Case-based Retrieval of Refugee Review Tribunal Text Cases

John Yearwood

**School of Information Technology and Mathematical Sciences,
University of Ballarat. Ballarat 3353, Australia.
jly@ballarat.edu.au**

Abstract

Reasoning from cases has always played a large part in law and so the ability to effectively retrieve and reuse old cases in formulating new decisions and arguments is important. This paper examines techniques for improving retrieval effectiveness by using the structure present in the text cases. The Refugee Review Tribunal cases are pre-processed into a consistent structure and evidence from the case components is combined using retrieval functions based on different strategies for determining the importance of components. The work indicates that a minimum success level in retrieval would be 54% based on a simple rule, the use of full text matching has a 66% success rate and the combination of case component evidence produces a similar success rate. Better results are achieved when a derived legal structure is used and even with a simple derivative process the results have greater depth.

1. Introduction

Many areas of law, health and commerce have large numbers of cases in text. In many instances there is an explicit document structure that is evident in the sectional structure of the text. For example the transcripts of many legal cases have a Headnotes section, a Body section and a Decision. It is not often that the sectional structure of these documents corresponds directly in a detailed way to the case structure as we may think of it from a knowledge engineering or reasoning point of view. Given this situation and the increasing need for the reuse of cases in formulating new decisions and judgements it seems appropriate to consider the following broad strategies: users may retrieve previous text cases using a full text search tool without regard to structure; they may retrieve cases based on good sectional text matches; they may retrieve cases based on good text matches in the important features of the cases; or they may consider converting the text cases to a knowledge base and carry out case-based retrieval on the formal highly structured cases. Currently many users are moving to the first strategy.

This work considers the second and, to a limited extent, the third approach in the domain of Refugee Law. The number of refugee cases that need to be processed has increased dramatically over the last ten years and better support tools for decision makers are needed. The Refugee Review Tribunal (RRT) consists of members who review the decisions of the Immigration Department on refugee status. Each RRT case documents the application, the background, considerations on jurisdiction, the legal framework the applicant's case and the findings. Unfortunately even this sectional structure is not uniform across the database. This work investigates semi-automatic uniform sectioning of the documents and retrieval based on the

resultant case structure. Hearst (Hearst, 1993) as well as Mittendorf and Schäuble (Mittendorf, 1994) have used more sophisticated techniques for segmenting text into coherent and motivated sub-topics which are then used to enhance retrieval. Their work has not dealt with case structure.

Once a consistent text case structure is established it is possible to formulate a matching model that is based on matching the individual text components and combining evidence from these. A training set of 50 queries from the case base of approximately 1000 cases is used for a regression model for combining the component evidence.

Previously, good results were achieved with applying case-based text retrieval results in the nursing domain with highly structured data. The RRT cases are different in that they contain much larger sections of text which are not so clearly delimited and there is not a clear alignment of the sectional document structure with the case structure. There is also a severe overlap in the vocabulary common to sections and the outcomes are of a dichotomous nature.

The paper is organised as follows: Section 2 examines different views of structure that may be helpful in designing a retrieval function; sections 3 and 4 discuss the effects of the user and queries on the retrieval model; sections 5 and 6 briefly discuss indexing and matching; section 7 outlines the methodology, section 8 describes how a coarse legal structure that is implicit in the cases is extracted and used in a component matching model as well as the experiments and results obtained from using this legal structure.

2. Case Structure

We could consider the structure of these cases from at least three viewpoints. First, each case is a document with headings above sections and secondly there is a discourse where some elements logically depend on others. Thirdly there is a legal structure to an RRT case.

2.1. Sectional structure

The RRT cases do not have a uniform sectional structure. That is, each case is not described according to prescribed sections in a template. For example there are almost always sections on THE APPLICATION and THE BACKGROUND, and usually sections headed JURISDICTION or JURISDICTIONAL FOUNDATION and THE LAW or LEGISLATIVE FRAMEWORK. There may then be a selection from many headings such as THE APPLICANT'S CASE, APPLICANT'S CLAIMS, CLAIMS, CLAIMS AND EVIDENCE, CLAIMS EVIDENCE AND FINDINGS and CLAIMS AND FINDINGS. Other sections are REASONS FOR DECISION, DISCUSSION OF FINDINGS and there is generally a CONCLUSION.

Although there are some differences in what some 'members (of the tribunal)' discuss under THE LAW and under LEGISLATIVE FRAMEWORK there is sufficient similarity to view these as one component (LF). Generally it also seemed reasonable to classify sections headed APPLICANT'S CLAIMS, EVIDENCE etc. as THE APPLICANT'S CASE and to cluster FINDINGS, REASONS FOR DECISION and DISCUSSION OF FINDINGS into FINDINGS. The difficult task was to separate the FINDINGS in CLAIMS EVIDENCE AND FINDINGS and CLAIMS AND FINDINGS into APPLICANT'S CASE and FINDINGS. This was done manually.

