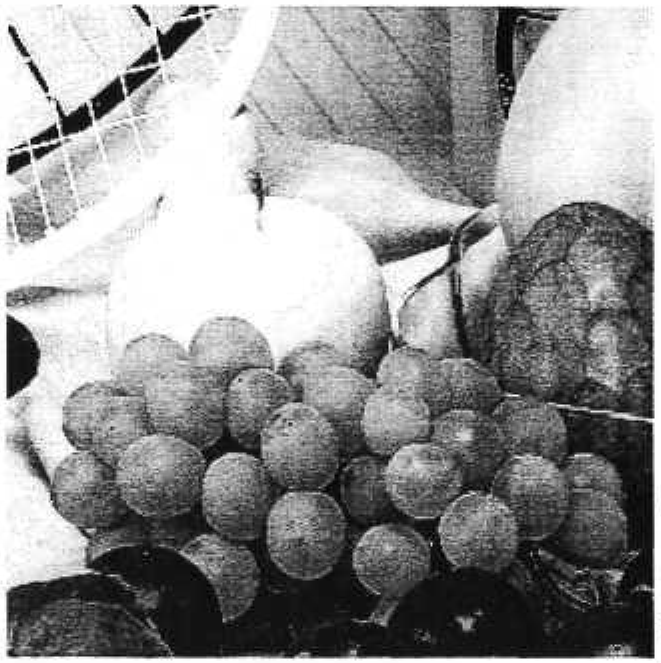

Figure 10: The original bike image

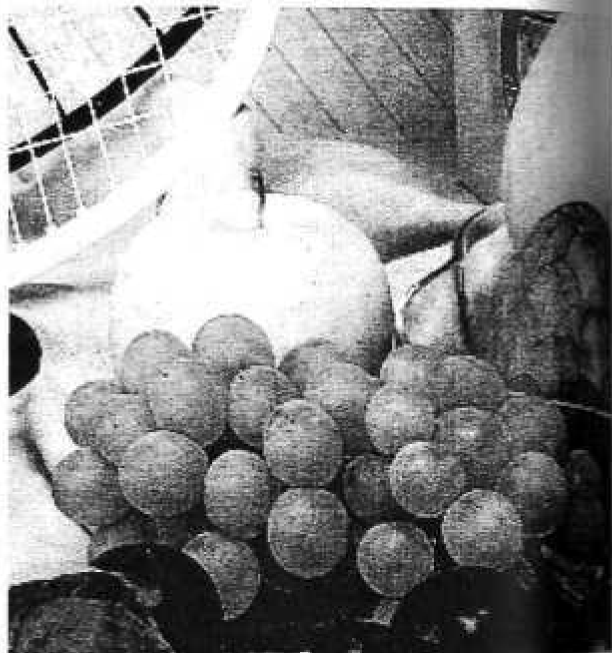

Figure 12: The watermarked bike image

Figure 11: The watermarked lena image

The use of argumentation to assist in the generation of legal documents

John Yearwood

School of Information Technology and Mathematical Sciences

University of Ballarat, Ballarat, Victoria, Australia

j.yearwood@ballarat.edu.au

Andrew Stranieri

Donald Berman Laboratory for Information Technology and Law, Dept of Computer Science and Computer Engineering

La Trobe University, Bundoora, Victoria, Australia

stranier@cs.latrobe.edu.au

Chaula Anjaria

School of Information Technology and Mathematical Sciences

University of Ballarat, Ballarat, Victoria, Australia

c.anjara@ballarat.edu.au

Abstract

Many text documents in the legal domain are created in order to express the reasoning steps a decision maker followed in reaching conclusions. For example, refugee law determinations are documents that express the reasoning steps a member of the Refugee Review Tribunal in Australia followed in order to infer conclusions regarding the status of an applicant. Although, it is reasonable to expect that a mapping between the reasoning steps used by a decision maker and the structure of the document produced would clearly be apparent, a number of authors have discovered that such a mapping is by no means obvious. In order to develop legal knowledge based systems that generate documents from their own reasoning steps, discourse analysis is invoked to bridge the gap and perform the mapping. In this paper, we articulate a heuristic that we use to generate a plausible document structure without the use of discourse analysis. Without discourse analysis, the heuristic cannot ***contribute*** *to our understanding of the process employed by decision makers to convert reasoning to text. Nevertheless, the heuristic can mimic the process. The heuristic has been trialed with a small sample of refugee law determinations by extracting the reasoning steps from each determination and applying the heuristic to reproduce each document's structure.*

Keywords Document generation, argumentation, refugee law.

