SELECTIVE RELEVANCE FEEDBACK
USING TERM CHARACTERISTICS

Ian Ruthven[1] *and Mounia Lalmas*[2]

[1]*Department of Computing Science, University of Glasgow, G12 8QQ Scotland*
[2]*Informatik VI, University of Dortmund, Dortmund D44227, Germany*
igr@dcs.gla.ac.uk, mounia@ls6.cs.uni-dortmund.de

Abstract: This paper presents a new relevance feedback technique; selectively combining evidence based on the *usage* of terms within documents. By considering how terms are used within documents, we can better describe the features that might make a document relevant and thus improve retrieval effectiveness. In this paper we present an initial, experimental investigation of this technique, incorporating new and existing measures for describing the information content of a document. The results from these experiments positively support our hypothesis that extending relevance feedback to take into account how terms are used within documents can improve the performance of relevance feedback.

1. Introduction

The particular characteristics of digital libraries - large, diverse collections - mean that the access methods to these collections must support *how* users search for information. In particular, access methods must provide a wider range of techniques to allow the user to specify what makes an object, such as a document, relevant. In this paper we investigate how relevance feedback can better describe a user's information need by taking into account how terms are used in documents.

Most relevance feedback algorithms attempt to bring a query closer to the user's information need by reweighting or adding/deleting query terms. The implicit assumption is that we can find an optimal combination of weighted terms to describe the user's information need at that point in the search. However, relevance as a user judgement is not necessarily dictated only by the presence or absence of terms in a document. Rather it is a factor of what concepts the terms represent, the relations between these concepts and how they relate to the information in the document. Looking at studies such as (Barry and Schamber, 1998) it is clear that current models of relevance feedback, although successful at improving recall-precision to an extent, are not very sophisticated in expressing what makes a document relevant to a user. (Denos et al, 1997), for example, make the good point that although users can make explicit judgements on why documents are relevant, most systems cannot use this information to improve the search.

Users judgements are affected by a variety of factors; relevance feedback algorithms, on the other hand, only consider frequency information or the presence or absence of terms in documents. They do not look further to see what it is about terms that indicate relevance, ignoring information on how the term is *used* within documents. For example a document may only be relevant if the terms appear in a certain context, if certain combinations of terms occur or if the main topic of the document is important. Extending feedback algorithms to incorporate the *usage* of a term within documents would not only allow more precise querying by the user but also allows relevance feedback algorithms to adapt more subtly to users' relevance judgements.

This paper describes an initial, experimental investigation of this approach by considering relevance feedback as a process of selection: selecting which characteristics of a term (e.g. frequency, context, distribution within documents) should be used to retrieve documents. The following sections outline our general methodology (section 2), the data we used in our experiments (section 3), definition of the term characteristics we used to describe the use of terms within documents (section 4), experiments on combining evidence of term use and relevance feedback (sections 5 and 6) and our conclusions (section 7).

2. Methodology

Our intention behind the set of experiments described in this paper is twofold: first to demonstrate that taking into account how terms are used within documents (which we refer to as *term characteristics*) can improve retrieval effectiveness; secondly that it is possible, for each query, to select an optimal set of characteristics for retrieval based on the relevance assessments. The second point is the main one considered in this paper. We are not only asserting that considering how terms are used *can* improve retrieval, but that the characteristics that *will* improve retrieval will vary across queries. For example, for some queries the context in which the query terms appear will be important, whereas for other queries it may be how often the query terms appear. For each query, then, there will be a set of characteristics that will best indicate relevance.

To investigate these issues, we have designed a set of restrictive experiments. Our experiments are restricted in that we are only attempting a form of *precision enhancement*. Rather than scoring each document according to one or more criteria, we retrieve a number of documents and then re-rank the retrieved documents according to various criteria. This allows us to run a large number of experiments quickly but it also means that we are manipulating the part of the ranked list that most users will be investigating - the top part.

In the experiments described in section 5 - section 7, we retrieved at most 100[1] documents using the *idf* weighting function (Sparck Jones, 1972). The *idf* function was chosen as an experimental baseline, because it only provides information about a term relative to the collection as a whole; it has the same weight in each document in which it appears. It does not supply any information about a term that is specific to individual documents. The three term weighting functions described in section 4, all provide information that is specific to the document in which they occur. In particular they provide information on how terms are used within the document and will be used in our experiments to differentiate between documents.

3. Data

In these experiments we used the Wall Street Journal (1990-92) (WSJ) collection from TREC-5 (Voorhees and Harman, 1996) and the Financial Times (FT) collection from the TREC-6 (Voorhees and Harman, 1997) set of collections. The details of these collections are summarised in Table 1.