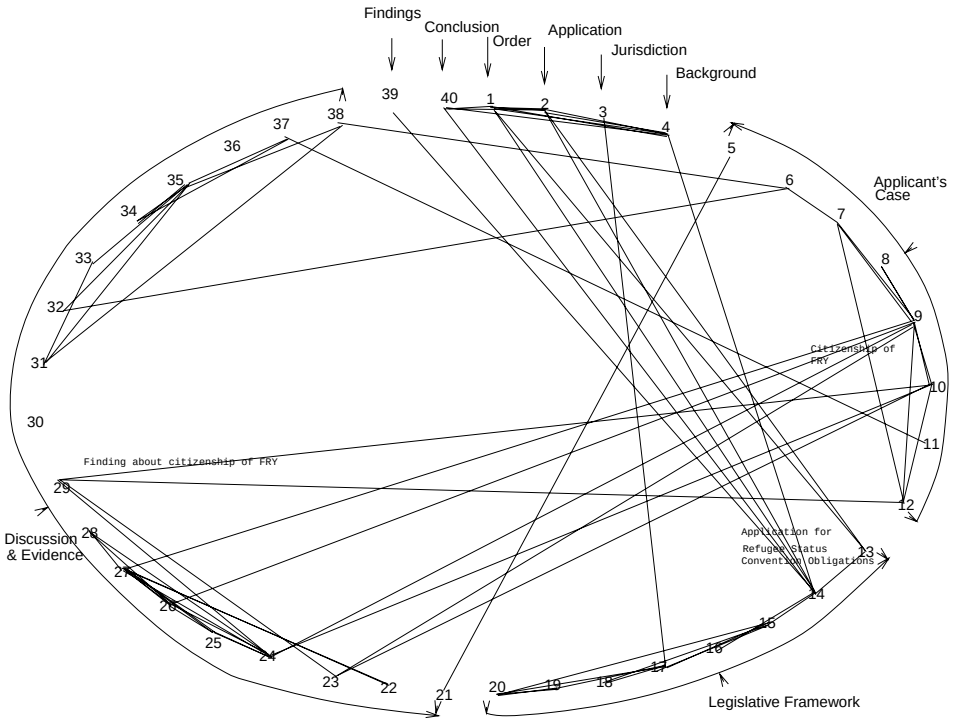

Figure 1: Relating paragraphs in an RRT case

Figure 1 is a diagram of the type used by Salton, Allan and Singhal (Salton, 1996) in their text structuring and decomposition experiments and shows the similarities between the paragraphs of an RRT case (case V95/03431) which are approximately greater than one third of the similarity of each paragraph matched with itself. The RRT cases have a structure as shown in Figure 2. The diagram of links does not indicate that the case components stand out as distinct segments although there is relatively little cross component linking. It seems that the larger components might consist of one or more segments and some of the smaller components are strongly connected.

- **It can be seen that it is difficult to break down evidence and findings in this way. Paragraph 29 & 39 are both findings. The triangle 29 (10,12) is a theme. 29 is really an interim finding about citizenship of FRY. The triangle 9 (24,27) is a theme about citizenship of FRY. The triangle 14 (4, 40) is a theme about the Application for Refugee Status and obligations according to the Convention.**
- **Some of the small case components are related.**
- **Some of the larger case components may be broken into segments and some have themes relating to aspects in other case components.**

2.2. Discourse structure

Text structure reflects a progressive accumulation of semantic information. Authors commonly start with points known to readers and progress to a state of communicating things that were not known. In text structure the theme represents elements that are in some way related to the preceding text. The rheme expresses information that is in a sense new or unpredictable from what has been said already (Hutchins,

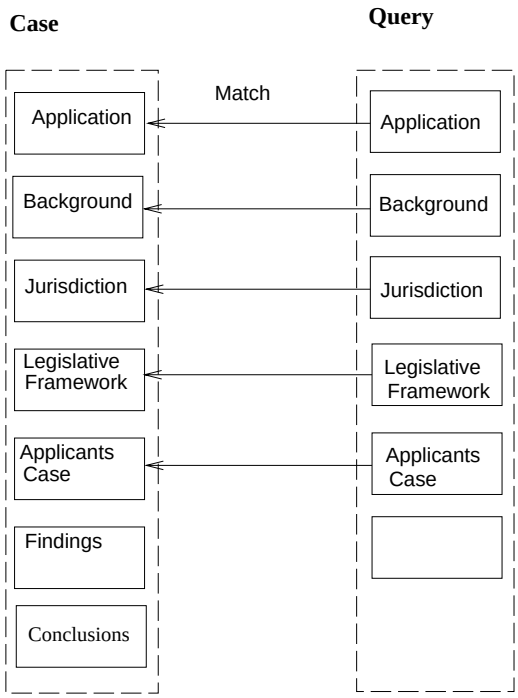

Figure 2: Case Components, query components and the matching model

1977). At the level of sentence structure, the linking and binding of elements is seen by the use of anaphoric devices like pronouns and definite articles.

There are two basic types of sentence progression from the theme-rheme point of view: linear progression, where the favoured thematic elements relate to the preceding rheme Figure 3; and parallel progression where the theme remains constant Figure 4.

Figure 3: Linear progression

The macro-structure of a text is similar in that it is in the initial sections of a text that the author establishes the foundations of what the reader may already know or what can be assumed. So the thematic part of a text is established early and is usually rich in clues as to what the text is 'about'. At this level the rheme will express what the author has to say about the theme.

