Integrating reading and writing of documents, with particular application to bioinformatics

A. P. Galton, A. Narayanan, M. R. Amos *

May 23, 2003

PART 1: Previous research and track record of Visiting Fellow and Department

1 Visiting Fellow

Professor Peter Brown has a track record of pioneering new ideas, and turning these into successful systems and eventually into products. The most successful of these has been the Guide system, which embodied new research ideas in hypertext, including expanding in-situ links or scripts and animated illustrations of search paths to prevent the user getting lost. Guide was turned into a product by OWL, a small company based in Edinburgh, and sold widely both to personal and corporate customers. In particular it was used by General Motors for their vehicle-bay information systems. Guide was a winner of the BCS Technical Award. In the mid-nineties Professor Brown started work in context-aware mobile applications; this work has been specifically related to the use of documents and annotations by mobile users [2], [3]. He created the technology of "stick-e notes" (essentially an annotation of a context), which is now widely used by researchers. Again this activity has produced a promising candidate for future exploitation: the system recently developed by Jason Pascoe, a research student working under Professor Brown. This system is used in fieldwork data-capture applications, and embodies new ideas in (a) stick-e notes, (b) using context and (c) user interfaces. The system is used by Earthwatch, an ecology charity.

The mobile applications work has more recently fed back into the hypertext work, and has led to new ideas in electronic documents. These have been published in an invited paper [4], and in a contribution selected as input to a recent workshop on Grand Challenges in Computer Science. The current proposal is based on this.

A brief c.v. of Professor Brown, highlighting collaborations with other research teams is as follows.

- 1968-present: (main employment) Professor of Computer Science at University of Kent at Canterbury; emeritus since 1999.
- 1980-1: Visiting Professor, Stanford University; worked on emerging ideas for user interfaces for document preparation. (Also worked on VLSI chips for string processing, research that subsequently won the Hartree Premium from the IEE.)
- 1984-92: consultant to OWL on making the Guide system a product.
- 1987: Erskine Fellow, University of Canterbury, New Zealand; worked with their database team on relationships between hypertext and databases.

*School of Engineering and Computer Science, University of Exeter

- 1992-3: Xerox EuroPARC, Cambridge; worked with staff from Cambridge and Palo Alto on human memory aids.
- 1996-7: Xerox European Research Centre; worked on novel methods for document delivery on mobile devices.
- 1999-present: Visiting Professor, Southampton University; worked on joint papers, primarily in document engineering.
- 1999-present: part-time contract work at Univ. of Exeter; self-employed author and consultant.

2 Relation to Departmental interests

Professor Brown is known to the Computer Science Department, and, indeed has been employed in it.

His proposed research programme fits well with the work of our Information Engineering subgroup. However we see an even greater benefit for the Department's bioinformatics research. The Department of Computer Science is currently at the forefront of developments in bioinformatics postgraduate teaching and research. The Department (in cooperation with the School of Biological Sciences) started an MSc in Bioinformatics in 1998 with HEFCE money for extra funded numbers based on excellence in teaching. The programme was awarded MRes status in 1999 by the EPSRC as part of its development of Master of Research training programmes and now receives approximately 100k per annum from EPSRC.

The Department has appointed two new bioinformatics lecturers in conjunction with the School of Biology (Dr Dalby and Dr Amos) and has two other researchers (Prof Narayanan and Dr Yang) active in the area. There are 6 research students currently studying or completing their PhD programmes in bioinformatics. The Department is also a member of consortia recently awarded 2.1m by the Wellcome Trust (Biocatalysis Centre) and 2.2m by the ESRC for the creation of a national centre in genomics research. Bioinformatics researchers in the Department are national and international leaders in the area of applying techniques from artificial intelligence, neural computation and evolutionary computation to bioinformatics problems. They have recently submitted tutorial proposals and review papers on the subject to *Intelligent Systems for Molecular Biology* (ISMB03, Brisbane), *Applied Bioinformatics* and the BCS-supported Pattern Analysis and Recognition Network. Bioinformatics researchers in the Department have research links with in the recently created Peninsula Medical School; special study units in gene chip technology are being made available to medical undergraduates and discussions have started on future cooperation with regard to microarray analysis applications.