Table 1: Details of collections used

Collection	**FT**	**WSJ**
Number of documents	204790	74580
Number of queries used[2]	38	30
Average words per document	215	283

Each collection comes with fifty TREC topics, each describing an information need and which criteria relevant documents should fulfil to be assessed relevant. A TREC topic has a number of sections, (see Figure 1 for an example topic). In our experiments we only

[1] We do not retrieve exactly 100 documents for each query as there may be less containing any of the query terms.

[2] Although each collection has 50 topics, not all topics have relevance assessments retrieve any relevant documents in the first 100 documents retrieved. Theref topics/queries in our experiments.

used the short Title section as a query, as using any more of the topic description may be an unrealistic user query.

Figure 1: Example of a TREC topic

Number: 301
Title: International Organized Crime
Description:
Identify organisations that participate in international criminal activity, the activity, and, if possible, collaborating organisations and the countries involved.
Narrative:
A relevant document must as a minimum identify the organisation and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organisation(s) involved would not be relevant.

4. Characteristics

Here we outline three alternative ways of describing term importance in a document: *term frequency* (how often a term appears), *thematic nature* (how a term is distributed within a document), and *context* (proximity of a query term to another query term). The latter two are very simple methods to give a rough estimate of the behaviour of the characteristics of a term.

Term frequency

Including information about how often a term occurs in a document - term frequency information - has often been shown to increase retrieval performance (Harman, 1992) . For this experiment we used the formula $tf_d(t) = \ln(occs_t)/\ln(n_{unique})$ where occs_t is the number of occurrences of term *t* in document *d* and $\mathrm{occs}_{\mathrm{unique}}$ is the number of unique term occurrences in *d*.

Theme

Previous work by e.g. (Hearst and Plaunt, 1993) and (Paradis and Berrut, 1996), demonstrates that information about the topical or thematic nature of the document can improve retrieval. Hearst and Plaunt present a method specifically for long documents, whereas Paradis's method is based on precise conceptual indexing.

We present a simple term-based alternative based on the distribution of term occurrences within the document. This is based on the assumption that the less evenly distributed the occurrences of a term are in the document, then the more likely the term is to correspond to a localised discussion in the document, e.g. a topic in one section of the document only. Conversely, if the term's occurrences are more evenly spread throughout the document, then we may assume that the term is somehow related to the main topic of the document. Unlike Hearst and Plaunt we do not split the document into topics and assign a sub- or main-topic classification, instead we define a *theme* value of a term, which is based on the likelihood of a term to be a main topic. The algorithm which we developed for this is shown in Equation 1.

Equation 1: *Theme* characteristic for term t in document d, where $\mathrm{distr}_d(t)$ is the expected distribution of term in the document, epos_i is the expected position of the ith occurrence of term t, and pos_i is the actual position of the ith occurrence. occs(t) is the number of occurrences of term t in document d. n is the number of query terms in the document.

$$theme_d(t) = (length_d - difference_d(t)) / length_d$$

where

$$difference_d(t) = first_d(t) + last_d(t) + \sum_{i=2}^{n-1} \left| epos_i(t) - pos_i(t) \right|$$

$$\begin{array}{lcll} first_d(t) & = & 0, & if \quad pos_1(t) \le distr_d(t) \\ & = & pos_1(t) - distr_d(t), & ow \\ last_d(t) & = & 0, & if \quad (length_d - pos_n(t) \le distr_d(t)) \\ & = & length_d - (pos_n + distr_d(t)), & ow \end{array}$$

$$epos_i = pos_{i-1} + distr_d(t)$$

$$distr_d(t) = length_d / occs_d(t)$$

(1)

This value is based on the difference between the position of each occurrence of a term and the *expected* positions. Table 2 gives a short example for a document with 1000 words, and five occurrences of term t. First, we calculate whether the first occurrence of term *t* occurs further into the document that we would expect, based on the expected distribution ($first_d(t)$ - line two, equation 1; column 7, table 2). Next we calculate whether the last occurrence of the term appears further from the end of the document than we would expect ($last_d(t)$ - line two, equation 1; column 7, table 2). For the remainder of the terms we calculate the difference between the expected position of a term, based on the actual position of the last occurrence and the expected difference between two occurrences ($\sum_{i=2}^{n-1} |epos_i(t) - pos_i(t)|$ - line two; column 4-6, table 2).

Table 2: Example calculation of *theme* value for a term

length	**occs**	**distr**	**epos**	**pos**	**diff**	**first**	**last**	**difference**	***theme***
1000	5	200	-	100		0			
			300	500	200				
			700	551	349				
			751	553	547				
			753	700	600				
			900				100		
					600	0	100	700	0.3

We then sum these values to get a measure of the difference between the expected position of the term occurrences and their actual positions. The greater the difference between where term occurrences appear and where we would expect them to appear, the smaller the *theme* value for the term. The smaller the difference, the larger the *theme* value for the term.

