Towards a General Theory for Information Supply

B. van Gils
University of Nijmegen
Sub-faculty of Informatics
Toernooiveld 1
6525 ED Nijmegen
The Netherlands
bas.vangils@cs.kun.nl

H.A. Proper
University of Nijmegen
Sub-faculty of Informatics
Toernooiveld 1
6525 ED Nijmegen
The Netherlands
E.Proper@acm.org

P. van Bommel
University of Nijmegen
Sub-faculty of Informatics
Toernooiveld 1
6525 ED Nijmegen
The Netherlands
PvB@cs.kun.nl

June 23, 2004

Abstract

The Web has grown considerably over the last few years, both in size and in nature. The usage of the internet as a source of information has grown considerably as well. However, finding the *right* information is not always straight forward. For example, how to go about finding out how many concerts the band *Golden Earring* gave in the year 1987, or finding out which webpages are inspired by Terry Pratchett's *Diskworld* novels? Even though much effort has been invested in this area, we feel that a solid, conceptual framework for *information supply* is still missing. In this paper we broadly define such a framework, and explain how it can be used in e.g. Web Retrieval by means of an example. We also explain which areas need more research and what can be expected in the (near) future.

PUBLISHED AS:

B. van Gils, H.A. Proper, and P. van Bommel. Towards a general theory for information supply. In C. Stephanidis, editor, *Proceedings of the 10th International Conference on Human-Computer Interaction*, pages 720–724, Crete, Greece, EU, 2003. ISBN 0805849300

1 Introduction

How do search engines deal with questions like: "Which books are based on the bible?", "What is the total number of concerts given by the band *Golden Earring* in the year 1987?" etcetera. Assuming that the query "books based on bible" is a good representation of the former question, the search engine GOOGLE returns about 856,000 hits[1]; the first of which points to the online bookshop `http://www.amazon.com`. Apparently, GOOGLE does not seem to know how to deal with our query. We feel that this is because the essence of *information supply* –which can loosely be defined as the total amount of 'information' available to us– has not yet been studied well enough (see also Secion 2.1).

[1] On December 16th, 2002

The world of information supply can be seen as a supply chain with three important phases: information supply, demand for information and brokering between demand and supply (HP99). In this article we primarily zoom in on the information supply and, to a certain extent, ignore the other two phases.

Much research has been conducted in the area of information modeling, representation and retrieval already. Fields of research include *library applications* (e.g. Z39.50), *relational databases* (e.g. (Ull89; MM97; FFLS00)), *meta data activities* (e.g. (WKLW98)), *relevance ranking* (e.g. (BP98)), *markup languages* (e.g. (ISO86; BPSMM00)) and finally *conceptual information modeling* (e.g. (Che76; Hal01; BRJ99; HPW93)).

Despite the effort that has been put in, most pre-existing viewpoints on information supply are either implementation oriented (in the sense that they do not abstract from underlying technological considerations) or are rather exclusively geared towards a specific class of information supply (such as structured information, textual information, the knowledge in people's heads etcetera). Although some attempts have been made to abstract from underlying technology without restriction to a specific class of information supply, the proposed models do not provide a complete solution yet.

We feel that a solid, *conceptual* framework (i.e. a conceptual model) should *at least* be able to represent the information landscape –a term coined in (PPY01)– as we know it today. Such a framework would cater for concise and precise statements about the information space. Fields that can benefit from such a model include information retrieval, digital libraries, data warehouses etcetera.

The following section introduces our model. The section there after explains the model by means of a detailed example. Finally, some conclusions are drawn and suggestions for future research are given in Section 4.

2 The model

This section gives an *informal* introduction of our model. In this short paper we limit ourselves to the essential elements of the formalisation.

2.1 Services and representations

An important distinction is that of data versus information: data *becomes* information as soon as it is found to be relevant to a given information need. A similar, and equally important, distinction is that of an information service and the technology that was used to store it. Technology, in this context, is used in a broad sense, it can be paper, a database, a flat file, but also the knowledge in people's heads (to be accessed using e.g. a conversation). The ability of these entities –in terms of our value-chain perspective– to provide information to some consumer, is viewed as a *service* that the entities may provide. This is why we have chosen to use the term *information service* for the entities that make up information supply. With this in mind, we view the Web as as a landscape of inter-related information providing entities, which are at the technology level.

2.2 Features

Obviously, there is a relation between information services and their underlying representations. However, there is more to this relation than meets the eye. For example, consider a document on the Web may have an abstract of the book *The color of magic*. Depending on ones perspective, this document is either "full content", or "an abstract". To be able to model these facts, we introduce the notion of *features*.

More specifically: we have chosen to view an information service as an abstract entity, which may possess several *features*. Each feature is presumed to have some concrete underlying *representation* associated to it. An example would be: an information service *Lord of the Rings*, with feature *full content as movie*, and a representation in MPEG, or an information service *Lord of the Rings*, with feature *keyword list* and a representation in ASCII, holding several keywords about the movie. Note that for some information services, a "full content" version is not available. It is highly unlikely that *everything* someone knows can be "dumped" (the term *memory dump* comes to mind) and stored.