In terms of effective matching of a user query with a text it may be suggested that the text is more likely to be relevant (answers the users information need usually by the new information in the document) if the balance between the foundation

Figure 4: Parallel progression

information in a document and the new information can be understood. A user query is more likely to use terms that match the index terms of the early thematic or foundational parts of the text but is more likely to be satisfied by the information or answers in the latter rhemic part of the text.

2.3. Legal Structure

In Refugee Law there are no firm decision guidelines or a sentencing model as in the criminal domain (Nash, 1991). However there are factors that have to be considered and addressed in determining refugee status. The primary standard of refugee status today is that derived from the 1951 Convention relating to the Status of Refugees and amended by the 1967 Protocol relating to the Status of Refugees. A Convention refugee is a person who is outside her country because she reasonably believes that her civil or political status puts her at risk of serious harm in that country, and that her own government cannot and will not protect her (Hathaway, 1991). The definition comprises five essential elements, each of which must be established before status is appropriately recognised. The five elements are:

- Alienage – only persons who have left their country of nationality, or in the case of stateless persons, their place of habitual residence;
- Well-founded fear – It is not enough to believe herself to be in jeopardy, there must be some objective facts to provide a concrete foundation for the concern; Objectively maybe:
  - forward looking assessment of risk;
  - Mr. Justice Stevens reference to the sufficiency of a 10% chance of persecution;
  - Test based on country's human rights record;
  - Claimant's testimony;
  - Evidence of individualized past persecution;
  - Evidence of harm to persons similarly situated;
  - Generalised oppression;
- Persecution – focuses on the existence of persistent harassment by or with the knowledge of the state of origin. It involves "constant infliction of some mental or physical cruelty", "persistent or urgent efforts to harm or cause to suffer" and "pursuit with enmity", such as to provoke "an irrepressible fear of asking the authorities ... for protection".
- Nexus to race, religion, nationality, membership in a particular social group, or political opinion;
- Cessation and Exclusion - Serious criminals and others who exhibit disregard for the purposes of the United Nations may face possibility of persecution in their state, but are not refugees.

3. Query Structure and User model

The query structure used is shown in Figure 2. It is conceivable that a query might not contain the JURISDICTION or the LEGISLATIVE FRAMEWORK and therefore alternatives are possible. In fact RRT members (judges) indicate that the JURISDICTION component would only be important in cases where jurisdiction is in issue, that is in about 1% of the cases. The LEGISLATIVE FRAMEWORK is usually the same for all cases and therefore this component is unlikely to be useful in retrieval. They see the ability to match component-wise as being useful and suggest that the important features for the retrieval of similar cases are "the applicant's claims and evidence", the country of origin, and the current situation in that country. Information about the current situation in the country, persecuted groups and ideological issues are collected from research, amnesty international, united nations and other sources. The discussion of these factors is usually present in the DISCUSSION AND FINDINGS section of a case. A member (at least one prospective user) would like to retrieve similar cases by matching on as many of these components as are known at the time. The country of origin is usually given in the BACKGROUND section, the applicant's claims and evidence in the APPLICANT'S CASE and information about the current situation in the country of origin in the FINDINGS section.

4. Case structure, the user and the retrieval model

The RRT transcripts are written to report the judgement of the tribunal to the applicant and as such attempt to present the applicant's details, the law, the reasons for the findings and decisions reached. The way in which the retrieval model relates to the case structure is not obvious and perhaps the simplest model that is suggested by analogy with retrieval techniques used in CBR is the nearest neighbour model. This is essentially weighted linear combination of case component similarities. This approach behaves in much the same way that the cosine measure of similarity (Salton, 1983) functions, in that it assumes that the components are independent contributors to similarity evidence that should be collected and pooled. When the conditional dependence of components on other components needs to be modelled inductive techniques and decision trees tend to be used (Quinlan, 1986), however there is a requirement of more exact matches at the component level. There has been previous work (Yearwood, 1996a,1996b) in using linear regression to learn the component weighting in a linear retrieval function that optimises retrieval of similar cases. Part of this work examined some non-linear retrieval functions but has not been able to indicate the general form that an optimal retrieval function would take based solely on case structure. In fact this general optimisation problem is mathematically intractable. Domain knowledge and user models may be able to suggest better models.

In seeking the form of a retrieval function we could consider the intentionality of the documentation, domain knowledge as well as how users actually go about querying such a case-base. From a discourse analysis point of view, the order of case components might give an indication of the likelihood of dependence. It is more likely that later components will depend on the earlier components. So in the RRT cases it is more likely that FINDINGS will depend on BACKGROUND than LEGISLATIVE FRAMEWORK. In fact most of the subsequent components have a discourse dependence on the BACKGROUND component. This is reinforced from a user model. The fact that users first seek cases which match on country of origin is not only useful domain knowledge but also highlights the importance of this feature in the case structure or the importance of the individual case component containing this information (BACKGROUND). Because members use this strategy there is the suggestion that a retrieval function should incorporate matching on country of origin first.

It is not clear how the conditional dependence of case components may be automatically learnt especially when there is not a strong uniform case structure across the database. In fact it is not clear what form cases would have to take to meet the combined goals of informing the applicant, keeping a record and serving as retrievable reusable documents for future case advice.

