Aptness based search on the Web*

B. van Gils, H.A. (Erik) Proper, P. van Bommel, Th.P. van der Weide

Institute for Computing and Information Sciences, Radboud University Nijmegen
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands, EU

Abstract. The Web has, in a relatively short period of time, evolved from a medium for information exchange between scholars to one of the most important media in modern times. This has had a major impact on the infrastructure supporting the Web. Retrieval systems that select relevant resources from the ever increasing volume of resources that are available to us are becoming more and more important. In our opinion, the traditional view on these systems (where 'topical relevance' seems to be the key notion) is too limited. The main contribution of this paper is an integral view on a more advanced scheme for search on the web called *aptness based retrieval*.

1 Introduction

The World Wide Web (the Web) has become increasingly important for performing our day to day activities. What started out as a medium for communication between scholars has evolved into one of the most important media in modern days. Several factors have contributed to this development.

First of all, the sheer volume of resources available to us has increased enormously over the last few years. In [1] it is called an "explosion of online information". Secondly, it is sometimes stated that anything can be found on the Web. Certainly, resources are available on many different topics. Not only the size of the Web, but also its usage, ranging from online communication via E-mail and instant messaging to E-governance and E-commerce has evolved. The Web is no longer a mere "static library" with information. Last but not least, the kinds of resources available online have evolved to include webpages, online databases, E-services and other interactive applications (See e.g., [2, 3]).

To cater for all these changes, the technical infrastructure supporting the Web has evolved over the years as well. The most prominent infrastructural changes are, in this respect, related to localization and transporting information over the Web. Examples in this include:

Localization – search engines, yellow-pages, service repositories
Transporting – HTTP, FTP, Jabber, Bittorrent, VoIP

* The investigations were partly supported by the Dutch Organization for Scientific Research (NWO).

It does not seem unreasonable to assume that the Web would collapse without proper tools for localizaing resources on the Web that are relevant to whatever task one has at hand. This is also reflected by the enormous amount of research that has been invested in *information retrieval* (IR) in the past [4, 5, 6] resulting in the IR paradigm. In this paradigm a users query is matched against the characterizations of a set of resources. Traditionally these resources were ranked only with respect to their topical relevance. Modern search engines such as *Google* indeed offer some additional capabilities. Finally, the ranked list is returned to the user who can select those resources that are of interest to him. As such, the main challenges in this field seem to be query formulation, characterization of resources, and matching of queries to characterizations.

A keyword based approach has traditionally been used for both the characterization of online resources as well as for query formulation. The underlying assumption is that keywords are a sufficiently good representation of the information conveyed by the document under consideration. Relevant examples include the vector space model where characterization is typically done by means of some word frequency measures and query by navigation with a much richer characterization scheme based on index expressions (See e.g., [7, 8]). The latter provides a rich characterization of the information conveyed by a document. For example, the index expression:

attitudes of (students of universities) to (war in Vietnam)

is a richer description than the keyword set consisting of the nouns in that expression. It is interesting to observe that in [9] it is stated that "most information retrieval systems on the Internet rely primarily on similarity ranking algorithms based solely on term frequence statistics". This implies that the traditional IR paradigm is still the predominant characterization mechanism for searching on the Web. By contrast, a review of the additional capabilities of modern search engines (such as *Google*'s advanced search) and information retrieval literature (e.g., the TREC conference[1]) suggests that other aspects of search are both recognized and being developed. Especially initiatives focussing on meta-data search and annotations are increasingly popular (see e.g., [10, 11]).

These additional capabilities that are being developed are, indeed, a step in the right direction. In our opinion even these more modern approaches do not go far enough; searchers should be able to express their entire information need and not only the informational aspects, either explicitly or automagically. In other words, users should be able to express such things as desired language, price, relations to other resources, required background knowledge, form, format, size, author, last modification date and so on. We dub this futuristic situation *aptness based retrieval*.

A radical change in thinking about valuation of resources in the Web is needed to be able to achieve such futuristic situation. The goal of this paper is twofold. Firstly, we want to present a thorough analysis of the problem domain which is firmly grounded in literature. Secondly we will show how such situation

