \mathrel{|\!\sim} \gamma}{\alpha \wedge \beta \mathrel{|\!\sim} \gamma} \quad \text{rational monotonicity}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta \quad \alpha \not\mathrel{|\!\sim} \neg\gamma}{\alpha \wedge \gamma \mathrel{|\!\sim} \beta} \quad \text{rational monotonicity}$$

As a small illustration of preference reasoning, consider the following: A user who has expressed a preference for information about surfing in Hawaii ($\mathsf{About}(\mathsf{surfing}) \mathrel{|\!\sim} \mathsf{About}(\mathsf{hawaii})$) and then refined this preference to surfing conditions in Hawaii:

$$\mathsf{About}(\mathsf{surfing}) \wedge \mathsf{About}(\mathsf{hawaii}) \mathrel{|\!\sim} \mathsf{About}(\mathsf{conditions})$$

Using the cut rule we can conclude that the user is interested in information about surfing conditions:

$$\mathsf{About}(\mathsf{surfing}) \mathrel{|\!\sim} \mathsf{About}(\mathsf{conditions})$$

Using the priority relations of Amati and Georgatos, we could assume that the term "wind" has a high priority in the relation $\prec_p$. As a consequence, the term "surfing" does not preclude the term "wind":

$$\mathsf{About}(\mathsf{surfing}) \not\mathrel{|\!\sim} \mathsf{About}(\mathsf{wind})$$

Using rational monotonicity permits the derivation:

$$\mathsf{About}(\mathsf{surfing}) \wedge \mathsf{About}(\mathsf{wind}) \mathrel{|\!\sim} \mathsf{About}(\mathsf{conditions})$$

In other words, the preference reasoner has derived a preference that the user is interested in wind surfing conditions. Such inferences could be shown to the user for relevance feedback (The formulae wouldn't be shown, but suitable textual representations of the intended preference represented by the formula). In this way the user could navigate the proof space generated by the preference reasoner as part of the information discovery process. This has the additional computational advantage that only a part of the proof space needs to be computed and shown to the user. Feedback from the user could then be used to guide further inferences. A complimentary approach to this has been advocated by Ohsawa & Ochida using abuctive reasoning [OY97].

Preferential reasoning can be used in an information filtering setting as well. As stated above, in our view information filtering involves two aspects: ranking and routing. For ranking, the above discussed ranking strategies for information discovery can be directly applied. Each message-box (refer to figure 2) may have associated a set of preferences (clues) C that determine the ranking of messages within that message-box.

Routing of messages can be expressed by means of predicates over information carriers. The predicate can be used to evaluate if a given message (or rather the underlying information carrier) should be routed through a 'message filter' to a next router or a message-box. In other words:

An information carrier c is accepted for a routing filter r iff its preferences are compatible with the user's preferences.

The actual predicate can be defined as follows:

$$\mathsf{Accept}(c, r) \Leftrightarrow \exists_P \left[P \mid\!\!\overset{\mathsf{C}}{=} \Pi(r) \cup \chi(c) \right]$$

where $\Pi(r)$ is a set of clues expressing the user's information interests defining filter r, and $\chi(c)$ provides the clues supported by the information carrier itself. The latter set of clues can be expressed by the following set of preferences:

$$\chi(c) \;\triangleq\; \left\{ \mathit{true} \mid\!\sim \phi \;\middle|\; c \mid\!\!\overset{\mathsf{C}}{=} \phi \right\}$$

4.3 What is aboutness?

An interesting aspect of aboutness is that we need other information carriers to express what a given information carrier is *about*. If we wouldn't do so, we would not be able to discuss with fellow humans (e.g. to a librarian), or computers for that matter, what an information carrier is about. When stating: "I want to know something about surfing", the information carrier surfing expresses what we want to be informed *about*. Aboutness can therefore be seen as a relationship between information carriers: $\mathsf{IsAbout} \subseteq \mathcal{IC} \times \mathcal{IC}$.

Those information carriers that are used to express what other information carrier are about play a special role. They will be referred to as *promises*. Stating that a given book is about surfing can be viewed as a *promise* with regards to the information provided in the book. The actual promise would be: "by reading this book, you'll be informed about the act of surfing". These promises can be used to express what information carriers are about.