2.3 Relations

With this machinery, we are able to model information services, their representations and the fact that there are different "views" on information services. That is not all there is to it, however, since information services (especially on the internet) have relationships with others. For example, a scientific paper may *refer to* other papers, a chapter is *part of* a book and a movie is *based on* a script/book. In case of the Web, these relations are usually *implemented* using hyperlinks; a mechanism which dates back to the notion of hypertext (Con87; Bus45).

We view these relations to be binary, with a unique *source* and *destination*. For example, a situation where a scientific paper A refers to another paper B. In this case, A is the source, B is the destination and "refers to" is the relation.

2.4 Typing

On closer examination, some observations can be made with regards to the features of information services and the relations between information services. Different information services may have similar *features*, yet with differing content. For example: "title", "authors", "full-content", "full-service", "description", etc. In other words, features may be classified in terms of a *feature type*. Furthermore, features may be represented according to different formats, referred to as *representation types*. For instance: "XML", "LaTeX" or "PDF", which results in the fact that these representation types can be thought of as MIME-TYPES, a standard developed to –among other things– facilitate the uniform recognition and handling of different media types across applications[2]. The representation types are more complex than they seem at first sight. For example, an ASCII file is quite different from a POSTSCRIPT file, which again are very different from a structured database, or a JAVA application. The way we will approach this diversity and heterogeneity is to treat representation types as

[2] MIME is an acronym for *Multipurpose Internet Mail Extensions* and is defined in (BF92)

abstract data types, which are represented as many-sorted algebra's providing abstractions of the underlying "implementation details" (BW90).

Finally, different semantic classes of relations between Information Services exist, for example: "refers to", "part of" and "based on", etc. In other words, relations may be classified in terms of a *relation type*. In other words, a *typing mechanism* must be introduced. Using such a mechanism would allow us to make very precise statements about (groups of) elements in the information space. The usual merits of introducing such a mechanism also apply.

When refining the situation with a typing mechanism for features, representations and relations, the ER model as shown in Figure 1 results. Note that information services do not receive an explicit type! A typing mechanism for information service, is really yet another *feature type*.

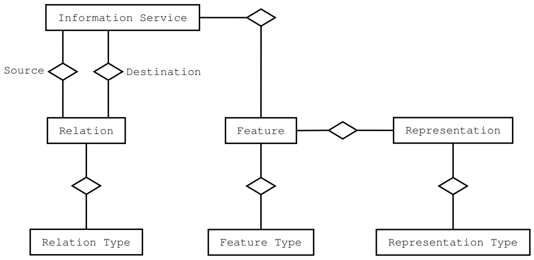

Figure 1: Typing of Relations, Features and Representations

With the typing mechanism in place, the model is complete. We feel that the model is expressive enough to be able to represent the information landscape as we know it today. In the following section a detailed example is presented.

2.5 Formalization

We are currently working on a formalization of the above model, based on descriptive mathematics. In our model we distinguish between information services, their relationships, features and representations that can be discerned are presumed to be contained in the sets: $\mathcal{IS}$, $\mathcal{RL}$, $\mathcal{FE}$, $\mathcal{RP}$ respectively.

Information service may have several features associated to it. However, each features has exactly one information service associated to it. This is modelled by the function

$$\mathsf{Service} : \mathcal{FE} \rightarrow \mathcal{IS}$$

Furthermore, the representations that may be associated to a feature are modelled by the function

$$\mathsf{Representation} : \mathcal{FE} \rightarrow \mathcal{RP}$$

The sources and destinations of the inter-service relationships in $\mathcal{RL}$ are presumed to be provided by the functions $\mathsf{Src}, \mathsf{Dst} : \mathcal{RL} \rightarrow \mathcal{IS}$ respectively. The typing mechanism is implemented in a similar way, using functions to associate descriptive elements to their underlying types.

$$\mathcal{DE} \quad \triangleq \quad \mathcal{RL} \cup \mathcal{FE} \cup \mathcal{RP}$$

Furthermore, we have introduced several Axiom's such as the fact that all instances in information supply are typed, and that all "types" that we recognize must actually be used:

Axiom 1 (Total typing) $\forall_{e \in \mathcal{DE}} \exists_t [e \ \mathsf{HasType}\ t]$

Axiom 2 (Total type usage) $\forall_{t \in \mathcal{TP}} \exists_e [e \ \mathsf{HasType}\ t]$

With these Axioms we will later be able to prove other properties of our model.

3 Example

This section has a small example, that illustrates introduced model so far. Assume there are two information services: J.R.R. Tolkien's *Lord of the Rings*, and Terry Pratchett's *Small Gods*; both novels. The former has the following features and representations associated to it:

- feature *full content in text* and representation `lotr.txt`
- feature *keyword list* and representation `lotr.kwl`. Note that keyword list says nothing on the *file type*. A keyword list can be in represented in any textual format (e.g. ASCII, PDF, PS)
- feature *full content as movie* and representation `lotr.mpg`. This movie *is based on* the book!
- feature *audio only* and representation `lotr.mp3`. Note that it is possible to extract that audio from the movie, and store this separately.