[1] `http://trec.nist.gov/`

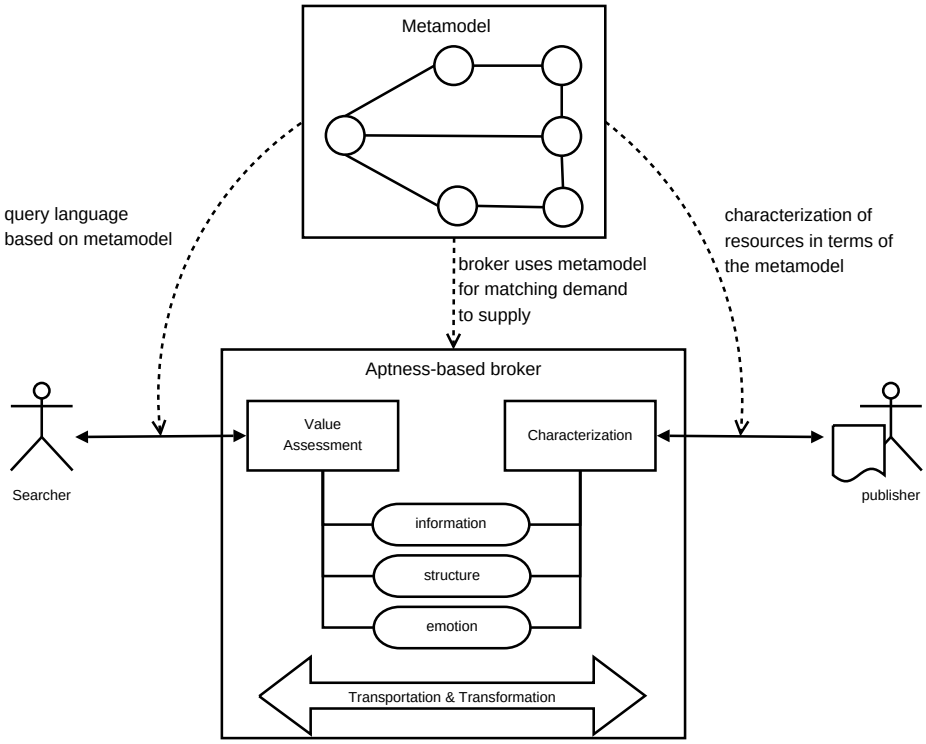

Fig. 1: Overview of our approach

might be implemented in practice. As such the main contribution of this paper is a comprehensive and novell view on search on the Web.

Figure 1 presents an overview of our approach. In previous work (e.g., [12, 13, 14, 15, 16]) we have (formally) described several (meta-) models pertaining to the information market. As a consequence, the examples in this paper (especially in Sections 3, 4 and, 6) can be interpreted / evaluated in a formal context.

2 The information market

When searching on the Web, queries are matched against (characterizations of) a set of resources that the search system knows about. As such, search systems attempt to estimate how valuable a resource is for a searcher by matching his query (which supposedly captures his information need) to the characterizations of the resources that it knows about. The claim that topic-based search is too limited thus means that we find the mechanism for valuing resources on the Web too limited. Issues in this area are both recognized and being addressed in contemporary research. For example, *Google*'s advanced search offers more than only topic based search. Similarly, the concept of meta-data search received

renewed attention with the advent of standard annotation languages such as RDF. Examples include [17], which describes meta-data filtering supported by RDF, and [18], which describes a query language that for RDF annotations of resources.

The concept of *value* is highly complex and is central to our discussion here. This notion is used in many fields, including mathematics, marketing, computer science and even personal and cultural values (see e.g., [19, 20, 21]). Before we discuss the notion of value on the information market in more detail we must firstly define what we mean by transactions. In our view a transaction is a specific, identifiable exchange of assets between two or more players. In the simplest case this would mean that one player exchanges one asset (a book) for another (€15) with another player. After all, money is just another asset! However, more complex transactions are possible as well. This also implies that concepts such as *buyer* and *seller* can only be identified in a transaction relative to the asset under consideration. In classic economic theory the focus on money suggests that whomever is paying money must be the buyer and whomever receives money must be the seller. One could, however, state that one sells €15. The assets on the information market are the resources that we typically find on the Web. Therefore, we define the searcher to be the player that receives a resource and the publisher to be the player that publishes it on the Web.

To explore this further, it is interesting to observe that there are two main classes of transactions of assets. Firstly, there is the trade of ownership, which implies that the ownership of a (physical) entity is transferred from one player to the next. This is, for example, the case when one buys a house. Note that it does not imply that the property itself is transported; it would seem rather difficult to transport a house that was bought, or the land on which it is built. The second class is that of execution of services which may be applied to entities. Examples would be the painting of a house, finding certain information etcetera.

Transactions on the information market typically fit in the second class. Even in the case where one downloads a file for a certain price. It would, perhaps, appear that this is a transfer of ownership but what is really transferred is the right (service!) to make a copy. Last but not least one should observe that transactions on the information market have a time aspect. To understand what we mean by this, recall that a transaction is a specific, identifiable exchange. This means that a transaction is not completed before a resource is both published and downloaded and there may be a (relatively) large gap between the moment a resource is published and the moment it is downloaded / consumed. By contrast, transactions on a 'normal' market are considered to be instantaneous.