1 Introduction

In many applications of human reasoning, conclusions ultimately reached and the reasoning steps employed to reach conclusions are expressed in written natural language. For example, the inferences that members of the Refugee Review Tribunal (RRT) make in assessing claims for asylum seekers to remain in Australia as refugees are natural language documents called determinations that vary from 6 pages to many tens of pages in length and only loosely conform to a pre-defined structure. However, although determinations reflect reasoning, each one is written in a style that is not obviously consistent with a representation of refugee reasoning that we have developed over the last two years in close collaboration with RRT members.

The disparity between a natural language document and a representation of reasoning that a document expresses has been noticed by a number of authors using different knowledge representation schemes so is likely to be an artifact of communication styles rather than a peculiarity of any one knowledge representation scheme. Dick [2], in translating legal decisions that spanned hundreds of years into conceptual graph frames initially attempted to do so by creating graphs directly from text components. This proved to be too difficult because of the disparity between the document text and a conceptual graph representation of the reasoning that the document expressed. She proceeded by reading each judgement in its entirety, then formulated a set of conceptual graphs from her understanding and not from the text.

Proceedings of the 4th Australasian Document Computing Symposium,
Coffs Harbour, Australia,
December 3, 1999.

More recently, Branting *et al* [1] reiterate the absence of isomorphism between an inference tree that represents domain reasoning and a rhetorical tree that represents the organisation of text in a document and go further, first, by clearly articulating the benefits of a mapping and then by utilizing discourse knowledge to bridge the gap. According to those authors, the discourse structure is particularly clear in documents such as contracts or wills and therefore well suited to natural language generation techniques aimed at planning and realising multi-sentential text generation.

Multi-sentential text generation has application in tasks such as contracts or wills because these tasks re-use similar documents in a manner that becomes difficult for a human to perform quickly and in large numbers. For these applications an automated system that generates as much of the finished document, as quickly as possible, is preferred. However, automated document generation systems can be useful to support decision makers in those tasks that involve less re-use of similar documents and more individual crafting of text to suit the case at hand. The drafting of a refugee determination is unlikely to ever be automated like a wills generator can be, because the life and death nature of a determination necessitates, for political if not for moral reasons, the full involvement of human decision makers. Nevertheless, a document drafting system can support, without crossing the imaginary line towards replacing, the decision maker by generating a document structure and leaving the creation of sentences to the author.

We agree with Branting et al [1] that the discovery of a mapping between reasoning steps and document structure based on discourse knowledge leads to improved document generation. Indeed, it seems unlikely that a complete mapping between reasoning steps and document structure can ignore discourse analysis. However, a rigorous mapping is not necessarily required if the sole objective is to convert a series of reasoning steps into a plausible document structure. In this study we propose a relatively simple heuristic for the traversal of an inference tree that realises a structure that is similar to that created by authors. The heuristic cannot explain why a particular document structure derives from an inference tree. Some invocation of discourse analysis would be required for that level of description. Nevertheless, in the restricted domain of refugee law determinations, the heuristic can be applied to generate a document structure that is similar to one that a human author would plausibly arrive at. The heuristic can therefore be used as the basis of an automated document generation system that is much simpler than one based on discourse knowledge.

There is quite a large variation between what are considered to be discourse segments and discourse relations in the field of discourse analysis. In the Rhetorical Structure Theory of Mann and Thompson [6] there are the notions of non-overlapping text spans called *nuclei* and *satellites* with rhetorical relations such as *elaboration* which holds between them. In the Grosz and Sidner Theory (GST)[5], the elementary text units are called discourse segments and the discourse is explicitly stated to be a tree. Each discourse segment is characterised by a primary intention called the *discourse segment purpose*. GST identifies only two kinds of intention-based relations between discourse segments: dominance and satisfaction precedence. A large amount of research in discourse planning is based on the speech act theory proposed by Searle [8]. Dietz and Widdershoven [3] identify limitations of Searle's theory put forward by Habermas [4] and go further by noting that communication support systems based on Habermas's theory are quite different from those based on Speech Act Theory. The mappings between rhetorical relations and speech acts are problematical as are the mappings between intentional and informational levels. Marcu [7] provides a melding of text structures and intentions by formalising a discourse structure with nodes characterised by four features: the status of the node, the rhetorical relation that holds between the nodes that are immediate children, the set of salient units and the primary intention. We propose investigating this structure with the cases studied because there have been many instances where the explanation for a discourse segment relates to underlying intentions of the member that augment the reasoning.