Similarly, the latter has the following features and representations associated to it:

- feature *full content in text* and representation `smallgods.pdf`. In this case, the PDF file does not hold any images, just text.
- feature *full content as audio* and representation `smallgods.mp3`. The story was read by *Tony Robinson*.

Figure 2 graphically illustrates this example.

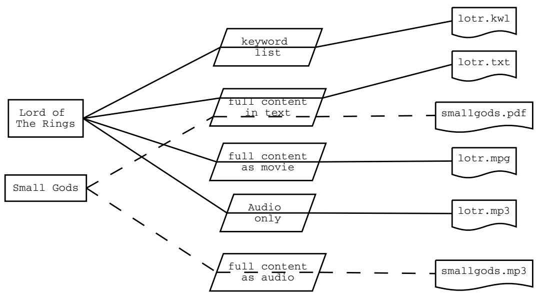

Figure 2: Illustrating the example

4 Conclusion

With the apperant rise of the Web as an important resource of information, new techniques for disclosing this information are needed. Over the last few years, several approaches have been proposed. The *nature* of these approaches vary a great deal, especially with regard to the way information was considered. For example, in the *relational model* the restriction is that data has to conform to a pre-specified conceptual schema, and in e.g. the GOOGLE approach, the assumption is that (Web)documents are linked using *hyperlinks* with *anchor text*.

We feel that a *conceptual model* –in which the information landscape as we know it today can be represented– is still missing. In this paper we (informally) introduce such a model, based on the notion that there is an important distinction between information and the technology that was used to store/represent this information. Furthermore, we recognize the fact that different *features* can be recognized in this context. Some information services can be considered as e.g. "full content", whereas others are (for example) abstracts.

Even though the model is highly expressive, much is missing still.We are currently working on a formalisation of the model using descriptive mathematics. With such formalization we are able to express our model more precisely, and cater for *proving* interesting properties of the model itself. One of the things we wish to demonstrate is the notion of *transformations* at the feature level (for example, it should be possible to transform from a feature "full content in text" to "keyword list") and at the representation level (for example, transform from HTML to ASCII).

Secondly, we are working on a constraint language, in order to be able to put some restrictions on our model. This language will be developed along the lines of LISA-D (HPW93), and is based on the notion of *path expressions*.

Last but not least, the model which we have presented must be validated. We are currently working on an implementation. With this implementation we will be able to run a series of experiments with will function as a proof of concept.

References

N. Borenstein and N. Freed. Mime: Multipurpose internet mail extensions. Technical Report RFC 1341, IETF Network Working Group, http://www.ietf.org/rfc/rfc1341.txt, June 1992.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, and Eva Maler. Extensible markup language (xml) 1.0 (second edition). Technical report, World Wide Web Consortium, October 2000.
{http://www.w3.org/TR/REC-xml}

G. Booch, J. Rumbaugh, and I. Jacobson. *The Unified Modelling Language User Guide*. Addison-Wesley, Reading, Massachusetts, USA, 1999. ISBN 0-201-57168-4

V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, Jul 1945.

K.B. Bruce and P. Wegner. An algebraic model of subtype and inheritance. In F. Bancilhon and P. Buneman, editors, *Advances in Database Programming Languages*, ACM Press, Frontier Series, pages 75–96. Addison-Wesley, Reading, Massachusetts, 1990.

P.P. Chen. The entity-relationship model: Towards a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, March 1976.

J. Conklin. Hypertext: An Introduction and Survey. *IEEE Computer*, 20(9):17–41, September 1987.

Mary F. Fernandez, Daniela Florescu, Alon Y. Levy, and Dan Suciu. Declarative specification of web sites with strudel. *VLDB Journal*, 9(1):38–55, 2000.

T.A. Halpin. *Information Modeling and Relational Databases, From Conceptual Analysis to Logical Design*. Morgan Kaufman, San Mateo, California, USA, 2001. ISBN 1-55860-672-6

S.J.B.A. Hoppenbrouwers and H.A. Proper. Knowledge discovery - de zoektocht naar verhulde en onthulde kennis. *DB/Magazine*, 10(7):21–25, November 1999. In Dutch.

A.H.M. ter Hofstede, H.A. Proper, and Th.P. van der Weide. Formal definition of a conceptual language for the description and manipulation of information models. *Information Systems*, 18(7):489–523, October 1993.

Information Processing – Text and Office Systems – Standard General MarkUp Language (SGML), 1986. ISO 8879:1986.
`http://www.iso.org`

Alberto O. Mendelzon and Tova Milo. Formal models of web queries. In *Proceedings of 16th Symp. on Principles of Database Systems – PODS 97*, pages 134–143, 1997.

M.P. Papazoglou, H.A. Proper, and J. Yang. Landscaping the information space of large multi-database networks. *Data & Knowledge Engineering*, 36(3):251–281, 2001.

J.D. Ullman. *Principles of Database and Knowledge-base Systems*, volume I. Computer Science Press, Rockville, Maryland, 1989. ISBN 0716781581

S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. Technical Report RFC 2413, Internet Engineering Task Force (IETF), http://www.ietf.org/rfc/rfc2413.txt, 1998.