The importance of Professor Brown's work to bioinformatics is that researchers can spend a huge amount of their time reading documents on the web, following various paths through them, and writing new documents based on what they have found or not found. Improved aids to reading and writing documents can therefore be a key to improving the productivity of bioinformatics researchers.

PART 2: Description of Proposed Research

1 The computer science challenge

When we read a document we often want to write as well, for example to annotate the document. When we write a document we often also want to read documents, for example we might want to be prompted about existing documents that are related to the one we are composing. Reading and writing should therefore be treated as complementary activities. Current computer science does not reflect this: we think in terms of reading tools, such as browsers, and entirely separate writing tools, such as editors or word processors. In terms of implementation, reading tools offer at best crude facilities for writing, and writing tools offer at best crude facilities for reading.

If we successfully meet the challenge of integration reading with writing, and implement corresponding software tools, we will have made an advance that benefits virtually every computer user.

2 Document model

A requirement for realising the challenge is to create new document models. The challenge implies that all components of such a model must be first class citizens. Thus anything can be written, anything can be realised electronically or on paper, anything (including an annotation) can be annotated, anything can contain hypertext links, "mark-up" can be changed or annotated by the reader in the same way as content, and everything has provenance including a record of who created it and when. The model must not only cover individual documents but collections of documents with links between them; the linked structure (which in implementation might be partly created automatically, e.g. by interlinking a collection of e-mails) should apply to all classes of document. Like any good model, the model must capture abstractions rather than realisations. This general model must be realisable in real systems that real people can use. A good starting point for new models may be the existing W3C document models.

3 Change

Two important aspects of documents are change and management of who can make changes. When a document changes and the reader decides (or is constrained) to use the new version, then previous annotations, links, etc., will not always work perfectly. Their behaviour should, however, degrade gracefully as the document changes.

Many original documents will be read-only as far as a particular reader is concerned. This typically applies to electronic documents written by others. Nevertheless the reader should be allowed to change and annotate such documents, without, of course, destroying the original. Such changes may be personal to one reader, or shared among a collaborative group of readers.

4 The nature of annotation

The purpose of annotation covers three aspects: (a) indicating the user's level of interest in each part of the material — this may be displayed using levels of highlighting, or, for non-interest, deletion; (b) adding commentary; (c) adding the user's personal links to other material (e.g. 'These results [a hypertext link] seem to contradict this'). Annotation can be applied at various levels: to an image on the screen, to a physical object, to a web page (many applications already exist for this), or to underlying data. A further property of annotations is that can apply to different communities. The community may be just one person, two persons (as often applies to Word annotation), or the whole world. It is important that the document model can capture the concept of annotation, like other document concepts, in its abstract form rather than as some way of representing it. Thus (a) above must be captured as 'this sentence is of the highest possible interest' rather than 'display this sentence in 18-point italics'. It is important that the model and its implementations incorporate the notion of automatically produced annotations, such as annotations showing the path a user takes through a set of web documents.

5 Reading

We have probably made more progress with tools for reading documents than tools for writing them. Nevertheless many parts of the technology need work:

1. Allowing the user to capture of all or part of a document for subsequent use: this is currently hard to do properly for extracts from electronic documents because of the possible presence of global declarations.

2. Allowing the user to create a link to all or part of the document being read, or perhaps to an annotated version of it.

3. Providing the user pre-emptively with suggested documents to read: thus if a user is reading a document and has indicated a high level of interest in some parts of it and has changed other parts, the system is well placed to understand the user's needs and to provide documents that meet these needs. This also applies when the user is composing a new document: the user can be shown documents that are relevant. These may relate to the content of the document or elements used within it. An example of the latter occurs when the name of a gene is used. The system can provide references to this gene in, say, their e-mails or in documents they have saved, bookmarked or linked to. (Obviously such pre-emptive behaviour must not be too intrusive, and requires a good interface — such as is provided by the work of Rhodes and Maes[10].)