This work achieves a transformation from a reasoning structure to a draft document structure that fits with the examples studied. Without the use of sophisticated discourse planning the user of a decision support system is able to refine a document which presents a relatively complete record of their reasoning. They may then be involved in a refinement process that consists of removing points that they would prefer to have implicit, leaving the salient points and possibly elaborating on these.

In this paper, we initially outline the knowledge representation schema we use in order to illustrate that it is relatively complex and therefore can capture intricate associations. Following that, we discuss our experience in extracting reasoning steps from a determination. Our objective is to demonstrate the effectiveness of the heuristic by reproducing the structure of the determinations in our sample. Results indicate that a relatively simple heuristic can mimic the structure of determinations and therefore be used as the basis of a system that supports the decision

maker by generating a document outline that displays a plausible structure.

2 Argument based knowledge representation

The knowledge representation we used is a variation from the Toulmin argument structure [10]. Despite the immediate appeal of the Toulmin argument structure (TAS) as a convenient frame for representing knowledge, most researchers that use Toulmin structures vary the original structure. A survey of the different variations of the Toulmin structure can be found in [9]. The statement "Most Saudis are Muslim" in Toulmin's now famous example, may be a warrant that convinces us that the assertion "X is a Muslim" follows from the knowledge that "X is a Saudi". However, this warrant communicates two distinct meanings. On the one hand the warrant indicates a reason for the relevance of the fact "X is a Saudi". On the other hand the warrant can be interpreted as a rule which, when applied to the fact that "X is a Saudi" leads us to infer that "X is a Muslim". These two apparent meanings are best perceived as different roles the warrant has in the structure of an argument. Drawing the distinction between the two roles a warrant has in an argument, leads us to explicitly identify four features that are left implicit in the Toulmin formulation. Figure 1 illustrates Toulmin's Saudi argument using our representation. The four features made explicit are:

- reasons that explain why a data item is relevant for a claim. The warrant in the original structure is seen as a reason for the relevance of the data item. In other examples that Toulmin uses, his warrant statement is a rule that we equate to an inference procedure.
- an inference procedure, algorithm or method used to infer a claim value from datum variable values. This must be made explicit if a computational model is to be built because conceivably any computable function can be used to infer a claim value from data item values.
- a {variable : value } formulation of each claim and data. This formulation is convenient for the application of an inference procedure.
- reasons that explain why the inference method used is appropriate. In some examples the inference procedure is a neural network. Reasons for the appropriateness of the inference procedure relate to how well the network was trained. In many legal arguments, a statute may make explicit an inference rule, in which case, the statute is the reason for appropriateness of the inference procedure.

We call the argument in Figure 1, a generic argument. An actual argument is an instance of a generic argument in that that values for variables are set. For example, an actual argument may infer X is almost certainly a Muslim from data that indicates X is certainly a Saudi. The Reasons for relevance of a data item and reasons for the appropriateness of a inference procedure apply to all actual arguments instantiated from a generic one regardless of variable values. However, a component we call a claim value reason is sometimes necessary in order to encode a reason for why a particular claim value follows from a particular data item value.

We have identified over 200 generic arguments that have been used by members of the Refugee Review Tribunal or applicants for refugee status by perusing determinations and interviewing members. Figure 2 illustrates the generic argument that makes the claim regarding well founded fear of persecution. No inference procedure has been specified for this argument as research is initially in progress toward the use of the knowledge representation to support human inferences and not to implement machine inferences.

The arguments combine to form a chain or tree of arguments because one argument's claim is another argument's data item. For example, another argument exists that has, as its claim, The applicant (does, does not have) a real chance of persecution , which is one of the data items in Figure 2. The ultimate argument asserts an applicant (is, is not) a refugee. In the next section we shall describe the representation of a sample determination using our schema and contrast this with the text of the determination.

3 Juxtaposition of reasoning steps with document structure

The first sample refugee determination involves a Sri Lankan woman who was determined to be in need of refugee status by the Refugee Review Tribunal. Her claims for refugee status were based on an incident some years ago in which her home was burnt down and more recent incidents of beatings and other reprisals by the armed forces following contact with persons suspected of belonging to the rebel army, LTTE. The determination is 5 pages in length and contains the following headings: *Jurisdictional Foundations* (Two paragraphs common to all determinations that establish the authority of the Tribunal to review the case); *The Law*. (Eleven paragraphs common to all determinations that outline leading cases, statutes and other relevant background information); *Findings of Fact*. (This section contains the member's summary of all claims an applicant has made. Associated with a summary of each claim is the member's own finding regarding the veracity of the claim and accompanying reasons.); *The Decision* (Two paragraph statement of the decision).