The procedure of first seeking cases that match on country of origin (of the applicant) suggests a filtering function but given that we have similarity evidence for each component other functions are possible. One model for implementing case component dependency that might be considered, is a product of the linear retrieval function with the dependent component similarities. So rather than use a gate function or switch use a modulation by the component similarities. For example if BACKGROUND is important as a condition to be matched then consider retrieval functions of the form

$$\sim_{Backgnd}(\beta_B \sim_{Backgnd} + \beta_{Apcase} \sim_{Apcase})$$

5. Indexing

One of the hypotheses that is being tested here is that structured localised indexing will be more effective for case retrieval than global indexing. Our baseline indexing model consists of indexing the text cases by terms which consist of all non stop word, word stems in a case, with importance weights the tf.idf weights. tf is the term frequency in the whole case and idf is log(1 + N/ df). A retrieval function based on matching with the whole case is based on this indexing model.

Another approach to indexing a case is to argue that only the terms used in a case component should be used to index that component. The case component is in fact a document whose scope is reduced or more focussed than that of the whole case and this may be captured more precisely by a more localised indexing model. The tf.idf scheme of attaching weights to indexing terms can again be used but this time tf indicates the frequency of occurrence in this particular component of the case. The idf part is calculated as $\log(1 + N/ df_i)$ where N is the collection size. (Just as improved similarity measures have been developed based on the 'verbosity hypothesis' (Robertson, 1994) and the 2-Poisson model of indexing (Harter, 1975) it is possible that more appropriate indexing models for case components will be developed based on the 'scope hypothesis').

6. Framework and Methodology

The starting point is the matching of an unstructured query consisting of the APPLICATION, BACKGROUND, JURISDICTION, LEGISLATIVE FRAMEWORK and APPLICANT'S CASE with an unstructured case consisting of all seven components shown in Figure 2 which is indexed as described above. The relevance of a retrieved case is not judged on its usefulness in assisting with the query case or its subjective similarity on most features as in (Yearwood, 1996a) but simply on whether its outcome in terms of refugee status is the same as the outcome in the query case. (This can be most inappropriate as it may also be the case that a lawyer will want the most similar case with the opposite outcome (Ashley, 1991).) Our underlying assumption then, is that similar cases should have similar outcomes. There is no gradation in similarity of outcomes as in previous work where a more continuous relevance scale was appropriate (Yearwood, 1996b).

Users are interested in retrieving a small number of cases which are similar enough to be suggestive of arguments, situations and findings that are relevant to the query case. They are generally trying to look at a few good cases and hence it is appropriate to concentrate on the proportion of retrieved cases that are relevant (precision) at low document levels.

There are only three possible tribunal outcomes. The decision is either: a refugee (IAR), not a refugee (NAR) or no jurisdiction (NJU).

Experiment 0. Counting the Outcomes

A script was used to detect the outcome of cases from THE ORDER section or CONCLUSION as either: is a refugee, is not a refugee, no jurisdiction or the script could not tell. In about one third of the cases the script could not tell and these were manually detected. For the 940 RRT cases the following counts were made: 420 cases out of 940, 45%, have the ruling Is a Refugee (IAR), 510 cases out of 940, 54%, have the ruling Not a Refugee (NAR), 10 (1%) are no jurisdiction cases.

7. Experiments

Experiment 1. Whole Case match

A whole case consists of (Application + Background + Jurisdiction + Legislative Framework + Applicant's Case + Findings + Conclusions). A full query consists of (Application + Background + Jurisdiction + Legislative Framework + Applicant's Case). Cases were indexed by their word stems after removal of a substantial number of stop words.

Fifty cases were randomly selected and manually structured into the above sections as queries. All other cases were transformed into this uniform sectional structure by a semi automatic process. The queries were then matched against the database using the cosine measure for ranking. The query case is removed from the rankings. The RRT cases frequently refer to sections of the Migration Act and often these section numbers were removed by the stop word removal program. Another experiment(1b) was performed in which this did not occur but the results (Table 1) show no real distinction between using the section and article references and not doing so.

	Document level						
	1	2	3	5	10	15	20
Experiment 1	Whole query matched with whole case						
1a	0.660	0.660	0.667	0.684	0.652	0.652	0.654
1b	0.660	0.650	0.673	0.680	0.658	0.652	0.656

Table 1: Average precision with whole query (A,B,J,L,AC) and whole case. Stopping & Stemming.

Experiment 2: Individual query component evidence

Each query is decomposed into its individual components (A, B, JU, LF, AC) and matched against whole cases. Therefore each case is indexed by all of the stemmed non-stop words occurring in the case. The argument for matching query component with full case might be based on the fact that the topicality to some extent will be selected in the matching process anyway. If we consider a term occurring in the query then depending on which indexing vocabulary we have (i.e one based on the vocabulary of just component documents or one based on whole documents) the weighting changes and therefore the computed similarities.

	Document level						
	1	2	3	5	10	15	20
Experiment 2	Query component matched with whole case						
2a Application	0.375	0.417	0.410	0.421	0.481	0.467	0.472
2b Background	0.476	0.500	0.500	0.500	0.517	0.494	0.515
2c Jurisdiction	0.289	0.329	0.316	0.384	0.408	0.421	0.411
2dLegal Framework	0.455	0.466	0.492	0.486	0.470	0.473	0.465
2e Applicant's Case	0.652	0.663	0.681	0.643	0.602	0.593	0.591

Table 2: Average precision for part query and whole case.

