ERCIM UI4ALL - Prague 7-8 Nov. 1996

Quality of Service for Information Access

Martin Prime
Rutherford Appleton Laboratory
Chilton, Didcot
Oxon OX11 0QX, UK
+44 1235 44 6555
martin@inf.rl.ac.ac.uk

Michael Wilson
. Rutherford Appleton Laboratory
Chilton, Didcot
Oxon OX11 0QX, UK
+44 1235 44 6619
M.Wilson@rl.ac.uk

ABSTRACT
Information is available in many forms from different sources, in distributed locations; access to information is supported by networks of varying performance; the cost of accessing and transporting the information varies for both the source and the transport route. Users who vary in their preferences, background knowledge required to interpret the information and motivation for accessing it, gather information to perform many different tasks. This position paper outlines some of these variations in information provision and access, and explores the impact these variations have on the user's task performance, and the possibilities they make available to adapt the user interface for the presentation of information.

Keywords
user interfaces for all, information retrieval, intelligent interface, automatic information presentation

INTRODUCTION

Question : Would you rather see the video from server X, read the text on server Y, experience the virtual world on server Z, or all three ? How do you answer such questions?

Distributed information retrieval systems can crudely be viewed as systems incorporating the following layers: Information storage, Transport Layer, Presentation, User Behaviour. At each of these layers there can be seen a variation in quality.

The user interfaces for all approach emphasises the variations in user interface for different user populations during system design. Within this approach a key element is to identify the variations in the potential system and the class of user for which each option is appropriate. This paper considers the variations in quality in information retrieval systems and tries to consider the effect these have at each of the layers of system, so that these variations can be mapped to user classes in order to guide the design and building of systems.

It is assumed that users retrieve information because they need either to a) perceive the form of the information and veridically memorise it; b) understand the propositional content of the information and maybe memorise it; c) understand the propositional information in order to make a decision; d) copy the information into an object they are creating where understanding may be irrelevant, but the accessibility to the users copy method is important. The ultimate judgement of quality of the information is its utility to serve in each of these four roles, and the judgement of the quality of the presentation of the information is its ease at perception, understanding and copying where required. The secondary task of information location introduces an efficiency measure of time to retrieve information.

Firstly the variations in quality at each of the layers are outlined, before considering the actions that can be taken to optimise the user's task performance.

INFORMATION STORAGE

The quality of information at the information storage level in respect of a users task can vary according to several parameters: Content, Form, Quantity, Recency, Accuracy, Availability as well as with the cost of accessing it. Metadata can be made available for each of these factors to inform the decision as to which source to choose. Each of these factors is presented in more detail below.

Content of Data: For any task the appropriate content is required in the data. Trivially this is obtained by matching a query to the data, and it is the form of the query and the matching process which determines the content of the data. When multiple information sources are available, queries may be issued to several sources, and the data must be integrated. Such an integration process gives rise to differences in the semantic interpretations used for each of the terms in the query by the constructors of the different information sources. Database schema and the indexes in information retrieval systems do not always use the terms in the same sense as that user's query so both inappropriate data can be returned and appropriate data

missed. Research systems attempt to overcome these problems by increasing the intensional information available to define the terms used, thereby supporting a richer process to match requests to information descriptions.

Form of Data: The form of data can be propositional, textual, tables including alphanumeric strings, or stored as graphic, video or audio media. In each case the abstraction used in the match will include the content, but may also include the form itself. The user may wish to obtain specified content, and either may or may not constrain the form of the data, which will be considered later as a presentation property of the information. Presentation will influence the usability of the information and therefore its quality in respect of task performance.

Availability of Data: Two aspects of availability are general availability and specific availability. General availability is the set of information available to the user; that is the set to which they have the required physical connections, and to which they are authorised access. The specific availability is the subset of these information sources which are currently available, others being unavailable due to communication, server or other failures. Availability governs quality since it limits the set of information sources across which any metrics of recency, accuracy etc can operate.

Quantity of Data: With 2 databases where the amounts of data in them differs, DB1: may contain the set X which contains 100 elements, whereas DB2 may contain a further 1000 elements. The user may only wish to receive 100 elements and therefore access will only be made to DB1 and not to DB2. The quantity of delivered information may be a constraint in terms of usability and task performance.

Recency of Data: A user may wish to access the IBM share price. This may be available in DB1 in the local Madrid finance house, it may also be available in DB2: the London Financial Times Database and thirdly in DB3: the New York Stock Exchange Database. DB1: may be reliable at close of business on the previous day; DB2 may be reliable to update every hour on the hour; DB3 may be reliable to the previous second. The user may only want an approximate figure for yesterday's close of business, and therefore only DB1 will be queried and only that data returned.