Figure 3 represents a partial tree of claims extracted from this determination. Reasons for the relevance of data items, appropriateness of inference procedures or for particular claim values, are omitted in Figure 3 for brevity. The claim labelled A with value (is) indicates the applicant is a refugee. This is made on the basis of the value of one data variable, B. B is the claim for an argument in its own right and has data items C and D. Claims in Figure 3 stop at N for brevity. In total there were 36 claims identified in the determination.

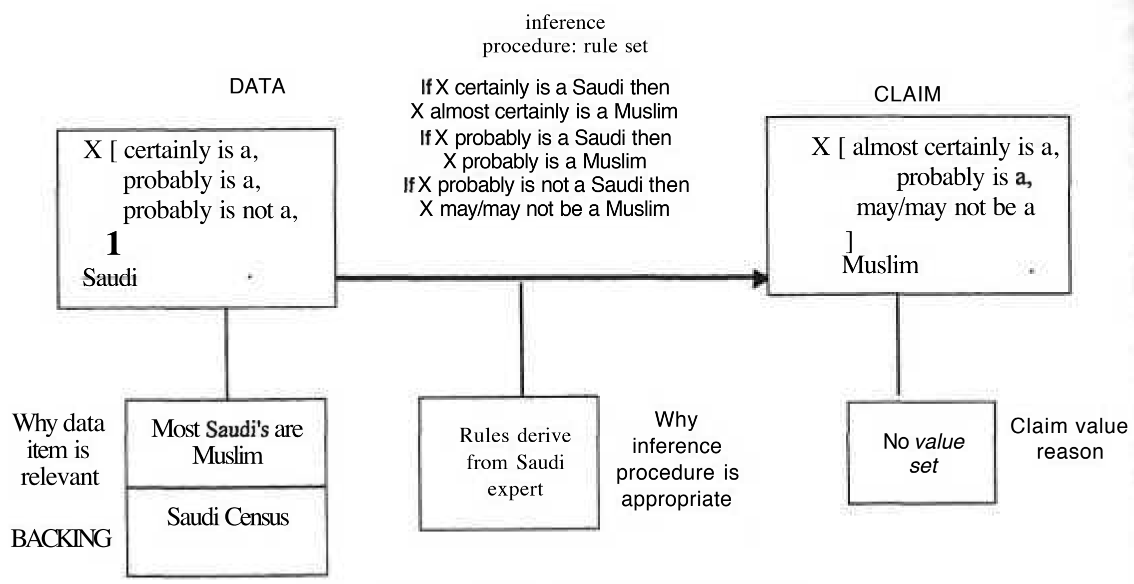

Figure 1. Our variation to the Toulmin Structure

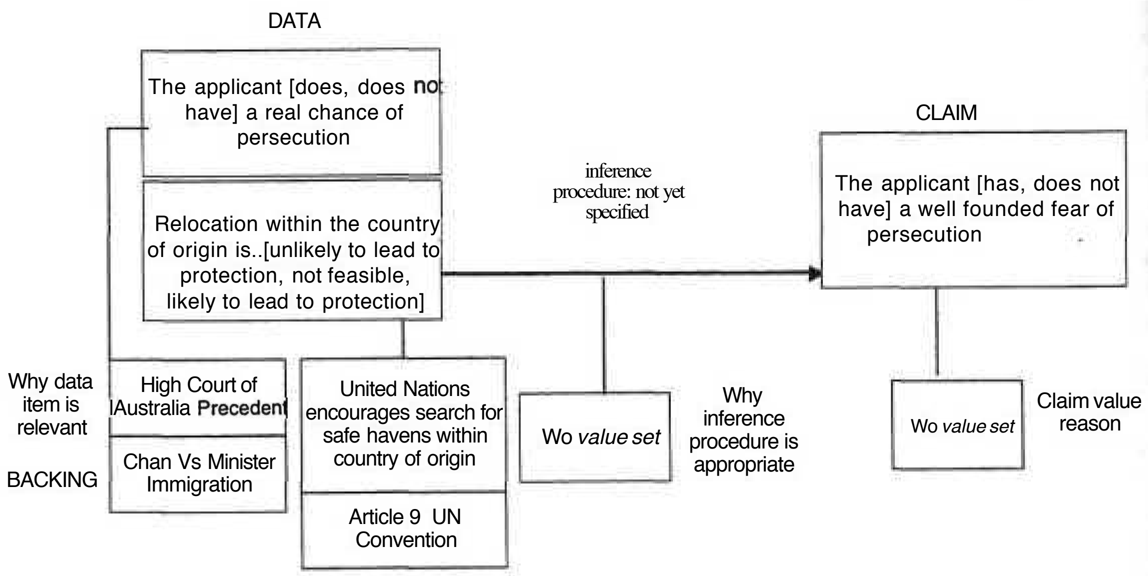

Figure 2: Generic *Well-founded fear* argument

The values for claims were extracted manually by reading the case, tagging passages relevant for a claim and assigning the value that seemed best to capture the member's intended meaning. For some claims the appropriate passage and value was obvious. For example, claim A, the variable *refugee* is unequivocally taken to have the value (*is*) from our reading of the following paragraph in the determination:

*The application **for** a domestic protection (temporary) entry permit is remitted for reconsideration with a direction that the criterion requiring the applicant to be a non-citizen who has been determined to be a refugee under the Convention and Protocol, is satisfied.* (1)

The variable, G, *Taken together all incidents constitute harassment* is associated with the text fragment:

*The burning of the applicant's home and belongings in Trincomalee, was a deliberate act designed to punish the applicant and her daughter for their perceived connection with the **LTTE**. Such an act does not carry the imprimatur of international approval as a reasonable use of force for purposes of self protection. Furthermore, I am satisfied that the applicant's daughter was actively persecuted thereafter by Sri Lankan military authorities in a totally unacceptable manner. I find that both the applicant and her daughter, XX, were persecuted by the Sri Lankan authorities prior to their departure.* (2)

There could be found no text fragment that directly corresponded to some variables in our representation. For example, variable J describes the extent to which an incident impacted on the applicant in a severe or persecutory manner. We surmise that the member has inferred that the house burning incident in Case 1 did impact severely on the applicant because of the tone of the text associated with variable G, depicted here in the paragraph above. We surmise that the member has left text fragments associated with this, and many other variables, implicit.

The reasoning expressed in the determination was found to map onto 36 variable:value tuples contained within 10 arguments. Passages of text such as the examples above mapped directly onto 18 of the 36 variables. Values for the remaining **18** variables were not stated explicitly in the determination, but like variable J, were left implicit. A value for the *well founded fear* variable is similarly not explicitly mentioned in the text but is assumed by the context of the entire document to have the value *has*

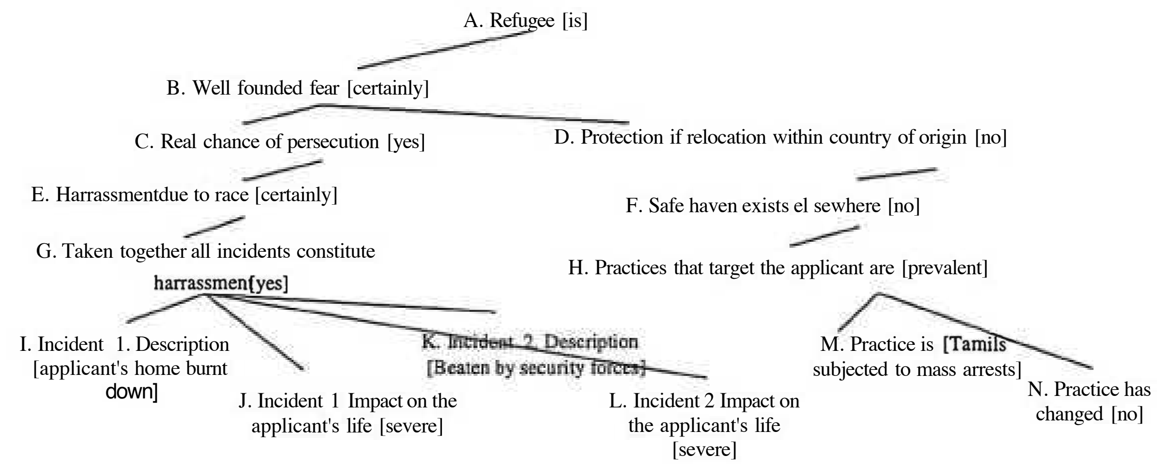

Figure 3. Partial tree representing reasoning steps in case 1

Traversal	Sequence
Top down, breadth first traversal	A, B, C, D, E, F, G, H, I, J, K, L, M, N
Bottom up	I,J,K,L,G,E,C, M,N,H,E, F,D, A
Actual	1..16, I,K,F,M,G,E,C,38,39, A.
Bottom up, sub-tree constrained traversal	I, {J}, K, {L}F,M,{N},{H},G,E,C,{B}{D}A