A *transactor* view on these transactions takes the player as starting point; a transactor can thus be seen as a model construct in which one player exchanges one asset for another. A *transactand* view on transactions takes the assets as the starting point. Thus, a transactand can be seen as a model construct describing how assets flow from one player to another. For our purposes the transactor point of view is most interesting because it allows us to study the valuing of resources from a searcher point of view.

The characterization of resources is highly complex. The first important realization in this respect is the fact that value is highly personal. In other words, value can only be considered with respect to a specific searcher. Secondly, value of an asset (to a player) can only be expressed in comparison to other assets. An example from the 'phyisical markets' clarifies this: the value of a bouquet of roses (an asset) can be compared to the value of the money one has to pay in order to obtain this bouquet. Similarly, the value of a document can be compared to the value of the time that one has to invest in order to locate and access it. Lastly, valuation goes beyond figuring out whether a document is about the right topic, as specified by the searcher. In our model the valuation should be based on *informational aspects*, *structural aspects* and *emotional aspects*. Most modern search tools (such as *Google*) do a pretty good job at the former but the latter two are usually not taken into account. An example information need could be:

> A searcher is looking for a document about (informational aspects) the pollution of rivers in Australia. The document must be (structural aspects) a lengthy, detailed report, preferably in the *Pdf* format. Last but not least, a highly complex and document with many statistics and calculations is preferred since the prospective reader is highly motivated to study (emotional aspects).

In our opinion the emotional aspects are difficult to work with in practice. However, it is an interesting topic of study. On the short term practical tools may benefit most from taking structural aspects into account. To this end we have developed a model for information supply.

3 Information supply

Our goal in developing a conceptual model for information supply is twofold. Firstly we want to gain a deeper understanding of the resources on the Web, their relations and so on. Secondly, and perhaps more importantly, we want to use the model as the basis for determining the structural value of resources with respect to an information need of a searcher in a given situation.

The main drive, thus, is at the conceptual level, which partly explains why we use a set theoretic approach for our models rather than RDF, RDFS, or OWL (a good introduction to these semantic web technologies can be found in e.g., [22]). Furthermore, we want to start from a clean slate, rather than "getting stuck" with the assumptions made by others. Despite all this, there are some striking similarities with our model. Hence, we do acknowledge that these technologies can definitely play an important role when building real tools such as aptness-based search engines! We will return to this discussion later.

Similar to the RDF approach we make a distinction between resources on the Web and values that are associated to these resources by means of attributions. Data resources are the entities on the Web which make up information supply. We presume that data resources are always *about* something. These 'somethinges'

are dubbed information resources in our model. Note that we do not state how real-life application should deal with this aboutness as there are many ways to deal with aboutness. As we have discussed previously, several keyword-based approaches exist such as a Boolean model (a keyword occurs in the text or not), term-frequency based models and so on. A good overview is presented in e.g., [23]. The choice for a specific approach depends on specific applications and situations.

We adopt the point of view that data resources implement the ideas associated to information resources. Even more, they may do so in different ways. For example, one data resource may be a *picture of* the Mona Lisa, whereas another may be a *textual description* of this famous painting. The observation that ideas can be represented in different ways can be quite important for valuation of resources in a retrieval setting. For example, in certain situations one may be interested in a highly detailed, complex, technical report whereas in another situation one may be interested in a management summary. To some extent one could argue that this fits in the realm of the emotional value domain as previously explained.

Also, we presume that data resources are typed. Typical examples are *Pdf* files, *Html* files, online databases, E-services, and potentially even humans! With regard to this typing we adopt a "types follows instances" approach, which contrasts with a population follows types approach that is normally found in the realm of (relational) database design. The typing of data resources may seem insignificant given the computing power that is available to us and the 'standardization' of the last few years. It indeed seems to be the case that certain file formats are dominant but one simply can not assume that everyone has the proper tools to view resources of a certain type. Issues that play a role in this respect are: the cost of software, different versions of software, incompatibility issues, file sizes, and so on.

An advantage of a formal model for information supply is the availability of formal query / constraint languages such as RIDL or LISA-D (See e.g., [24, 25, 26]). This allows us to state (the structural and informational aspects of) an information need formally as:

```
Data Resource being a Representation
  (of type "textual description" AND-ALSO about "Mona Lisa")
AND-ALSO having Attribute (of type "Language" AND-ALSO with value "Dutch")
```

The biggest advantage of this style of formulation is that it is both a semi natural language (also called *restricted language*) while being formal at the same time. Algorithms for computing the population (of the schema) that adhers to these queries are readily available. Most approaches, however, are 'binary' in the sense that for a particular instance either holds or not; gradations are not possible. We are working on an approach where it is possible to take these gradations into account. To this end we will use the concept of a linguistic variable as introduced by Zadeh (see e.g., [27, 28]). We will get back to this discussion in Section 6.