A number of known mechanisms exist to characterise the aboutness of information carriers in terms of promises. This ranges from sets of keywords [Rij79], (weighted) vectors of keywords [SM83, Sal89], index expressions [Bru93], term phrases [Lew92], to conceptual graphs [Mya92]. These are all different ways of defining the promise language $\mathcal{PR}$. Each keyword, vector of keywords, index expression or conceptual graph makes a promise about the contents of the information carrier.

Let $\mathcal{PR} \subseteq \mathcal{IC}$ denote some set of promises, then this would allows us to more precisely define aboutness (in the context of a user U and information need N) as a relationship over information carriers $\mathsf{IsAbout} \subseteq \mathcal{IC} \times \mathcal{PR}$ where:

$$c \,\mathsf{IsAbout}_U\, p \;\triangleq\; c \mid\!\!\overset{\mathsf{U}}{\underset{\mathsf{N}}{=}} \mathsf{About}(p)$$

The properly define its semantics, we need to dig deeper into th issue of aboutness.

In determining if an information carrier c is about promise p, we are at the mercy of the users. If we were to confront a user with information carrier c and promise p, then why would this user say that c is about p? Conversely, when confronted with a promise p, why would a user select p as a good description of their information need? To try to answer this question we need to study the effect that a given promise p has on a user.

When a user experiences (usually reads) a promise p, then this promise will relate to existing knowledge of the user. This can be modeled by a function:

$$\mathsf{Activates}_U : \mathcal{PR} \rightarrow \wp(\mathcal{IO})$$

This function should yield the objects (in the user's current knowledge) that relate to an experienced promise. Using the popular saying 'that rings a bell', this function returns the set of 'bells' that 'start ringing'. For example, when the promise is *surfing*, bells like *wind*, *wave conditions*, *Hawaii* may ring (assuming the user is switched onto the sport surfing at that point).

The information supplied to the user when experiencing c is *not yet* part of the user's knowledge when experiencing the promise p. So the infons *activated* by p may be different from the infons supplied by

c, especially when p is close to the user's knowledge gap. Therefore, we should only refer to the actual objects involved in the activation. The *object match* between a carrier c and promise p can be defined as:

$$\mathsf{ObjectMatch}_U(c,p) \triangleq \mathsf{Involve}(\mathsf{Supply}_U(c)^*) \cap \mathsf{Activates}_U(p)^*$$

This leads to the following refined infomantics of aboutness:

$$c \,\mathsf{IsAbout}_U\, p \text{ iff } \mathsf{ObjectMatch}_U(c,p) \neq \varnothing$$

So, if there is an overlap between the set of objects activated when confronted with promise p and the information supplied by carrier c, user U is expected to consider c about p.

Similarly, if a user provides p as a clue of their information need N, then this has the following infomantics:

$$N \,\mathsf{HasClue}_U\, \mathsf{About}(p) \text{ iff } \mathsf{ObjectMatch}_U(N,p) \neq \varnothing$$

where

$$\mathsf{ObjectMatch}_U(N,p) \triangleq \mathsf{Involve}(\mathsf{Demand}_U(N)^*) \cap \mathsf{Activates}_U(p)^*$$

The infomantics of IsAbout and HasClue also illustrates the difficulty of the indexing task. When indexing an information carrier, promises must be selected that would be acceptable to most users as proper descriptions of the carrier's aboutness. Additionally, these promises should be chosen such that its is expected that users would actively (active memory) use them to provide clues on their information needs.

4.4 Summary

This section has formalized the notion of information preference. It is natural in the information discovery process, the user will prefer some information carriers over others. It is argued that information discovery systems must be able to reason with user preferences in order to be effective. User preferences can be gleaned by relevance feedback and reason with using the inference rules of model preference logic developed within the AI community. The aboutness relation between information carries is clarified by introducing the notion of *promise*. Aboutness is defined in terms of the infomantics by an object overlap function. Although this function cannot be implemented directly, the assumption is that it can be approximated via the preference reasoning system.