Accuracy of Data: Data on the exact location of stars my be available from several astronomical databases. Each database is associated with an error in the accuracy of the instruments used to collect the data. A large database (DB1) may use less accurate measuring devices (R) to collect large amounts of data, a second small database (DB2) may only contain less quantity of data but it was collected with more accurate devices (P). A user may say that they wish to access the locations of stars in a portion of the heavens, but it must be more accurate than some measure Q. The accuracy calculation shows that P>Q>R, therefore only DB2 is queried for the data, and only that subset is returned.

Reliability of data: If there are two databases which contain data that could be the information that you require, and one is known to contain more errors in its entries, which would you choose ?

Response time: If there are two databases (DB1 & DB2) which contain subsets of answers X & Y to the query, and DB1 takes 3 minutes to respond whereas DB2 takes 40 minutes to respond, then a user may only want a quick answer from DB1 (X) and not want to wait for the answer to DB2(Y). In this case the user will state a time limit on the retrieval process which will be used to select data from DB1 only, and the set X will be returned rather than the larger set X&Y. Some servers may take a long time to connect to a service so it may be worth keeping the connection open in order to achieve fast responses even when requests are not being made.

Cost of Data: Data may be available on vacancies at Hotel D from the individual hotel (DB1) or from a booking agency (DB2). Because the booking agency provides many services on the same machine the cost of using DB2 is more expensive than DB1 where the overheads are less. A travel agency trying to book a holiday for a customer does not want to spend more money than it needs, and therefore wants the cheapest way to retrieve the information. Therefore they use DB1 only and not DB2 to provide the data. Different charging models are used on different services, some charge by the session, some by the time connected, others by the amount of information retrieved, and yet others may even include the amount of processing performed on the server in the charge. The cost of data is not a measure of quality, but since it is normally the inverse of quality and a factor in the overall task performance it is mentioned here.

TRANSPORT LAYER

Networks vary in their performance and the delivery of information. Ubiquitous (Weiser, 1991) and mobile computing systems will self evidently move about with the user. Therefore they will have access to different networks at different times, each of which can provide different performance. For example, a user may connect to an office ethernet when seated at a desk; change to a WaveLAN when walking through the building; move to GSM when in a car, and move to a modem when at home.

Table 1 shows the simple bandwidth available from a range of networks which may be used. GSM is the current European mobile cellular digital telephone network available for those who connect PC's to portable phones. This provides the lowest bandwidth available to most practical applications for data transfer and phone quality voice. The most popular current modem link used provides a higher bandwidth sufficient for image transfer on a web browser. Local area (300 m radius) radio based local area networks (e.g. the WaveLAN product) provide a wider bandwidth sufficient for jerky video with frame loss when a sound track is used too. The E1 telephone line used to domestic subscribers has sufficient bandwidth to the subscriber, but the return bandwidth is very low (about 1.6Kbit/sec) so that although video on demand can be supplied through such cabling, the return signal is enough to control this, and not enough to support a video server in the home.

NETWORK	**NET**	**BANDWIDTH**
GSM	9.6	Kbit/SEC
MODEM LINK	28.8	Kbit/SEC
WAVELAN	1.2	Mbit/SEC
T1 ETHERNET LINK	1.44	Mbit/SEC
E1 TELEPHONE LINK	2	Mbit/SEC
ETHERNET	6	Mbit/SEC
FAST ETHERNET	100	Mbit/SEC
ATM	155	Mbit/SEC

Table 1: Practical bandwidth supplied by different networks

APPLICATION	**BANDWIDTH**
Phone Quality Sound - 4kHz	9Kbit/sec
3D Audio - 7kHz	48, 56, 64 Kbit/sec
Motion JPEG @10/sec	589 Kbit/sec
MPEG Video	2 Mbit/sec

Table 2: Bandwidth required by different streamed media

When users send packets with bandwidth that exceeds the capacity of the bottleneck link, the router will simply throw away or drop the surplus packets. For example, if a user tries to send 64 kbits/s PCM-encoded voice over a 14.4 kbit/s modem link, only 22% of the packets will get through, and 78% will be lost. As users move around an intelligent network may change their connections to optimize the bandwidth available for the users task thereby changing the bottleneck link. The mechanisms available to adapt the network performance to the available bandwidth are shown in Table 3. However, this network technology cannot be relied upon to always provide sufficient bandwidth for a user's task, and the application and user must themselves make adaptations at the application or user behaviour levels shown In the table.