4 Transformations

A richer and broader way of expressing ones information need is a first step towards aptness based retrieval. Obviously, such rich queries only make sense if the broker knows how to deal with them. In our opinion the key to achieving this lies in a transformation framework. Transformations provide us with the opportunity to maninupate resources such that their aptness is increased for specific searchers with a specific information need. In this section we will firstly outline our transformation framework. Then, in the next section, we will explain how this fits in a retrieval architecture.

Transformations are, in essence, pieces of software that transform resources from one type into resources of another (possibly the same) type. Transformations can, thus, be used to change (the values of) properties of resources. Typical example would be the conversion of *Doc* files to *Pdf* or changing the resolution of a picture. In the former example the type-property of a resource is changed, whereas in the latter example the value of the resolution-property is changed. Note that the *aboutness* of a resource is never changed; it seems impossible that a document about dogs is suddenly about cats after it is transformed!

If our transformation framework is to be deployed by an actual system in a concrete setting then a lot of information about transformations must either be gathered or learned by this system. First and foremost the input type and output type of a given transformation must be known. This may seem obvious but if we do not know that transformation T transforms *Doc* files *Pdf* then the transformation is essentially useless as we would never know when we could use it. The situation is similar to having a tool in ones toolbox of which one doesn't know what it is for.

Transformations can be combined to form complex transformations. The idea is simply to construct a labelled and directed graph where the nodes are types and the edges are possible transformations such that the output type of one transformation becomes the input type of another. For example, if we can transform *Doc* files to *Html* and we can transform *Html* files to *Pdf* then we can create a transformation with which we can transform *Doc* files to *Pdf*. Similarly, one may combine a transformation that is essentially an abstract generator for *Doc* files with a transformation from *Doc* to *Pdf*.

It may be the case that there are different transformation paths from one type to another. Even more, each of these paths may have different effects on the properties of resources that may be transformed. This puts to the fore the problem of selecting a transformation: which transformation is "best"? At first sight it seems that a *shortest path algorithm* may help out. This indeed makes sense, but only if the possible effects on properties are taken into account. Therefore, the second piece of knowledge that we must gather is the possible effects of transformations on properties of resources. In our view, properties can be formulated in terms of the above mentioned language for information supply. An example would be:

Data resource being source of a relation having type "hyperlink"

which expresses the property that data resources have outgoing hyperlinks. Since we consider transformations as black boxes (that is, we do not look under the hood to examine the actual code of the software), the only way to learn the effects of transformations is to actually apply them and to observe what happens. This is exactly what we meant when we stated that a lot of information must be learned about transformations by systems that deploy our transformation framework. Learning is achieved by adopting a 'learning by doing' approach. By actually executing transformations on a specific data resource we may, for example, observe that all hyperlinks (the property that we formulated previously) are removed. This observation can then be generalized to the typing level which would result in a rule such as: *this transformation always removes all outgoing hyperlinks*. Observe that it may be the case that we find contradicting evidence. For example, in one case all outgoing hyperlinks are removed and in another case they are not removed. In that case we can only conclude at the typing level that the transformation *may* remove hyperlinks.

5 Transformations & retrieval

We will now shift our attention to the deployment of a transformation framework in a retrieval setting, which will be the basis for aptness based retrieval in a practicall setting. In our opinion, there are essentially two ways to consider this. The first option is rather theoretical in nature and is based on the idea of an extensional database versus an intensional database as introduced in e.g., [29]. The idea is that the data resources available on the Web form an extensional database. Given this extensional database and a set of transformations, the intensional database is defined to be anything that can be generated from the extensional database using these transformations. Note that this intensional database is potentially infinitely large! Either way, in an ideal world one would want to be able to query this intensional database while using the concepts of our model for information supply. It is indeed likely that searchers do not care whether a document is available somewhere on the disk of some server, or if that same document is generated by means of a (possibly complex) transformation.

The second option is more practical in nature. The main idea is to adopt a "push-down selection" approach[2] where one firstly selects the (topically) relevant data resources and then tries to increase their aptness. With the input types and output types of transformations one essentially has a labelled and directed graph. At this level the transformation selection algorithm has to select a path through this graph based on a query (i.e. to select a path from node a to node b where a and b are specified in the query). With the additional knowledge of the effects of transformations this task becomes more complex as we should now select the "best" path which is the path that will most likely result in a resource that matches the query. In other words, the input of the transformation selection algorithm is a query (consisting of the desired properties), the transformation

[2] In query optimization for relational databases on usually performs *select* statements before performing expensive *joins*. See e.g., [29]

graph and the list of effecects that a transformation may have. The output is a single transformation path which can be applied to a resource to make it more apt.