5 Conclusions

This article sketches the fundamentals of information discovery in terms of a logic/information-based framework. The main contribution of this article is a conceptual model of information discovery comprising relevant concepts and their inter-relationships. A feature that has repeatedly appeared is the user-centred nature of information discovery and thus its inherent subjectivity. Although this aspect has long been acknowledged in the field of information retrieval, few attempts have been made to integrate this troublesome aspect into a formal framework. We claim that this article has made some headway in this area, via (preferential) orderings and user-based functions.

In a sense, this article has raised more questions than it has answered. In particular, the model theory and the aboutness relation are areas deserving more attention. When an information carrier is treated as a model, what are the characteristics of this model? We have left this question unanswered as the model-theoretic underpinnings of information discovery are still under active investigation with little consensus reached. Aboutness is also an issue that is still crystallising. A number of logic-based approaches have recently emerged for studying this relationship and its associated properties. Once again no consensus has yet emerged, though some work using nonmonotonic logic has led to some interesting and comparable formal results [Hun95, BL97]. A thorough exposition of an axiomatic approach to aboutness can be found in [Hui96a]. Our future work will build on this article by proposing a model theory and defining the aboutness relation within this theory.

References

[AG96] G. Amati and K. Georgatos. Relevance as deduction: a Logical View of Information Retrieval. In F. Crestani and M. Lalmas, editors, *Proceedings of the Second Workshop on Information Retrieval, Uncertainty and Logic WIRUL'96*, pages 21–26. University of Glasgow, Glasgow, Scotland, 1996. Technical Report TR-1996-29.

[Bar89] J. Barwise. *The Situation in Logic*. CSLI Lecture Notes, 1989.

[Bar92] L.W. Barsalou. *Cognitive Psychology: an overview for cognitive psychologists*. Lawrence Erlbaum, Hillsdale, New-Jersey, 1992.

[BE90] Jon Barwise and John Etchemendy. Information, infons, and inference. In Robin Cooper, Kuniaki Mukai, and John Perry, editors, *Situation theory and its applications*, volume 1 of *CSLI Lecture Note Series*, pages 33–78. Center for the study of language and information, CSLI, 1990.

[BH95a] F.C. Berger and T.W.C. Huibers. A framework based on situation theory for searching in a thesaurus. *The New Review of Document and Text Management*, 1:253—276, 1995.

[BH95b] P.D. Bruza and T.W.C. Huibers. How nonmonotonic is aboutness? Technical Report UU-CS-1995-09, Department of Computer Science, Utrecht University, The Netherlands, March 1995.

[BL94] T. Berners-Lee. Universal Resource Identifiers in WWW. Technical Report RFC1630, IETF Network Working Group, June 1994.

[BL97] P.D. Bruza and B. van Linder. Preferential Models of Query by Navigation. In F. Crestani, M. Lalmas, and K. van Rijsbergen, editors, *Information Retrieval and Logic*. 1997. Forthcoming book chapter.

[Bru93] P.D. Bruza. *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia*. PhD thesis, University of Nijmegen, Nijmegen, The Netherlands, EU, 1993.

[CHSS98] H. Chen, A. Houston, R. Sewell, and R. Schatz. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7):604–618, 1998.

[CL96] F. Crestani and M. Lalmas, editors. *Proceedings of the Second Workshop on Information Retrieval, Uncertainty and Logic WIRUL'96*. Technical Report TR-1996-29. University of Glasgow, Glasgow, Scotland, 1996.

[Cle91] C.W. Cleverdon. The Significance of the Cranfield Tests on Index Languages. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, Illinois, October 1991. ACM Press.

[Des97] B.C. Desai. Supporting Discovery in Virtual Libraries. *Journal of the American Society for Information Science*, 48(3):190–204, 1997.

[HB96] T.W.C. Huibers and P.D. Bruza. Situations: A general framework for studying Information Retrieval. In R. Leon, editor, *Information retrieval: New systems and current research, Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialists Group*, pages 3–25. Taylor Graham, Drymen, Scotland, 1996.

[Hui96a] Theo Huibers. *An Axiomatic Theory for Information Retrieval*. PhD thesis, University of Utrecht, Utrecht, The Netherlands, 1996. ISBN: 90-3931207-9

[Hui96b] Huibers, T.W.C. and Lalmas, M. and Rijsbergen, C.J. van. Information Retrieval and Situation Theory. *SIGIR Forum*, 30(1):11–25, 1996.