6 The application area

The research will be to investigate and develop these ideas. This will not, however, be done in isolation, but will be related to an application area where there is a real need: bioinformatics. It will specifically relate to the use of hypertext (web) documents by bioinformatics researchers, and to documents they consequently write.

Consider the example of a researcher investigating whether a particular gene is related to cancer. The researcher will use a web search engine to search though many levels of information, starting with properties of the gene itself, and progressing through to published papers relating to the gene, especially if these papers also relate to cancer. We call the overall process a *search session*. During a search session the researcher will be continually presented with a choice of what links to follow. Sometimes she will select a particular link based on experience or intuition, sometimes she will make one choice with the intention of following another later, sometimes she will intend to proceed systematically through a whole set of links, e.g. a list of links to related genes.

Typically such search sessions are 'needle in a haystack', and apply to very sparse data (see below). The researcher will spend a lot of time rejecting paths, retracing steps, correcting previous assumptions, and following new paths. Occasionally the researcher will strike gold.

An important part of the researcher's output is likely to be a report, ideally of positive results. It is also useful to record failed search sessions ('I tried this path of searching and failed to find anything useful'). Negative results may be useful to others or to the researcher herself: she might like to revisit a previous session and try some small variant or correct a suspected mistake.

We plan to look at how read/write technology can help this process. It is a truism that any particular technology is likely to be a *partial* solution to a real-world problem; for a full solution it needs to be combined with other technologies. That certainly applies here. We will start by outlining how read/write technology can help, and will then outline how other technologies can also be used.

The uses of read/write technology to the example are:

1. the user may want to annotate a web page by recording brief comments on her reactions to the material in the page and her reasons for selecting a particular link to another page. Later she may want to extend or correct these annotations.

2. in writing a subsequent report she might wish to collect these annotations together (we assume each annotation comes with a specification of the web page(s) it is attached to).

3. when writing the report she might wish to be prompted automatically with web pages — either found on previous search sessions or not previously seen — that appear to relate to the report.

A key to the practical success of the above is, of course, the HCI: everything must be easy to do — semi-automatic if possible — and non-intrusive. HCI is, however, a separate issue and not the focus of our research.

Further technologies, complementary to read/write technologies, are also needed for our example:

1. a means of recording all the paths taken during a search session.

2. a means of *generating* suggested paths, e.g. following a list of links one by one.

3. an aid to more intelligent search, e.g. by following synonyms for names of biological objects (though this must be done in a controlled way to avoid a combinatorial explosion). This is especially important when the search process encompasses old papers, perhaps 40 years old.

We will look at how the above can fit with read/write technology in providing an overall solution to the bioinformatics researcher's needs.

7 Programme and methodology

Professor Brown has agreed, if this application is successful, to spend two days a week on this work. Thus there will be a total of about 100 days; given this relatively modest amount of time, the programme is concerned with concepts and feasibility, not with building complete systems.

7.1 Project objectives

The objective of the programme is to produce a feasibility report. This will consists of two parts:

1. *the Concepts Part* covering the viability of read/write technologies.

2. *the Applications Part* covering the application of read/write technology as an aid to bioinformatics researchers.

The report will cover multimedia material, though there will be an emphasis on text, since this offers the best near-term potential. The Concepts Part will examine underlying concepts to a reasonable depth, producing model(s) and taxonomies. It is possible that the Concepts Part will be partly negative (e.g. annotation [at a certain level] is not practical unless the original author provides a change notification service that offers ...').

The Applications Part will be based on input from real bioinformatics researchers (the two co-applicants and many more) working on real problems. Again the conclusions may sometimes be negative.