Experiment 2 results: **Experiment 2 indicates precision figures that largely agree with the intuition, that information relating to the applicant is the most important factor. Formalities such as the Jurisdiction and Application contribute less. It is also interesting to note that if the rule of, always NAR was applied then the success rate would be 54%. If the rule was to use the outcome of a randomly selected document then the success rate would be 0.54 x 0.54 + 0.45 x 0.45 + 0.1 x 0.1 = 0.5. The results above suggest that APPLICATION and JURISDICTION (and possibly LEGISLATIVE FRAMEWORK) are not good evidence.**

Experiment 3: Combining individual query components. Component vs whole case.

In this experiment similarities of individual case components matched with whole cases are combined according to the precision weights[1] (at the 1 document level). So Experiment 3a uses the formula

$\sim = 0.17 \sim_{App} + 0.21_{Backgnd} + 0.13 \sim_{Juris} + 0.20 \sim_{Legal} + 0.29 \sim_{ApCase}$

Experiment 3b uses all equal weights

$\sim = \sim_{App} + \sim_{Backgnd} + \sim_{Juris} + \sim_{Legal} + \sim_{ApCase}$

Experiment 3c uses the precision figures for the 20 document level as weights. The formula is

$\sim = 0.19 \sim_{App} + 0.209 \sim_{Backgnd} + 0.167 \sim_{Juris} + 0.192 \sim_{Legal} + 0.24 \sim_{Apcase}$

	Document level						
	1	2	3	5	10	15	20
Experiment 3	Combining query component vs. case sililarities						
3a 1 Doc. precisions	0.680	0.650	0.607	0.616	0.602	0.608	0.599
3b Equal Weighting	0.640	0.610	0.593	0.580	0.576	0.572	0.570
3c 20 Doc. precisions	0.640	0.630	0.607	0.596	0.596	0.597	0.593

Table 3: Average precision for combining query component vs. whole case similarities based on average precision scores from Experiment 2.

Experiment 3 results:

Experiment 3 indicates some improvement at the 1 document level only.

1) These are normalised by the sum of the precision values at the 1-document level.

Experiment 4: Combining individual case components. Component vs component.

There are several ways in which case components may be combined. Previous work by Yearwood and Wilkinson (Yearwood, 1996a, Yearwood, 1996b) has considered combining case component similarities according to precision weights or based on linear regression models. Experiment 4a combines case component-component similarities using precision weights from Experiment 2 at the 1-document level. So

$$\sim(Q,C) = 0.17\sim_{App} + 0.21\sim_{Backgnd} + 0.13\sim_{Juris} + 0.20\sim_{Legal}$$

The results are shown in Table 4.

Experiment 4b uses a training set of 50 randomly selected queries from the database and for each query 200 random documents are selected. Case component-component similarities are calculated for each of these 10,000 pairs and judgements made. A Logistic regression model is computed using the 5 independent variables (the component-component similarities for APPLICATION, BACKGROUND, JURISDICTION, LEGISLATIVE FRAMEWORK and APPLICANT'S CASE). The model is logistic in

$$-0.101 + 0.1671\sim_{App} + 0.2414\sim_{Backgnd} - 0.2847\sim_{Juris} = 0.0802\sim_{Legal} + 0.3161\sim_{ApCase}$$

In this model the Chi-Square test is significant at the 0.0004 level and the classification table indicates little better than random classification based on this model. A look at correlations between variables indicates that there are significant correlations between most of the independent variables. The strongest correlations between independent and dependent variables are between BACKGROUND and Relevance (0.0287) and APPLICANT'S CASE and Relevance (0.0196). There is also a strong negative correlation between JURISDICTION and Relevance (-0.0241). The strongest correlation between independent variables is between BACKGROUND similarity and LEGISLATIVE FRAMEWORK similarity (0.4816). Individual single predictor logistic models indicate that APCASE BACKGND and JURIS are the most significant with the coefficient in JURIS being negative.

Experiment 4c uses the same training set as 4b but takes a stepwise approach (Forward Stepwise Conditional) with Logistic Regression. It is interesting to consider the development of this model. The procedure starts with a model which is simply constant(Overall chi-sq is 50.32%) and then adds in the next most highly correlated variable BACKGROUND (chi-sq 51.37%). If a variable were now to be removed it would be BACKGROUND. The next added variable is JURIS (Overall chi-sq 50.72%). The variable that would be removed is now the constant. The next added variable is APCASE (Overall chi-sq is 50.66%). The interesting thing to note is that overall the models are becoming poorer, however, in the last model with A PCASE the percentage of 1s correctly predicted has risen from 0, to 34, to 33 to 43. This is also higher than in the full model in Table 4. In the interest of increasing the percentage of 1s for the top end of the distribution. The model generated is logistic in

$$0.3321\sim_{Backgnd} - 0.2439\sim_{Juris} + 0.3659\sim_{ApCase}.$$

Many factors now suggest the exclusion of JURISDICTION from the model but regression is not excluding it.