Context

There are various ways in which one might incorporate information about the context of a query term. For example, we might rely on coocurrence information, information about phrases, or information about the logical structures, e.g. sentences, in which the term appears. Instead we have gone for the simplest option which is to define the importance of context to a query as being measured by its distance from the nearest query term relative to the average expected distribution of all query terms in the document.

Equation 2: *Context* characteristic for term t in document d, where $\mathrm{distr}_d(q)$ is the expected distribution of all query terms in the document, $\mathrm{pos}_d(t)$ is the position of term t and $\mathrm{min}_d(t)$ is the minimum difference from any occurrence of term t to another, different query term.

$$\begin{aligned} &context_d(t) = (distr_d(q) - \min_d(t)) / distr_d(q) \\ &\min_d(t) = \min_{t \neq t'} |(pos_d(t) - pos_d(t'))| \\ &distr_d(q) = length_d / occs_d(q) \end{aligned} \tag{2}$$

All values for *theme*, *context*, *tf* and *idf* are scaled to fall between 0-50 to allow direct comparison of the scores. Each value (*idf*(t), tf_d(t), $context_d$(t), $theme_d$(t)) gives a score to a term describing its importance to the document which can be used to score documents for ranking.

5. Experiment One - retrieval by single characteristic

To test the relative effectiveness of each measure before attempting any combination of characteristics, we ranked each of the documents for a query by each characteristic.

Table 3: Recall-Precision (RP) figures for each characteristic based on re-ranking a maximum of 100 documents for each query

FT				
Recall	***idf***	***tf***	***theme***	***context***
0.100	0.408	0.582	0.441	0.327
0.200	0.384	0.582	0.415	0.316
0.300	0.355	0.530	0.351	0.277
0.400	0.335	0.512	0.333	0.260
0.500	0.309	0.478	0.318	0.241
0.600	0.314	0.435	0.287	0.233
0.700	0.297	0.404	0.268	0.234
0.800	0.294	0.397	0.265	0.230
0.900	0.264	0.349	0.239	0.195
1.000	0.265	0.338	0.225	0.193
average	0.322	**0.461**	0.314	0.251

WSJ				
Recall	***idf***	***tf***	***theme***	***context***
0.100	0.321	0.180	0.248	0.276
0.200	0.333	0.181	0.181	0.256
0.300	0.311	0.146	0.156	0.250
0.400	0.301	0.129	0.160	0.235
0.500	0.292	0.124	0.161	0.203
0.600	0.218	0.100	0.132	0.155
0.700	0.213	0.094	0.136	0.151
0.800	0.193	0.072	0.121	0.141
0.900	0.169	0.065	0.109	0.141
1.000	0.168	0.065	0.107	0.141
average	**0.252**	0.115	0.151	0.195

Table 3 shows the results of the original ranking (*idf*) against each of the characteristics (*theme*, term frequency *(tf)*, *context*) described in section 4. For the FT collection the *tf* scheme works better than all the others, with *idf* better than *theme* and *theme* better than *context*. All differences are statistically significant (paired t-test, $p < 0.05$, holding recall fixed and varying precision). In the WSJ collection the *idf* scheme works best, followed by *context*, *theme* and *tf* (all differences except between *idf* and *theme* are statistically significant). There are two explanations for this difference between the relative effectiveness of the characteristics over the two collections. Either the cut off at 100 documents unnaturally biases in favour of documents that display certain characteristics or these characteristics better describe the relevant set for the queries on these collections. The next section looks at combining the characteristics to test whether combinations of characteristics alter the differences between collections.

6. Experiment Two - retrieval by combination

Our stated hypothesis is that relevant document retrieval will be improved if we take into account more of the characteristics that indicate relevance. These experiments combined all combinations of two and three characteristics. In this set of experiments we simply added the score of each characteristic of each query term that occurred in the document to get the document score.

The results from this experiment are reported in table 4 and table 5. Table 4 (combining two characteristics) shows that, for the FT collection *idf* performance is generally improved by the addition of new evidence (except *context*), *tf* performance is only improved by *idf*, *context* and *theme* performance is improved by the addition of any information. However the only combination of evidence that outperforms ranking by *tf* is *tf* + *idf*.

Table 4: RP for each pairwise combination of characteristic based on re-ranking maximum of 100 documents for each query. The final row (% change) shows the % increase of the combination strategy over the best single characteristic retrieval for that collection. The best average precision is shown in bold.