7.2 Methodology

The two parts of the report will proceed in parallel, with examination of applications hopefully leading to refinement of concepts. There is, of course, the practical difficulty that read/write technology is a new idea not yet implemented. However we plan to alleviate this by using existing tools — e.g. web page annotation systems such as Annotea[7] — which provide some element of read/write technology.

The Concepts Part will involve trying to find common threads in disparate existing work that relates to read/write technology. This includes the JIT projects at MIT[10] (their Remembrance Agents are, we believe, the most advanced work in aiding composition by prompting about existing documents), intelligent paper[6] (which provides electronic annotation of a physical object — in this case paper), Buneman's Grand Challenge proposal[5] (which relates annotation to the underlying data, rather than to some instantiation of it), and several projects from Xerox PARC. The PARC projects include Sensemaker[1] (which embodies ideas in fusing parts of existing documents into newly composed ones) and Fluid Links[11] (which pioneers a form of lightweight interactive annotation).

It is planned that Professor Brown will work with a range of staff, both inside and outside the Computer Science Department, and that the current programme will be a catalyst for new lines of research in documents/bioinformatics.

7.3 Bioinformatics background

This sub-section sets the context of the research within the intended application area.

Microarrays (gene chips) have made it possible to measure experimentally the expression levels of many genes simultaneously as they change over time and react to external stimuli. Many diseases, because they are caused by problems such as chromosomal imbalance and gene mutations, give rise to abnormal gene expression patterns. These patterns provide a great deal of information about the underlying genetic processes and states of various diseases. If these patterns can be appropriately analysed they can be useful for detecting disease and identifying the extent to which a patient is suffering from the disease (early or late stages), which in turn can help in the management of the disease. There are three major problems for current gene analysis techniques. The first is that of dimensionality. Typically, about 7000 genes are measured on a microarray, and each sample or individual will therefore be associated with about 7000 measurements. The second problem is that of sparsity. Typically, while thousands of genes are measured, only a few dozen samples are taken.

The Department of Computer Science is at the forefront of research in these two problem areas of gene expression data analysis[8],[9]. However, there is a third problem area which has so far not received much attention in bioinformatics research: this concerns the relevance or 'interestingness' of the set of genes identified as being potentially directly related to the disease. Once the 7000 or so genes are reduced to, say, 20, the question arises as to how relevant these 20 are, given the current state of knowledge concerning the human genome. Narayanan has identified that typically 50% of a reduced gene set as identified by current statistical techniques can be shown, through searching of web-based literature such as MEDLINE and PubMed, to have no previously recorded relevance to the particular disease in question. That is, there have been no previous references by clinicians studying the molecular basis of diseases to many of the genes in the reduced gene set. Identifying each gene in the reduced gene set and tracing the relevance of that gene through the existing literature is proving to be a major bottleneck in reporting the results of applying data reduction techniques. While data reduction of 7000 genes for 100 samples to around 20 genes can take about 20 minutes using the Stuttgart Neural Network Simulator, tracking each resulting gene manually can take 2 or 3 hours each, given the number of papers which have been previously published on this gene and the need to search these papers for references to this gene.

There is great scope for confusion and error in this tracking process, given the number of cross-references and the sheer volume of web pages visited during this process as well as the number of accession codes (gene identifiers) used. We hope the programme described here represents a new direction towards alleviating these problems. In particular, we must be careful not to throw away new discoveries concerning the role of specific genes in disease determination and prediction just because there has been no previous recorded clinical reference found in MEDLINE and PubMed. While the creation of automatic system which will distinguish genuine discoveries from irrelevant associations is still a long way off, the first step must be the development of a gene tracking tool which will record a bioinformatician's attempt to link genes in reduced gene sets to previously reported results. Such a record can then be made available to other researchers who wish to replicate the results, or who wish to integrate the results of several gene expression dataset analyses.