Along with these variations in quality, network charging models vary so that some networks charge by the time connected, others include a large payment for each connection made, with lower charges for used time, some use a fixed price for the service per year or month. The calculation of best service will always trade-off the cost against the quality required.

PRESENTATION LAYER

Comprehension of information is maximised when the user has prior knowledge of the terms used and the relationships that hold between them. The relationships can be presented in various ways, as textual language, in tabular form. in various visualisations. If the information is to engage the attention and interest of the user the structure of the presentation should maximise the information salient to the users task, and minimise the effort required by the user to locate that information. Therefore a user forced to read a large amount of text which explains all the terms and relations prior to stating the required information may be required if the user does not have prior knowledge of the terms and relationships, but the effort required to reach the required information is great. If the information can be presented in a way where the user's knowledge of the world is used implicitly interpret the objects and relations then the comprehension effort can be lower and the engagement higher.

Static graphics implicitly code the spatial relationships between objects so that these do not need to be explictly presented. Representation of objects in graphics consistent with real world experience promotes the identification of those objects real world properties and relationships with the presented object. Video implicitly presents temporal relations so these do not have to be stated. Virtual reality models implicitly present the interaction properties of objects so that the relations can be explored. Analogies can be used to present non-spatio-temporal semantics as though they were spatio-temporal (e.g. different colours to represent different values at locations on a map). Specific spatial visualisations can be used to represent other relations in diagrams (e.g. business graphics, organisational hierarchies etc). Each of these manipulations of the presentation transforms a predicate relation into a spatial or temporal construct for presentation.

Architecture Level	Adaptation	Example
Transport and below	Change or introduce new protocols	New protocols can be selected which suit the characteristics of a particular network or appropriate protocols can be introduced (e.g. injecting a reliable data link layer).
	Optimise Data for the network	Protocols can adjust their packet sizes to suit different networks. The oprtating system can adjust the queue sizes onto the network interfaces which impacts on latency, particularly of multimedia streams.
	Optimisation of Multicast	Multicasts can be mapped onto the network technology, particularly those with partial or full hardware multicast support.
	Optimise for the characteristics of the network	There are a number of cost and network structure optimisations. For instance batching data to spread the dialing delays or transferring additional information while the time is already paid for.
	Reordering of Data	The priority or urgency of data may require that it is handled preferentially in scarce bandwidth situations.
	Demultiplexing to multiple networks	If multiple technologies are available simultaneously it may be advantageous to use several at once.
Middleware	On-demand Cache Management	Information can be fetched only when needed, instead of speculatively, e.g. opening the first page of a document and transferring successive pages later, or retreiving message headers before bodies (for e-mail etc.).
	Pre-fetching into the cache	The application can fetch information while the link is good, in case it is required when the link degrades or becomes expensive.
	Apply filtering and compression	The volume of information to be transferred can be reduced by compression or filtering non-essential frames from hierarchically encoded data.
	Efficient protocol utilisation of the channel	The transport mechanisms can to match channel characteristics, e.g. retransmission/back-off strategies, header compression, error control and handling of asymmetric channels.
Application	Restructure using agents or delegation of processing	Processing of network intensive tasks can be off loaded to remote sites or pre-processing or filtering applied to remote data (reducing bandwidth requirements and freeing the host for other tasks/dozing to save power).
	Use proxy services	The application can use local substitute services based upon cached information (often with reduced functionality) while disconnected.
	Change model of interaction	Interactions can be adjusted from polling to event based structures or from RPC to an alternative (perhaps asynchronous) paradigm.
	Reorganise application structure	One example of application restructuring is to change from using distributed state to a centralised architecture to simplify consistency management in unreliable conditions.
	Re-bind to new services	The application may be able to re-bind to equivalent serviceswhich are easier/cheaper to access. Alternatively, it may be possible to migrate the service or application component.
	Change Application demands	New QoS requirements can be negotiated or non-essential bindings dropped. Alternatives may be possible, e.g. lossy encodong.
	Adjust consistency requirements	Groups may be able to tolerate weaker consistency or adjust operations to achieve quorum, yet avoid hard to reach members.
User Behaviour	Change of working practices	The user can alleviate demands on the network, e.g. change task, swap from synchronous to asynchronous collaboration, or specify which tasks are most important to them.

Table 3: Adaptations at different levels of the architecture to accommodate varying network quality of service (Davies, 1996).