Recall that search engines can be seen as *brokers* on the information market. As such they must be value adding. Regardless of how these brokers implement a transformation framework, it is easy to see that they are, indeed, value adding by saving us the trouble to manually perform transformation operations to get exactly the results that we want. It is now, finally, time to zoom in on the concept of aptness, and aptness based retrieval in specific.

6 Towards aptness based search on the Web

We started this paper by observing that the topic-based search mechanisms on the Web are, in our opinion, somewhat limited. We argued that this is mainly due to the fact that the valuation mechanism is too limited when only topicality is used. We then presented our views on the information market. The main result from this exploration was the complex value notion which was based on the three dimensions *information*, *structure*, and *emotion*. This complex value notion can be said to be the basis for aptness based retrieval as we will show shortly.

Recall that the concept of value is highly personal, that the value of some asset can only be expressed in terms of other assets and that there are many views on what value means. As such it can be stated that this value notion can have different manifestations. In a retrieval setting one could state that the main goal of search engines, which in essense perform a brokering role on the information market, is to assess the value of resources to searchers with an information need. The value would then be a metric for how *apt* the resource is for this specific searcher. This may, however, be tricky since value of assets can only be expressed in its comparison to others! We therefore propose to use the notion of *quality* (which entails a specific view on value, see e.g., [19]) as a metric for aptness.

The notion of quality is also used in many different contexts such as philosophy, e-commerce, operations management, software engineering, data quality, library information systems and so on. An extensive survey of the literature (e.g., [30, 31, 32, 33, 34, 35]) has shown that there are two main views on quality. The first view encompasses the qualities of artifacts; the properties that artifacts have. To some extent this can be objectively measured. The second view has to do with how good something is. As such it is personal in nature, similar to the value notion. Obviously, the quality in this sense depends on the qualities (properties) that an artifact has, and how desirable these properties are. In other words, if we can measure the properties that an artifact has then we may be able to derive, or estimate its quality for a person!

Again, an example from the world of physical artifacts best illustrates our intensions. Consider the situation in which a person has to assess the quality of a mug. Firstly, he has to decide which properties will be used as the basis for

the quality assessment. Let's assume that only the *volume* is used. Secondly, he has to come up with a decision rule. For example:

I consider the mug to be of high quality if its volume is bigger than 20cc.

Finally, the actual contents of the mug is measured and a quality assessment can be made. From this example we can derive that properties have associated values and these values are expressed in a domain. In the above example, the property *volume* is expressed in the domain *cc* and for the mug under consideration we measure a value of 20. In practice, decision rules are usually not as concrete as in the above example. They tend to be more fuzzy in nature. For example:

I consider the mug to be of high quality if its volume is high.

The question is what 'high' means in this context. It certainly isn't a value in the domain of cc's. We now, briefly, enter the realm of *fuzzy logic* and *linguistic variables* [27, 28]. Simply put, this works by adding a "translation layer" between the assessment rules as expressed by humans and the measurements with respect to artifacts. This is done by introducting a linguistic variable which may have terms such as 'high' and 'low' as its values. This linguistic variable has an underlying concrete domain. Furthermore, for each of the terms that can be assigned to this linguistic variable we have to define a membershipfunction which expresses the degree membership of an observed / measured value from the concrete domain to the semantic class as defined by the linguistic term.

Continuing the above example, "high" is in the termset for the linguistic variable *volume*. The underlying concrete domain is that of cc's. Assume we measure that the mug has a volume of 20cc and that the membership degree of 20cc for the linguistic term "high" is 0.8. In this case we can conclude that we're 80% sure that someone will assess the observed 20cc to be a high volume for a mug[3]. Note that it may not be trivial to determine the membership functions in practice.

We now shift our attention back to assessing the quality of resources on the Web. More specifically, we focus on the task of assessing the quality of an asset to a searcher who has expressed his information need in terms of a query. As such the inputs for the quality assessment process are: the asset itself and the user query which can be seen as a list of properties that are used to assess the quality of an asset. In the previous section we already outlined a language for information supply which is well suited for expressing these properties. The example query as expressed in this language was:

Data Resource being a Representation
  (of type "textual description" AND-ALSO about "Mona Lisa")
AND-ALSO having Attribute
  (of type "Language" AND-ALSO with value "Dutch")

[3] Strictly speaking this line of reasoning is not 100% correct since a membership degree only translates to a probability under certain conditions. In a more thorough and mathematical treatment of these measurements we have to take this into account.