Experiment 4d uses the training set of 4b but leaves APP and JURIS out of the Logistic model. The model is logistic in

$$0.2693\sim_{Backgnd} + 0.0210\sim_{Legal} + 0.2351\sim_{ApCase}.$$

Experiment 4e uses the training set of 4b and logistic regression with BACKGROUND as a gate. The model is logistic in

$$0.1293 \sim_{Backgnd} + 1.4696 \sim_{B} \sim_{ApCase}$$

Experiment 4f uses the training set of 4b and logistic regression on BACKGROUND and APCASE only. The component evidence, user opinion, and in some ways regression now suggest the exclusion of all components except BACKGROUND and APPLICANT'S CASE. The model is logistic in

$$-0.0905 + 0.2842 \sim_{B} + 0.2450 \sim_{ApCase}.$$

	Document level						
	1	2	3	5	10	15	20
Experiment 4	Combining component vs. component similarities						
4a 1Doc. Precision weights	0.740	0.660	0.607	0.576	0.576	0.577	0.582
4b Logistic regression	0.620	0.610	0.593	0.548	0.602	0.589	0.579
4c Stepwise logistic	0.660	0.560	0.580	0.572	0.566	0.561	0.563
4d Logistic (no juris or app)	0.700	0.680	0.653	0.596	0.578	0.568	0.554
4e Logistic (Bgnd∗ (Bgnd + Apcase)	0.720	0.650	0.613	0.580	0.558	0.541	0.540
4f Logistic in Bgnd & Apcas	0.700	0.660	0.647	0.612	0.580	0.563	0.550

Table 4: Average precision for combining component similarity evidence.

Experiment 4 results

Combination of case component similarities according to a linear model based on 1 document precisions from Experiment 2 has produced an improvement over the whole case match at the lower document levels. The improvement is not significant when a Wilcoxon test is carried out the p-value is 0.222. The combination models based on logistic regression and linear regression achieve exactly the same results. Stepwise logistic regression based on a random 200 documents performs similarly to the whole case model at the 1 document level. Experiment 4d produces an improvement over experiment 1b at the 1 and 2 document levels but thep-value is 0.32 in both cases, so this is not a significant improvement. The incorporation of some user knowledge (specifically the importance of country of origin and hence BACKGROUND) by the gate or product model shows some promise.

Experiment 5: Using derived components in the combination process.

In experiment 4 with component-wise matching much of each case in the case-base is not used in the retrieval process. Each case hasFINDINGS and CONCLUSIONS and there is nothing to match in the query component. In (Yearwood, 1996a) derived components were used in the matching process.

When the FINDINGS component of the best whole case match was used as a derived FINDINGS section 'dFINDINGS' the regression coefficient is negative, e.g β_{ApCase} = 0.2872 , $\beta_{Backgnd}$ = 0.2891 , β_{dFIND} = - 0.224 and const = -0.0795. A simple regression of the outcome against dFINDINGS also gives a negative coefficient.

Experiment 6: Combining with global evidence

The component-wise techniques are showing some promise at the low document level end. The work by Katzer and others suggests that if the document sets retrieved by different strategies are different then there is something to be gained by combining evidence (Katzer, 1982). Therefore combination with global matches would possibly improve retrieval effectiveness.

Experiment 6a: Uses the same training set as 4b. The model is Logistic in

$$-.04 + 0.014 \sim_{ApCase} + 0.1395 \sim_{Backgnd} + 1.2023 \sim_{Wc}.$$

Experiment 6b: Uses the same training set as 4b. It combines the use of BACKGROUND as a gate with global evidence. Logistic in

$$-0.03829 + 0.0554 \sim_{Backgnd} \sim_{Backgnd} + 0.9235 \sim_{Backgnd} \sim_{ApCase} + 1.0764 \sim_{Wc}$$

	Document level						
	1	2	3	5	10	15	20
Experiment 6	Combining component and whole case similarities						
6a	0.700	0.680	0.673	0.680	0.642	0.651	0.657
6b	0.740	0.700	0.673	0.652	0.624	0.629	0.630

Table 5: Average precision for combining component-component with whole case.

Results for Experiment 6

For 6b, the Wilcoxon test on the 1 document level gives a p-value of 0.117. the t-test gives a p-value of 0.80. So this result is significant at the 8% level.

8. Making use of the implicit Legal structure

The legal structure in an RRT case document is not explicit and therefore it is not straightforward to exploit structure based on the major legal considerations. The area of Refugee law is sufficiently discretionary as to encourage models of similarity based on matching these major components and possibly avoid the detailed arguments and their evolution.

One way of extracting case components corresponding to the four most obvious legal considerations: Alienage, Persecution, Well founded fear and Nexus to race, religion etc. is to extract from each case a chunk of text based on a match with text that describes each of these components. The text used as a description of each of these legal components was taken from the LEGISLATIVE FRAMEWORK section of some cases. The matching of this text was actually with the texts of whole cases minus the JURISDICTION and LEGISLATIVE FRAMEWORK sections so that simple restatements of the law were not regenerated. The text matching was carried out using the vector space model with the usual cosine measure of similarity (Salton, 1983). The issue of what quantity or portion of text constitutes a match arises, as there seemed to be little uniformity across different RRT cases in terms of the numerical values of similarity. (Lewis also makes this point in the case of probabilistic models where the Evaluation strategies are not tuned to the purpose and context of the exercise (Lewis, 1995)). Preliminary examination of the similarities at the paragraph level suggested no clear strategy other than selecting a set number of top ranking paragraphs.