The transformation of predicates into spatial or temporal constructs can take place on the retrieved information at the client side to generate a multimedia visualisation, but in many cases the multimedia presentation has already been created at the server, and should be retrieved in that form for direct presentation. In the discussion of factors influencing quality at the information server, the form of the information storage was one that could be used to explicitly retrieve information in desired forms. However, it is obvious that the multimedia form of the information contains explicit structural information which is not present in other forms, and therefore requires more storage, with the consequent impact on the transport layer.

In conflict with the need to retrieve the maximally engaging form of the information sought after is the delay and cost required over that needed for the less engaging ones. System response time is the dominant component of users satisfaction with the usability of computers. Table 4 shows a classic set of results for the toleration of users for delays in response. Although it is possible to overcome user dissatisfaction by attempting to fill the delay with clear indications that the system is busy fulfilling the request such as active cursors or rotating world logos, to provide explicit feedback on progress of the task, or to provide the partial information as it is received, these are all explicit attempts to placate the user having acknowledged that the system cannot fulfill the task in the desired time.

User Reaction	**Delay**
assumed when no computation expected	0.2sec
tolerated when computation expected	2 sec
uneasy and unacceptable	< 20 sec
user gives up	> 20 sec

Table 4 : User tolerance of computer response delay to requests (from Shneiderman, 1984).

When the information conveyed by the medium is understood it may be that the medium is indeed conveying more information than was presumed, and information which is required for the efficient performance of some tasks. For example, table 5 presents the bandwidth required to present signal quality of speech required to convey different information that is normally carried when we speak face to face. At low telephone bandwidths we are able to hear the words and understand the propositions; we are normally able to identify familiar speakers, but turn taking cues conveyed by intonation changes at the end of speakers turns require higher bandwidths. For conferencing applications this class of information becomes important in managing the interaction of different speakers.

Information	**Bandwidth**
propositional content	6 Kbit/sec
speaker identification	9 Kbit/sec
turn taking intonation cues	32 kbit/sec

Table 5: Bandwidth required to convey different information in speech.

Similarly, table 6 shows a classic social psychology result describing the source of information understood by a person on first meeting another. All this information is conveyed by a video presentation, but considerably less by a written text of somebody's speech. Yet maybe the information derived about the speaker is important for the performance of the user's task, and not just the information contained in the propositional content.

Source	**% Information**
Physical Appearance	20
Body Movement & Posture	15
Facial Expression	15
Eye contact	30
Intonation	10
Word Choice & Grammar	5
Propositional Content of Speech	5

Table 6: The percentage of information from different sources understood about a person on first meeting them (recalled from memory, so the numbers are inaccurate).

If the propositional content or the text is retrieved from the information source, and a anthropomorphic character is generated to present the information, is this extraneous information being added to that retrieved ? Recent development of intelligent assistants, or personal agents do exactly this, in using a character to present information. Microsoft have developed the Peedy parrot character to present information about their CD service as an animated agent which is able to communicate both through speech and through body movements; a more human agent has been developed by Sony which has a human face which includes lip synchronised speech and facial expressions to convey information. Neither of these reach the humanity of the agent in Apple's (1992) envisionment of the Knowledge Navigator video but they are attempting to reach that goal. However, several studies have shown that

although the use of human faces appear to increase engagement, they also take more effort to interact with the system (Koda and Maes, 1996), lowers performance and lead to the user generating too high expectations of the systems abilities (Takeuchi et al, 1995; Walker et al, 1994).

The use of human faces is an extreme case in the generation of multimedia information presentations, but many intermediate visualisations exist where users generate false implicatures about the presentation. That is, they assume that parts of the display have been presented to convey some information which in fact was not intended (Reiter & Marks, 1990) because there is information in the presentation which is redundant to the message, indeed worse than this, the extra information reduces the quality of the presented information.

Let us return to the retrieval of information of different media and the adapting presentations to reduced quality information. If the user requests to retrieve some information in the form of sound, and the bandwidth is not sufficient for the transmission, there will be a reduction in quality. Initially this reduction will come from packet loss and the introduction in jitter as packets arrive out of order. There are two general actions that can be taken, to either compress or transform the information. The compression option is enacted by reducing the sampling rate in order to reduce the bandwidth required, or to reduce the packet size to reduce the effects of jitter. At some point, the transformation of the sound will reduce the quality less that the further compression, and the sequence of transformations shown in table 7 becomes possible, ultimately choosing to transmit text instead of sound as a result of speech to text conversion at the information source.