This query has three properties; an aboutness property, a property that asserts what kind of representation type is desired and finally a property that asserts which language the resource must be in. Another example property could be expressed by the linguistic variable named importance with termset low, medium, and high. The underlying value domain could be *Google*'s pagerank, assuming that this is a good metric for relative importance of resources on the Web [36, 37].

A query thus contains the criteria with which one can assess the quality (and thus the aptness) of data resources on the Web. By using linguistic variables (with concrete underlying value domains and proper membership functions) we have essentially developed a metric with which we can quantify quality.

7 Putting it to practice

It is now time to look in more detail at a possible application of aptness based retrieval. The setting is a digital library with scientific papers. For example, in our research group we have collected a database of (at the time of writing) 3518 scientific papers. For each of these papers we have the bibliographic data, possibly an abstract and the actual paper itself. Finding a paper in such collection is often tricky, to say the least. It seems apparent that a search system (a broker that adds value) can help. Obviously, we are not the first to tackle this problem[4]. It does, however, provide us with a good setting to exemplify the theory as introduced so far.

Figure 1 shows that the characterization of resources is done in terms of our metamodel. The data resources, in this case, are scientific papers. Given the nature of these data resources it seems logical that the emotional dimension of valuing can safely be neglected. The information dimension of valuing can be implemented in several ways. For the time being we have chosen for a rather simple keyword based approach. Users can enter keywords as part of their information need, and documents are considered to be topically relevant if the keywords occur in their characterization.

We can now use our meta-model to define an application specific language for searching. In this case the language mainly pertains to the structural dimension of value. To this end we must define the representation types, data resource types, attribute types, and relation types. We will discuss each of these in turn before presenting an example session with our system.

We chose the set of representation types to be *full text*, *abstract*, *meta-data*, and *keywords*. Furthermore, the relevant data resource types that we chose are *Ascii*, *Html*, *Pdf*, LaTeX, *Postscript*. This means that we could ask the search system to give us an *abstract* of a paper that has certain keywords, and present it to us in the *Pdf* format. We now turn our attention to the selection of possible attribute types that we take into account in our application. Since we have the meta-data of publications readily available we can easily include those in our search. This means that we include attribute types such as *author*, *publication*

[4] See for example the digital library of ACM at `http://portal.acm.org/dl.cfm` or Citeseer at `http://citeseer.ist.psu.edu/`.

type (is it a "conference publication", a "technical report", or a "journal publication"?), *year of publication*. More specifically, we used the fields from *Bibtex* to guide us in the selection of these attributes. With these it becomes possible to, for example, search for all papers authored in a certain year. Last but not least we include relations between papers in our search system. In case of scientific papers this is given shape by scientific references. This means that we add the relation types *cites* and *cited by* to our language which allows us to search for a paper that cites *a* and is cited by *b*. With the application specific language that we have introduced so far we can, thus, combine content search with meta-data search using a uniform query language. It allows searchers to accurately specify their information need. If the set of concepts that we have introduced so far is too limited then we can always extend the language without having to re-design the entire application.

It is now time to shift the focus from query formulation and characterization of scientific papers to the actual aptness assessment process, which constitutes the middle part of Figure 1. We have previously explained that roughly two approaches can be taken: the extensional/intensional database approach versus the push-down selection approach. In case of our digital library we have adopted the latter approach. This implies that we firstly select the (topically) relevant papers and then try to increase their aptness by means of transformations. To this end we use transformations that convert from one data resource type to another (for example: `pdflatex` which transforms from LaTeX to *Pdf*) as well as an abstract generator and a keyword list generator[5].

A typical session with our system could go as follows. After logging into the system a searcher formulates the following query:

> Data Resource of type "Pdf" AND-ALSO
> being a Representation (of type "abstract" AND-ALSO about "OWL")
> AND-ALSO having Attribute (of type "author" having value "John Doe")
> AND-ALSO having Relation (of type "cites" having value "RDF.pdf")

which supposedly selects an abstract (in *Pdf* format) of a paper about OWL that is written by a *John Doe* and that also cites a paper with the name *Rdf.pdf*. The system parses this query and firstly selects all papers of the proper topic. It then performs a further selection by removing all papers that were not of the correct author and that do not cite the proper document. Finally, the system will check to see if an abstract of this paper is readily available or not. In the latter case it will generate one. Last but not least, it will make sure that the paper is presented in the proper format. As such, the value mechanism in our system can considered to be binary. If a paper conforms to all properties that are specified in the query then it will end up in the list of resources; if (one or more of the) properties are not matched then the paper is no longer considered and

[5] The latter two transformations are considered to be plug-ins for our application. If we find 'better' versions of these applications then they can easily be plugged into the system.

will, thus, not be listed. Basically this means that we have not yet implemented the fuzzy quality assessment as we have explained previously.