Table 1. Sequence of passages associated with reasoning steps of Example 1

Twenty cases were randomly selected for analysis. Text fragments from three of the cases were associated with variables from our representation independently by three assessors. The majority of text fragments were associated with the same variable:value pair by each assessor. However, there were some anomalies. Each anomaly originated from some ambiguity concerning the discourse intention of the member. For example, a fragment in a case involving a Timorese applicant was associated with the *well founded fear* variable by two assessors and with the *credibility* variable by the other assessor. The disagreement was due to subtleties in interpretation of the intention underpinning the text. This is a clear indication that rigorous analysis of text requires discourse analysis representation and that speakers intentions need to be taken into account. However, the anomalies occurred sufficiently infrequently to suggest that, in the majority of refugee determinations, the member's intention is clear. This obviates the need for sophisticated discourse analysis if the objective is solely to support the member in generating a base determination for further refinement.

Table 1 illustrates two possible traversals of the tree in Figure 3; a top down, breadth first and a bottom up variant. The same table also indicates the sequence within the determination that the passages actually appeared and a traversal we call bottom up, sub-tree constrained that we describe below.

Paragraphs 1..16 and 38..39 do not correspond to any of the generic arguments we used to model refugee determinations. Paragraphs 1..15 comprise the standard paragraphs in the *Jurisdictional Foundation* and *The Law* sections that provide background information but do not relate specifically to the reasoning steps in the case at hand. Paragraphs 16, 38 and 39 relate to the age of the applicant and although a generic argument does not exist we can represent the reasoning made in these paragraphs with the inclusion of an additional and new argument.

As expected, a top down or bottom up traversal of the inference tree does not obviously map neatly onto the actual structure of the document. However, a traversal that we call *bottom up, sub-tree constrained* represents a heuristic traversal that does map the document structure neatly onto the inference tree. This traversal is described informally as follows:

1. Begin at any node near or on a leaf
2. Visit all descendent nodes, that is, all nodes within the sub-tree defined by the node, in any order until there are no nodes unvisited in the sub-tree
3. Where possible, insert implicit nodes adjacent to siblings
4. Select any unvisited node and repeat.

The traversal I, {J}, K, {L} F, M,{N},{H},G, E, C,{B}{D}A is consistent with the heuristic as follows:

- Begin at node I because it is a leaf,
- I has no sub-tree
- Insert implicit node {J}
- Visit any node not visited: K
- K has no sub-tree, Insert implicit node {L}
- Visit any node not visited, F
- Visit any node in the sub-tree of F: M
- M has no sub-tree, Insert implicit node {N}
- Visit any node in the sub-tree of F: {H}
- Sub-tree F completed so visit any unvisited node: G
- All nodes in sub-tree of G have been visited so visit any unvisited node: E
- All nodes in sub-tree of E have been visited so visit any unvisited node: C
- All nodes in sub-tree of C have been visited so visit any unvisited node: {B}
- Visit any node in the sub-tree of {B}: {D}
- All nodes in sub-tree of {B} have been visited so visit any unvisited node: A

The traversal heuristics do not produce a document structure isomorphic to the inference tree. This is not a limitation because the expression, in text, of the same reasoning by different persons is not expected to be identical. The objective of a traversal algorithm is to produce any good document structure and not the one optimal document however optimal is defined.

The traversal heuristics have intuitive appeal because Members, in weaving a determination, do not commence at the conclusion because there are too many issues that have contributed to a value at that level. Instead an issue closer to the facts (leaves) are selected and all facts and issues relevant to that issue are discussed (ie. all sub-tree nodes). Once that issue is covered, the author has creative freedom to select any other issue. Discussion pertaining to that issue continues until all issues and facts relevant to that are covered. While it is certainly possible that an author exercising a skillful degree of flair may construct a document using heuristics other than these, our contention is that the heuristic presented here can be used to generate a document that appears at least, coherent to a reader, even if it lacks dramatic emphasis.

The algorithm that describes this traversal relies on a global data structure that contains the list of nodes that have been visited in a tree. This is initialized to null before the algorithm, traverse, is called for the first time, with the root of the tree as argument.