3D sound

Stereo High Bandwidth Sound

Stereo High & Low Bandwidth

Stereo Low Bandwidth

Mono Low bandwidth

Speech to Text translation

Table 7: Successive reduction in bandwidth needs of sound, with concomitant reduction in quality.

If the example chosen were video instead of sound, the same alternatives of compression and transformation exist, and the scale of transformations is shown in table 8, once again ending in the transmission of text where that is the last resort.

MPEG Video 25 frames/sec

Motion JPEG 10 frames/sec

storyboard

single image

text description

Table 8: Successive reduction in bandwidth needs of video, with concomitant reduction in quality.

In each of these cases of transformation information from the originally engaging form of information is being lost. A match between an information request and an information content description could still be satisfied and the presented information could still meet the user's need. The information which is lost is redundant to the matching process. Indeed, as shown by the false implicature example in visualisation generation, the redundent information may indeed cause users problems in their tasks, and the transformed information may be easier to understand and use. However, it is also possible that for some tasks as illustrated in table 5, task relevant information may be lost.

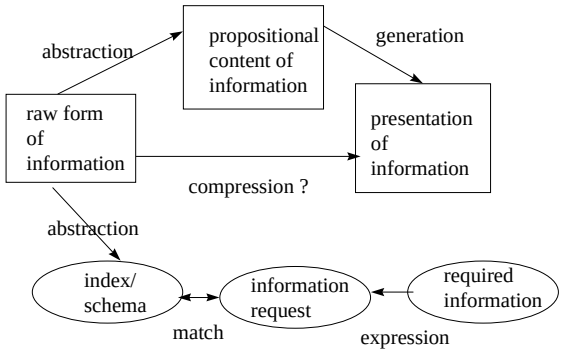

Figure 1: A representation of the information retrieval process.

CONCLUSION

This position paper has attempted superficially to analyse the information retrieval process in the spirit of UI4All by extracting the variables over which an intelligent information interface (I^3) could adapt. Effective retrieval and presentation of information requires the accurate retrieval of the required information and the presentation so that it can be understood, and copied for some tasks. Efficient retrieval and presentation requires that this process is performed at minimum time and cost. This paper has illustrated each level of analysis with examples in order to show how the factors can interact: indeed the

combinatorial explosion that results when attempting to produce effective and efficient retrieval shows some of the difficulties in the problems that I^3 research needs to address.

It is possible to rely on the information abstraction process that produces indexes/schema at a server, and present the information that is retrieved in response to an information request as present systems do if the bandwidth is available for transmission. However, the composition of information from different sources, or even video from the same source quickly shows that many forms of information include redundant information which causes user problems.

If the transport bandwidth is not available, then it is possible to compress or transform the information, hopefully only loosing redundant components, but for some tasks these may serve a significant role.

It is possible to retrieve the 'raw' information and elaborate it at presentation in order to visualise it increase engagement; but the risk of introducing false implicatures exists there depending on the excellence of the design rules used.

Each of these possibilities should be accommodated by the I3 system, within the limits of the combinatorial explosion of adaptable factors, but a large amount of research is required to investigate both the interaction of the different factors and role of possibly redundant information in the performance of real user tasks. Many of these thoughts have probably been better expressed by Card et al, (1991) and in other papers elsewhere.

REFERENCES

1. Shneiderman, 1984.
2. Weiser, The computer for the 21st century. Scientific American, 265(3) 94-104, 1991
3. Koda, T and Maes, P, Agents and Faces : The Effects of personification of Agents. Poster at HCI '96, the annual conference of the BCS HCI group, August 1996.
4. Ball et al, Lifelike Computer Characters: the Persona project at Microsoft Research. In J. Bradshaw (Ed.) Software Agents, MIT Press, 1996.
5. Walker et al, Using a Human Face in an interface, In Proceedings of CHI'94, 85-91.
6. Takeuchi, A. Et al (1995) Situated Facial Displays: Towards Social Interactions, In Proceedings of CHI '95, 450-454.
7. Reiter, E. And Marks, J. Avoiding Unwanted Conversational Implicatures in Text and Graphics. In Proceedings of AAAI '90, Boston, July 1990.
8. Clarke, Fitzpatrick and Coulson Adaptive System support for Multimedia in Mobile End-Systems. In Proceedings of the Third IEEE Communication Networks Symposium, Manchester, July 1996, 54-57.
9. Davies, N. Presentation at the First EPSRC Review of the ADAPT project, Lancaster University, October 1996.
10. Card, S.K., Robertson, G.G., Mackinlay, J.D., The Information Visualiser, an information workspace. In Proceedings of CHI '91, New Orleans, 1991, 181-188.