8 Relevance to beneficiaries

Assuming the feasibility study subsequently leads to applications being built, the most tangible benefits are to bioinformatics researchers in (a) making their search sessions more productive, and (b) in facilitating, we hope semi-automatically, the keeping of records.

There are also benefits to the computer science community in creating a better model for documents, and ultimately in fostering new applications that successfully bring document reading and writing into an integrated whole, thus offering huge usability benefits.

9 Dissemination and exploitation

We intend the feasibility study to be in the public domain. This, plus supporting detail, will be on the web site. We also wish to present our results at conferences, both to the hypertext community and to the bioinformatics community.

10 Justification of resources

Since Professor Brown already lives within 45 minutes drive of Exeter University he has agreed not to ask for travel costs. To publicisize our results we are asking for funds to attend two conferences. One of these is Hypertext'04 (which may happen after the end of the Fellowship itself), which will take place in North America. The other will be the bioinformatics conference that best appears to provide contact with the intended beneficiaries.

References

[1] M. Q. W. Baldonado and T. Winograd, 'Sensemaker: an information exploration interface supporting the contextual evolution of a user's interests', *Proc. CHI'97*, Atlanta, Ga., pp. 11-18, 1997.

[2] P. J. Brown and G. J. F. Jones. Context-aware retrieval: exploring a new environment for information retrieval and information filtering. *Personal & Ubiquitous Computing*, **5**, *4*, pp. 253-263, 2001.

[3] P.J. Brown and G.J.F. Jones, 'Exploiting contextual change in context-aware retrieval', *Proceedings of the 17th ACM Symposium on Applied Computing (SAC 2002)*, Madrid, ACM Press, New York, pp. 650-656, 2002.

[4] P. J. Brown. 'From information retrieval to hypertext linking', the invited paper in *New Review of Hypermedia and Multimedia*, Vol. 8, pp. 1-?, 2002.

[5] P. Buneman and M. Steedman, 'Annotation – the new medium of communication', proposal to the UKCRC *Grand Challenge* workshop, Edinburgh, 2002.

[6] M. Dymetman and M. Copperman 'Intelligent paper', in Hersch, André and Brown (Eds.) *Electronic Publishing, Artistic Imaging and Digital Typography*, LNCS 1375, Springer, pp. 392-406, 1998.

[7] J. Kahan, M-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, 'Annotea: an open RDF infrastructure for shared web annotations', in *Proc. WWW10 International Conference*, Hong Kong, May 2001.

[8] A. Narayanan, S. S. Tatineni and J. Gamalielsson, 'Extracting causal knowledge from multiple myeloma gene expression data' Paper submitted to *Bioinformatics*, 2002.

[9] A. Narayanan, E. C. Keedwell and B. Olsson, 'Artificial intelligence techniques for bioinformatics', Paper invited for submission to *Applied Bioinformatics*, 2002.

[10] B. J. Rhodes, and P. Maes. Just-in-time information retrieval agents, *IBM Systems Journal*, 39(4), pp. 685-704, 2000.

[11] P. Zellweger, N. O. Bouvin, H. Jehoej and J. D. Mackinlay, 'Fluid annotations in an open world', *Proc. 12th ACM Hypertext Conference*, pp. 9-18, 2001.

PART 3: Management Workplan

We do not think a lot of extra management is needed for a project employing one person for 100 days. In particular we do not see a need for milestones and timescales. Clearly the Grantholders and Professor Brown will meet regularly to monitor progress.

However we are keen to ensure that the benefits of Professor Brown's time here are fully exploited both (a) inside and (b) outside the Department. We do not see (a) as a problem, since we are a small close-knit Department, and have a regular programme of discussions/seminars to present and evaluate each other's research. Given the bioinformatics focus of the work we plan to put effort into (b). We have just set up a University-wide *Bioinformatics Network* that brings interested researchers together. This will provide an excellent forum both to allow Professor Brown to keep up with the latest practical detail of how bioinformatics researchers seek information, and for Professor Brown and the Grantholders to present ideas for criticism.