The function pickANode selects any node in the tree with root, R that has not already been visited and is near the leaf nodes. The function nodesToVisit terminates the traversal of the tree/sub-tree when there are no nodes unvisited in the sub-tree.

traverse(N)

Purpose: visit nodes in a tree using bottom up, sub-tree constrained traversal

Arguments: R, node label that is the root if the tree to be traversed

Updates: variable visitedNodes

Calls: nodesToVisit, pickANode, traverse

```
While nodesToVisit (visitedNodes, R) = true
   Do N gets pickANode(visitedNodes, R)
      If N <> null
         Then begin
               Append N to visitedNodes
               Traverse (N)
            end
      endif
endwhile
end traverse.
```

In order to illustrate our traversal we take a second example, drawn at random from a pool of determinations. In the second case, the applicant is an ethnic Tamil citizen of Sri Lanka from Jaffna with a heart condition who experiences some incidents of harassment whilst in Colombo. Return to his home in Jaffna would involve a real chance that his life would be endangered because of a government embargo on medical supplies to the area (directed at LTTE sympathizers). Relocation to the most likely area of Colombo is not possible because none of the threshold criteria for returnees to establish themselves without serious security problems can be met.

Table 2 illustrates that the bottom up, sub-tree constrained traversal described by our heuristics reproduces the actual argument if the split passages associated with, variable N are united. Although quite different factually, examples 1 and 2 both contained 10 Toulmin arguments and over thirty variable:value attributes. In both examples there was a high proportion of implicit variables (18/36 for example 1 and 18/31 for example 2). In summary, example 2 confirms the observations we made from the first example.

To date, the reasoning steps in twenty cases drawn at random from a pool of RRT determinations have been extracted using the knowledge representation frame described above. The heuristic we described has been applied to each of the twenty cases in order to ascertain whether the structure of the actual document could be produced using our heuristic. The heuristic was found to account for the actual structure of the document in each of the twenty cases. We plan to survey member's opinions of document structures generated with the use of the heuristic. Table 3 illustrates a case that is one of the more intricate we sampled. In this case the reasoning steps involved 58 variable:value tuples. Twenty eight were explicitly represented by text fragments and 30 were implicit. The sequence of fragments once again conforms to the heuristic described above.

Traversal	Sequence
Actual	N,K,N,9-16,N,L,M,C,25,F,I,O,P,J,D,A
Bottom up, sub-tree constrained traversal	N, K, L, M, {G}, {H}, C, F, I, 0, P, {Q}, J, D, {B}, A

Table 2. Sequence of passages associated with reasoning steps of Example 2

Traversal	Sequence
Actual	P1-11,22,24,25,Y,27,26,1,29,15,3,B,5,32,7,35,F,9,21,F,C,H,F,P36-P39, C, F, W, X,L,F, C, B,D,P51,A
Bottom up, sub-tree constrained traversal	22, {23},24, 25, {11},{V}, L, W, X, {M} {I}, Y, 26, 27, {28}, {12}, {13}, {2}, {N}, 1, 29, 30, {31}, {16}, {15}, {2}, {O},3, {4}, {P}, 5, 32, 33, {34}, {6}, {Q}, 7, 35, {36}, {37}, {18}, {19}, {8}, {R}, 9, {20}, 21, 10, S, {J}, {E}, F,C, {T}, {V},{K}, {G}, H, D, B, A

Table 3. Sequence of passages associated with reasoning steps of a third case

4 Conclusion

In performing a fine-grained analysis of the mapping between a document structure and the reasoning steps expressed in a document we noticed that a mapping is not obvious. The heuristic that we use to generate a plausible document structure without the use of discourse analysis has proved to be successful in the tests so far. Without discourse analysis the heuristic cannot contribute to our understanding of the process employed by decision-makers to convert reasoning to text.

Nevertheless, the heuristic can mimic the process sufficiently well to be useful in the task of supporting document drafting in a complex domain where there will always be a need for human interaction and refinement. Future research in this direction is aimed at formalising the heuristics and engaging in more rigorous evaluations. An evaluation should include comparisons of documents generated with our heuristics with actual documents in addition to studies that measure the readability or coherence of documents generated with our heuristics.

5. Acknowledgements

This research was supported by the Refugee Review Tribunal, Australia and the Australian Research Council.

6. References

[1] Branting, L., K., Callaway, C, B., Mott, B., W., and Lester, J., C, 1999. Integrating Discourse and Domain Knowledge for Document Drafting. *Proceedings of Seventh International Conference on Artificial Intelligence and Law ICAIL'99.* ACM Press. pp214-220.