8 Conclusions

In summary, we have argued that it is time to reconsider the way we think about search on the Web and move to a situation that we have dubbed *aptness based retrieval*. The basis for such an approach lies in the valuation of the resources under consideration. In practice this means that, if possible, all aspects of an information need should be taken into account during the search process. We have approached the search process from an economic point of view which results in two important conclusions. First of all, we propose to use three dimensions for valuing on the information market: informational value, structural value and emotional value. These three dimensions form the basis for assessing how apt resources are with respect to an information need. The second conclusion in this respect is that search engines on the Web can be seen as value adding brokers. In our opinion this value addition can be achieved by using transformations, as exemplified by the example presented in the previous section.

Our initial experiences with searching are promising; in case of the digital library for scientific papers we have seen that it indeed seems possible to move towards aptness based retrieval. A lot of work remains to be done in this area, though. As we already stated, we have to extend our system with fuzzy quality assessment. Furthermore, we also intend to test our ideas in a less "controlled" situation. The (meta-data of) scientific papers is indeed fairly structured and thus provides us with a good environment to test the general idea. The logical next step is to try similar techniques on, say, the intranet of our faculty or the resources available in a large enterprise.

There are also, still, some theoretical issues that remain interesting. First of all, we intend to further explore the possibility of using RDF annotations. Even more, it seems interesting to explore the relation between our approach and the user modelling community. For example, it would be interesting to *learn* from the behavior of individual searchers and to use this knowledge in the search process. If a searcher appears to have a strong preference for *Pdf* then we could include this in our search. Similarly, we could also use knowledge about the searcher for such things as word sense disambiguation. In short, aptness based retrieval is indeed achievable but some interesting challenges remain.

References

1. Sahami, M., Yusufali, S., Baldonado, M.Q.: Sonia: a service for organizing networked information autonomously. In Witten, I., Akscyn, R., Shipman, F.M., eds.: Proceedings of DL-98, 3rd ACM Conference on Digital Libraries, Pittsburgh, Pennsylvania, USA, New York, New York, USA, ACM (1998) 200–209. ISBN 0897919653
2. Day, M.: Resource discovery, interoperability and design preservation. some aspects of current metadata research and development. VINE (2000) 35–48.

3. Papazoglou, M., Proper, H.E., Yang, J.: Landscaping the information space of large multi-database networks. Data & Knowledge Engineering **36** (2001) 251–281.
4. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York, New York, USA (1983).
5. Rijsbergen, C.v.: Information Retrieval. Butterworths, London, United Kingdom, EU (1975). ISBN 0408709294
6. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Reading, Massachusetts, USA (1999). ISBN 020139829X
7. Bruza, P.: Hyperindices: A novel aid for searching in hypermedia. In Rizk, A., Streitz, N., Andre, J., eds.: Hypertext: Concepts, Systems and Applications; Proceedings of the European Conference on Hypertext - ECHT 90. Number 5 in Cambridge Series on Electronic Publishing, Paris, France, EU, Cambridge University Press, Cambridge, United Kingdom, EU (1990) 109–122. ISBN 0521405173
8. Bruza, P., Weide, T.v.d.: Two level hypermedia - an improved architecture for hypertext. In Tjoa, A., Wagner, R., eds.: Proceedings of the Data Base and Expert System Applications Conference (DEXA 90), Vienna, Austria, EU, Berlin, Germany, EU, Berlin, Germany, EU, Springer (1990) 76–83. ISBN 3211822348
9. Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: Proceedings of the 23rd annual international ACM SIGIR conference on Researchand development in information retrieval. (2000) 288–295. ISBN 1581132263
10. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin Core Metadata for Resource Discovery. Technical report, Internet Engineering Task Force (IETF) (1998) Last checked: 13-Sept-2005.
`http://www.ietf.org/rfc/rfc2413.txt`
11. Brasethvik, T.: A semantic modeling approach to metadata. Internet Research **8** (1998) 377–386.
12. Gils, B.v., Proper, H.E., Bommel, P.v.: A conceptual model of information supply. Data & Knowledge Engineering **51** (2004) 189–222.
13. Bommel, P.v., Gils, B.v., Proper, H.E., Vliet, M.v., Weide, T.v.d.: The information market: Its basic concepts and its challenges. In Ngu, A., Kitsuregawa, M., Neuhold, E., Chung, J.Y., Sheng, Q., eds.: Web Information Systems Engineering (WISE), New York, New York, USA. Volume 3806 of Lecture Notes in Computer Science., Berlin, Germany, EU, Springer-Verlag (2005) 577–583. ISBN 3540300171
14. Bommel, P.v., Gils, B.v., Proper, H.E., Schabell, E., Vliet, M.v., Weide, T.v.d.: Towards an information market paradigm. In Belo, O., Eder, J., Pastor, O., Falcao e Cunha, J., eds.: Forum proceedings of the 17th Conference on Advanced Information Systems Engineering, Porto, Portugal, EU, FEUP (2005) 27–32. ISBN 9727520782
15. Gils, B.v., Proper, H.E., Bommel, P.v., Weide, T.v.d.: Transformations in information supply. In Grundspenkis, J., Kirikova, M., eds.: Proceedings of the Workshop on Web Information Systems Modelling (WISM'04), held in conjunctiun with the 16th Conference on Advanced Information Systems Engineering, Riga, Latvia, EU. Volume 3. (2004) 60–78. ISBN 9984976718
16. Gils, B.v., Proper, H.E., Bommel, P.v., Weide, T.v.d.: Typing and transformational effects in complex information supply. Technical Report ICIS–R05018, Radboud University Nijmegen, Institute for Computing and Information Sciences (2005).
17. Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G., Cho, J.: Beyond document similarity: understanding value-based search and browsing technologies. SIGMOD Rec. **29** (2000) 80–92. ISSN 01635808

18. Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., Scholl, M.: Rql: a declarative query language for rdf. In: WWW '02: Proceedings of the 11th international conference on World Wide Web (Honolulu, Hawaii, USA), New York, NY, USA, ACM Press (2002) 592–603. ISBN 1581134495
19. Holbrook, M.b., ed.: Consumer value, a framework for analysis and research. Routledge, 270 Madison Ave, New York, NY 10016, USA (1999). ISBN 0415191939
20. Varian, H.R.: Intermediate Microeconomics, a modern approach. 4th edn. Norton, New York, New York, USA (1996). ISBN 0393968421
21. Shannon, C., Varian, H.: Information Rules, a strategic guide to the network economy. Harvard Business School Press, Boston, Massachusetts, USA (1999). ISBN 097584863X
22. Antoniou, G., Harmelen, F.v.: A Semantic Web Primer. The MIT Press, Cambridge, Massachusetts 02142, USA (2004). ISBN 9780262012102
23. Paijmans, J.: Explorations in the Document Vector Model of Information Retrieval. PhD thesis, Tilburg University, Tilburg, The Netherlands, EU (1999). ISBN 9036100240
24. Halpin, T.: Information Modeling and Relational Databases, From Conceptual Analysis to Logical Design. Morgan Kaufmann, San Mateo, California, USA (2001). ISBN 1558606726
25. Hofstede, A.t., Proper, H.E., Weide, T.v.d.: Formal definition of a conceptual language for the description and manipulation of information models. Information Systems **18** (1993) 489–523.
26. Meersman, R.: The RIDL Conceptual Language. Technical report, International Centre for Information Analysis Services, Control Data Belgium, Inc., Brussels, Belgium, EU (1982).
27. Zadeh, L.: The concept of a linguistic variable and its application to approximate reasoning – i. Information Science **8** (1975) 199–249.
28. Zadeh, L.: From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions. International Journal of Applied Mathematics and Computer Science **12** (2002) 307–324.
29. Ullman, J.: Principles of Database and Knowledge–base Systems. Volume I. Computer Science Press, Rockville, Maryland, USA (1989). ISBN 0716781581
30. Lala, V., Arnold, A., Sutten, S., Guan, L.: The impact of relative information quality of e-commerce assurance seals on internet purchasing behavior. International Journal of Accounting Information Systems **3** (2002) 237–253.
31. Harrison, M.: Principles of operations management. Pitman, London, United Kingdom, EU (1996). ISBN 0273614509
32. Davis, G., Olson, M.: Management Information Systems: Conceptual Foundations, Structure and Development. McGraw–Hill, New York, New York, USA (1985).
33. Gertz, M., Özsu, M.T., Saake, G., Sattler, K.U.: Report on the dagstuhl seminar: data quality on the web. SIGMOD Rec. **33** (2004) 127–132.
34. DESIRE: Quality Selection Criteria for Subject Gateways. (2005) Last checked: 27-Oct-2005.
`http://www.sosig.ac.uk/desire/qindex.html`
35. Hernon, P., Calvert, P.: E-service quality in libraries: Exploring its features and dimensions. Library & Information Science Research **27** (2005) 377–404.
36. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117.
37. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998).