[Hun95] A. Hunter. Using default logic in information retrieval. In C Froidevaux and J Kohlas, editors, *Symbolic and Quantitative Approaches to Uncertainty*, volume 946 of *Lecture Notes in Computer Science*, pages 235–242, 1995.

[Lal96] M. Lalmas. *Theories of Information and Uncertainty for the modelling of Information Retrieval: an application of Situation Theory and Dempster-Shafer's Theory of Evidence*. PhD thesis, University of Scotland, Glasgow, Scotland, 1996.

[Lal98] M. Lalmas. Logical models in information retrieval: Introduction and overview. *Information Processing & Management*, 34(1):19–33, 1998.

[Lan86] F. Landman. *Towards a Theory of Information*. Foris, 1986.

[LB] M. Lalmas and P.D. Bruza. The Use of Logic in Information Retrieval Modeling. In press.

[Lew92] D.D. Lewis. *Representation and Learning in Information Retrieval*. Technical report 91–93, Computer Science Department, University of Massachusetts, Amherst, Massachusetts, 1992.

[Los90] R.M. Losee. *The Science of Information*. Academic Press, 1990.

[Los97] R.M Losee. A Discipline Independent Definition of Information. *Journal of the American Society for Information Science*, 48(3):254–269, 1997.

[LR92] M. Lalmas and C.J. Rijsbergen. A logical model of information retrieval based on situation theory. In *Proceedings of the 14th Research Colloquium of the British Computer Society Information Retrieval Specialist Group*, pages 1–13. Springer Verlag, April 1992.

[Lyn95] C.A. Lynch. Networked Information Resource Discovery: An Overview of Current Issues (Invited paper). *IEEE Journal on Selected Areas of Communications*, 13(8):1505–1522, October 1995.

[MSST93] M. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A Model of information Retrieval based on Terminological Logic. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–307, 1993.

[Mya92] S.H. Myaeng. Using Conceptual Graphs for Information Retrieval: A Framework for Representation and Flexible Inferencing. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, pages 102–116, Las Vegas, Nevada, March 1992.

[Nie90] J. Nie. *Un modèle logique g'en'eral pour les systèmes de recherche d'informations - Application au prototype RIME*. PhD thesis, Institut National Polytechnique de Grenoble - Univeristé Joseph Fourier Grenoble, France, 1990.

[Nie92] J. Nie. Towards a Probabilistic Modal Logic for semantic-based Information Retrieval. In N. Belkin, P. Ingwersen, and A.M Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–151, Copenhagen, Denmark, June 1992.

[OY97] Y. Osawa and M. Yachida. An index navigator for understanding and expressing user's coherent interest. In *Proceedings of 15th International Joint Conference on AI - IJCAI-97*, pages 722–728, 1997.

[Rij79] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, 2nd edition, 1979.

[Rij93] C.J. van Rijsbergen. What is Information Anyway? In *Two Essays in Information Retrieval*. University of Glasgow, Glasgow, Scotland, 1993. Research Report IR-93-03.

[RL96] C.J. van Rijsbergen and M. Lalmas. Information calculus for information retrieval. *Journal of the American Society for Information Science*, 47(5):385–398, May 1996.

[Sal89] G. Salton. *Automatic Text Processing–The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts, 1989.

[Seb98] F. Sebastiani. On the role of logic in infromation retrieval. *Information Processing & Management*, 34(1):1–18, 1998.

[Sha48] C. E. Shannon. A mathematical theory of communication. In *The Bell System Technical Journal*, volume 22, pages 379–423,623–656, 1948.

[SM83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill New York, NY, 1983.

[SM94] K. Sollins and L. Masinter. Functional requirements for uniform resource names. Technical Report RFC1737, IETF Network Working Group, http://www.ietf.org/rfc/rfc1737.txt, December 1994.

[WGMD95] S. Weibel, J. Godby, E. Miller, and R. Danierl. Metadata workshop report. Dublin, Ohio, March 1995.

[Won96] B. Wondergem. Preferential Structures for Information Retrieval. Technical Report INF-SCR-96-21, Department of Computer Science, Utrecht University, The Netherlands, 1996.