[2] Dick, J. P. 1991. *A conceptual, case-relation representation of text for intelligent retrieval*. Ph.D Thesis. University of Toronto. 1991. Canada.

[3] Dietz, J. L. G., and Widdershoven, G. A. M., 1992. A comparison of the linguistic theories of Searle and Habermas as a basis for communication supporting systems in van de Richt, R. P. and Meersman, R. A. (Eds)., 1992. *Linguistic Instruments in Knowledge Engineering*. Elsevier Science Publications. Pp121-130.

[4] Habermas, J., 1987. The theory of communicative action. (tr Thomas McCarthy). Boston : Beacon Press

[5] Grosz, B. J, and Sidner, C. L. 1986. Attention, Intentions and the structure of discourse. *Journal of Computational Linguistics*. Vol 12, No 3. Ppl75-204.

[6] Mann, W, C. and Thompson, S. A., 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text* Vol 8, No. 3. pp 243-281.

[7] Marcu D. 1997. The Rhetorical Parsing, Summarisation, and Generation of Natural Language Texts. Ph D Thesis, Department of Computer Science, University of Toronto.

[8] Searle, J., 1969. *Speech Acts: An Essay in the philosophy of language*. Cambridge University Press. Cambridge.

[9] Stranieri, A., and Zeleznikow, J. 1999. A survey of argumentation structures for intelligent decision support. *Information Systems and Decision Support Systems ISDSS'99* Melbourne July 1999 Monash University Press.

[10] Toulmin, S. 1958. *The Uses of Arguments*. Cambridge University Press. Cambridge

Automatic document metadata extraction and manipulation: a working system for the Intelligence Analyst

Mark Burnett
DSTO
Dept of Defence
Fern Hill Park, Bruce, ACT

Richard Jones
Lloyd-Jones Consulting
PO Box 6155
Philip ACT 2605

mark.burnett@dsto.defence.gov.au jonesrl@compuserve.com

Abstract

This paper discusses the design and implementation of an operational system to aid health intelligence analysts. The HINTS system provides automated support to undertake tasks such as specific health-related research and report writing in the face on an ever-growing body of electronic information, available on the web, and on local file systems. Our approach is to provide automated support for document analysis and discovery from technologies that support ad-hoc searching, consistent filtering for specific pieces of information such as hospital facilities, diseases and locations, and that provide document summarisation and keywording. Document metadata is stored in XML in a data structure that allows a variety of searches and views of the document space to be performed. The user interfaces to the system by web browser and a map-based geospatial application.

Keywords
Document Analysis, Document Databases
Information Retrieval, XML,
Information Extraction

1 Introduction

Intelligence analysts typically operate in two modes. In the first mode, they scan, on a regular basis, a wide range of documents from a range of sources, in case they contain something that might be useful. A selection is made on some general criteria and those documents are put to one side for a rainy day, organised in some way so that they can be found later. They move into the second mode of working when called upon to prepare a specific brief. They obtain the source material for the brief by accessing this repository of information, and combining this with the results of specific searches from other sources. Invariably the information that is key is a side issue in a document that may be discussing some other topic. This information extraction requirement is, of course, the focus of the MUC series of experiments [1].

This paper describes a repository system to support both modes of operation described above, but with more emphasis on the first. The system, known as HINTS, is based upon an XML store containing three types of information, standard bibliographic information, domain specific information - in this case the health domain, and information specific to the intelligence analysis requirement. A key part of the system is to assign specific values to metadata elements automatically. To this end an information extraction process looks for specific entities (e.g. disease names, locations), and a document summariser assigns keywords and extracts a summary.

A strong goal of the design was to make the system generic and easily applicable to other intelligence domains. To this end we selected technologies that operate on text at a surface level, and a component-based architecture for integration.

2 XML Data Storage

We chose an RDF implementation of XML to provide a generic storage mechanism, and to allow concepts, and relations among concepts, to be defined. RDF also supports a class system analogous to Obect-Oriented programming and modeling systems. The RDF specification has now progressed to a "Proposed Recommendation" to the W3C (see http://www.w3c.org/TR/PR-rdf-syntax).

3 Metadata Extraction

It is often useful for users to read some sort of condensed or structured surrogate for a document. The purpose may be varied: it may be to determine if the document is worth reading in full, or to extract specialised information from the document, either by reading specific portions, or even without reading it at all. Some standard structures to support these requirements, (e.g. MARC) and abstracts, keywords, etc. have been in place for a long time. This

Proceedings of the 4th Australasian Document Computing Symposium,
Coffs Harbour, Australia,
December 3, 1999.