Examination of the top paragraphs retrieved in some cases suggested that a reasonable starting point would be: the top 5 paragraphs for Alienage and Nexus and the top 10 paragraphs for Persecution and Well founded fear. It is probably better to take an experimental approach and see how each performs at choices of 5, 10, 15 and 20 paragraphs, however, the issue here is whether this structured representation and technique can act as a basis for improving retrieval.

8.1. The matching model

Only the four main factors in determining refugee status are extracted for matching in this model. Figure 5 illustrates the overall process of extracting the main legal factors of Alienage, Persecution, Well-founded fear and Nexus and then the component-wise matching.

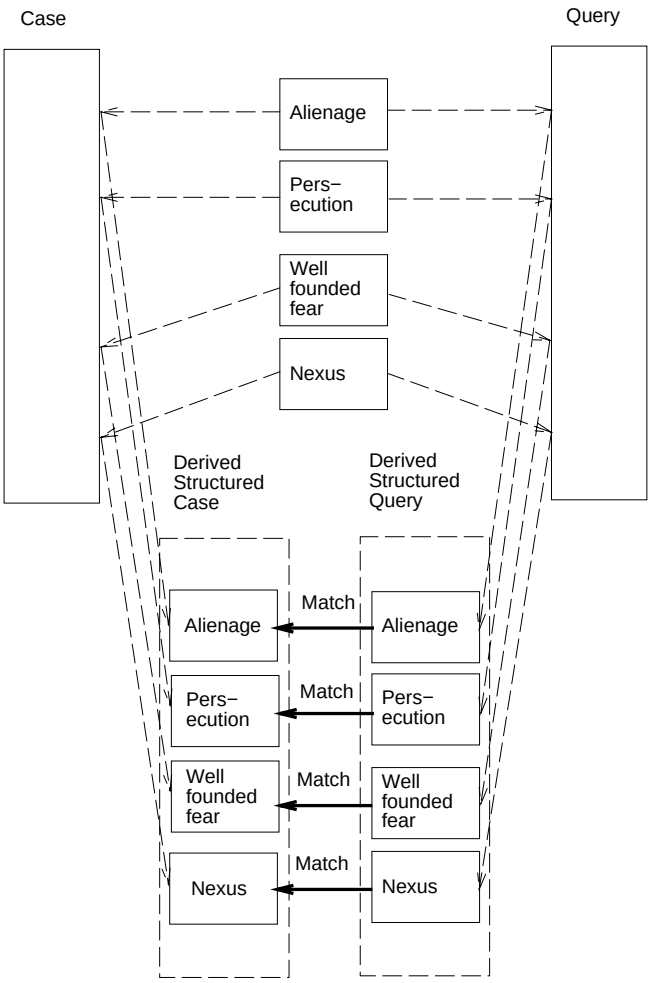

Figure 5: Capturing and using the legal structure in matching

8.2. Experiments on Legal Structure

Experiment 7: Individual case component evidence

Each query is decomposed into its individual components (Alienage, Persecution, Well-founded fear and Nexus) using the process described above and matched against the corresponding component. Therefore each case is indexed by all of the stemmed non-stop words occurring in the component. The indexing vocabulary used is one based on the vocabulary of just the component documents.

Experiment 7 results:

The results of the individual component versus component matches are shown in Table 6. It is interesting to note the performance of the Well-founded fear factor which at the 1-document level is as good as has been achieved. The performance at the higher document levels falls more rapidly than in the result of Experiment 6b.

	Document level						
	1	2	3	5	10	15	20
Experiment 7	Derived legal component matched with derived legal component						
7a Alienage	0.580	0.620	0.600	0.592	0.586	0.576	0.567
7b Percution	0.660	0.680	0.700	0.676	0.656	0.643	0.640
7c Well-founded fear	0.740	0.710	0.700	0.672	0.622	0.624	0.612
7d Nexus	0.700	0.640	0.620	0.620	0.594	0.584	0.576

Table 6: Average precision for legal component vs. legal component.

Experiment 8: Combining individual case components matches.

In this experiment similarities of individual case component matches are combined first according to the precision weights from Experiment 7 (at the 1 document level) and then equally. So

Experiment 8a uses the model

$$\sim(Q,C) = 0.58 \sim_{Alienage} + 0.66 \sim_{Persecution} + 0.74 \sim_{WFF} + 0.70 \sim_{Nexus}$$

Experiment 8b uses all equal weights

$$\sim(Q,C) = \sim_{Alienage} + \sim_{Persecution} + \sim_{WFF} + \sim_{Nexus}$$

Experiment 8c uses logistic regression with a training set of 200 random documents for each query. The model is

$$\sim Q,C) = 0.2189 \sim_{Alienage} + 0.9448 \sim_{Persecution} + 1.6795 \sim_{WFF} + 0.2738 \sim_{Nexus}$$

Experiment 8d uses stepwise linear regression with a training set of 200 random documents for each query. The model is given by

$$\sim(Q,C) = 0.2818 \sim_{Persecution} + 0.4458 \sim_{WFF}$$

Experiment 8 results:

The results are shown in Table 6. Again it can be observed that precision weighting has produced better performance than equal weighting (although the difference is not statistically significant). The regression techniques and in particular stepwise regression produce the best results with increased depth over the WFF result from Experiment 7. This result is better than the best result with the explicit case structure.

	Document level						
	1	2	3	5	10	15	20
Experiment 8	Combining legal component vs. legal component similarities						
8a 1 Doc precisions	0.700	0.670	0.647	0.684	0.660	0.639	0.623
8b Equal Weighting	0.680	0.670	0.660	0.668	0.664	0.639	0.621
8c Logistical regression	0.700	0.720	0.693	0.660	0.656	0.643	0.632
8d Stepwise	0.740	0.710	0.680	0.664	0.638	0.636	0.632

Table 7: Average precision for combining legal component vs. legal component.

9. Conclusion

The basic hypothesis being tested is whether combining evidence from individual case components can improve retrieval effectiveness. However it is not clear with this database of RRT cases that either of the techniques used previously has been solely successful. The strategy of using precision weights in work with an Occupational Health and Safety database (Yearwood, 1996a) has notgenerated significant improvement here. Linear regression was used successfully with a highly structured nursing database in (Yearwood, 1996b) where there was no overlap in the index terms used in each component; the case base was large; the learning sample was large and the outcome variable was approximately continuous. The model was built using stepwise regression. In the RRT database there is (as with any natural language data) a huge overlap in the index terms used in case components; the learning sample is much smaller (this need not be, but it seems appropriate to put it to the test as relevance judgements are always expensive); the outcome variable is dichotomous. The dichotomous nature of the outcome variable indicates the use of Logistic regression (Gey, 1994). However this strategy alone really only led to results that were comparable with the baseline whole case match. However there were aspects of the overall statistical and regression analysis that led to decisions about which components should be excluded and which included.

Other considerations such as the user and the nature of the text in terms of its discourse form have been interpreted to influence the form of the retrieval function. Both statistical results and user information suggested the exclusion of APPLICATION, JURISDICTION and LEGISLATIVE FRAMEWORK. The user model and a little understanding of these cases in discourse terms suggest that the BACKGROUND is an underpinning or thematic component (although this is not seen in the similarity diagrams).

Finally the results of combining case component evidence and global evidence produce significant results at the 8% level for the 1-document level. It should be noted however that there has not been any examination of the content similarity of the cases to the queries. The results are simply assessed in terms of similar outcome. The use of larger learning samples may be indicated and it seems that techniques based on small 'top-end' training have again not been successful.

9.1. Implicit Legal Structure - Automatically Derived

Manually identifying and extracting the main legal factors in these cases would be an expensive task. The technique used here is a simple first stage automatic one and shows promise. The results of the subsequent case matching using these components are comparable with and slightly better than the best results from combining the explicit structural components even without the incorporation of global evidence. The results at the higher document levels are nearly as good as the original whole case match and the lack of improvement may simply reflect the fact that the amount of text used in the derived components is small relative to the amount of text available in the cases.

10. References

Kevin D. Ashley. Reasoning with cases and hypotheticals in HYPO. International Journal on Man-Machine Studies, 34: 753-796, 1991.

B. T. Bartell, G. W. Cottrell, and R. K. Belew. Automatic combination of multiple ranked retrieval. In Proceedings of the 17 th International Conference on Research and Development in Information Retrieval, pages 173-181, 1994.

Fredric C. Gey. Inferring probability of relevance using the method of logistic regression. In Proceedings of the 17 th International Conference on Research and Development in Information Retrieval, ACM, 1994.

Stephen P. Harter. A probabilistic approach to automatic keyword indexing: Part 1. On the distribution of specialty words in a technical literature. Journal of the American Society for Information Science, 26: 197-206, 1975.

James C. Hathaway. The Law of Refugee Status. Butterworths, 1991.

Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In Proceedings of the 16 th International Conference on Research and Development in Information Retrieval, pages 59-68. Association for Computing Machinery, 1993.

W. John Hutchins. On the problem of 'aboutness' in document analysis. Journal of Informatics, 1(1), 1977.

J. Katzer, M. J. McGill, J. A Tessier, W. Frakes, and P.Dasgupta. A study of the overlap among document representations. Information Technology: Research and Development, 1(2):261-274, 1982.

David D. Lewis. Evaluating and optimising autonomous text classification systems. In Proceedings of the 18 th International Conference on Research and Development in Information Retrieval, pages 246-254. ACM, 1995.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. Proceedings of the 17 th International Conference on Research and Development in Information Retrieval, ACM, 1994.

Elke Mittendorf and Peter Schäuble. Document and passage retrieval based on hidden Markov models. Proceedings of the 17 th International Conference on Research and Development in Information Retrieval, ACM, 1994.

A. Nash. Sentencing Act 1991. Butterworths, Melbourne, 1991. J. R. Quinlan. Induction of Decision trees. Machine Learning, 1:81-106, 1986.

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the 17 th International Conference on Research and Development in Information Retrieval, ACM, 1994.

Gerard Salton, James Allan and Amit Singhal. Automatic text decomposition and structuring. Information Processing and Management, 32(2): 127-138, 1996.

Gerard Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

John Yearwood and Ross Wilkinson. Case-based retrieval of highly structured cases using Text Representation and Retrieval. In Proceedings of the First Australian Document Computing Symposium, 1996.

John Yearwood and Ross Wilkinson. Combining case compo nent evidences for text based retrieval of cases: An Experimental Study. In Proceedings of the Nineteenth Australasian Computer Science Conference, 1996.