FT						
Recall	**_idf_ + _tf_**	**_tf_ + _theme_**	**_tf_ + _context_**	**_idf_ + _theme_**	**_idf_ + _context_**	**_theme_ + _context_**
0.1	0.588	0.472	0.386	0.472	0.352	0.491
0.2	0.586	0.476	0.390	0.464	0.346	0.430
0.3	0.538	0.409	0.348	0.432	0.312	0.371
0.4	0.524	0.398	0.309	0.424	0.285	0.360
0.5	0.492	0.369	0.277	0.403	0.266	0.338
0.6	0.440	0.315	0.267	0.349	0.273	0.314
0.7	0.411	0.294	0.257	0.334	0.243	0.287
0.8	0.407	0.286	0.254	0.328	0.241	0.281
0.9	0.376	0.271	0.233	0.287	0.221	0.249
1	0.361	0.257	0.228	0.279	0.222	0.240
Average	**0.472**	0.355	0.295	0.377	0.276	0.336
% Change	**+2.5**	-23.0	-36.0	-18.1	-40.1	-27.0

WSJ						
Recall	**_idf_ + _tf_**	**_tf_ + _theme_**	**_tf_ + _context_**	**_idf_ + _theme_**	**_idf_ + _context_**	**_theme_ + _context_**
0.1	0.495	0.332	0.382	0.351	0.309	0.349
0.2	0.481	0.291	0.352	0.306	0.285	0.325
0.3	0.392	0.229	0.280	0.311	0.270	0.292
0.4	0.365	0.215	0.254	0.267	0.257	0.266
0.5	0.366	0.216	0.237	0.253	0.229	0.255
0.6	0.310	0.174	0.191	0.225	0.177	0.215
0.7	0.268	0.173	0.183	0.221	0.170	0.216
0.8	0.248	0.158	0.170	0.190	0.155	0.194
0.9	0.201	0.146	0.162	0.160	0.154	0.191
1	0.200	0.145	0.163	0.158	0.154	0.190
Average	**0.333**	0.208	0.237	0.244	0.216	0.249
% Change	**+32.1**	-17.4	-5.8	-3.0	-14.2	-1.0

For the WSJ collection *idf* performance is only improved by the addition of term frequency information. *tf*, *theme* and *context* performances are improved by addition of any new information. However, again, the only combination of evidence that outperforms the best single ranking is *tf*+*idf*. In both collections *theme* and *context* are improved by addition of any new evidence but none of the combinations in which they appear outperform all other combinations.

We should note here that it may not be appropriate to treat all evidence (all characteristics) as *equally* important for each query. For example the fact that one term shows a strong thematic relationship may not be as important as the fact that it occurs frequently in a document. To test this we tried scaling the separate characteristics, adding the characteristic score for a term but treating it as less important than the other characteristics, e.g. halving the *theme* value, or doubling the *tf* value. This highlighted two issues: first that treating characteristics differently may improve the retrieval performance but also that it is difficult to get one set of scaling factors that will improve retrieval independent of the combination of characteristics that are being considered. We will discuss this in more detail in section 7.

Table 5: RP for each combination of three characteristic based on re-ranking maximum of 100 documents for each query. The final row (% change) shows the % increase of the three-way combination strategy over the best two-way combination strategy for that collection. The best average precision is shown in bold.

Recall	***tf + idf + theme***		***theme + idf + context***		***tf + idf + context***		***tf + theme + context***	
	WSJ	*FT*	*WSJ*	*FT*	*WSJ*	*FT*	*WSJ*	*FT*
0.1	0.274	0.477	0.270	0.483	0.248	0.390	0.278	0.487
0.2	0.201	0.490	0.225	0.436	0.221	0.399	0.232	0.459
0.3	0.201	0.462	0.222	0.389	0.193	0.360	0.223	0.409
0.4	0.207	0.458	0.202	0.387	0.143	0.326	0.182	0.389
0.5	0.210	0.432	0.203	0.378	0.144	0.315	0.186	0.372
0.6	0.177	0.357	0.169	0.345	0.113	0.306	0.150	0.326
0.7	0.146	0.343	0.139	0.309	0.102	0.271	0.138	0.305
0.8	0.136	0.335	0.134	0.304	0.100	0.270	0.130	0.303
0.9	0.126	0.310	0.132	0.255	0.094	0.248	0.128	0.274
1	0.124	0.294	0.132	0.251	0.095	0.240	0.128	0.265
Average	0.180	**0.396**	**0.183**	0.354	0.145	0.313	0.177	0.359
% Change	-45.88	**-16.18**	**-45.04**	-25.11	-56.31	-33.83	-46.65	-24.00

Table 5 shows that the general performance achieved by combining any three strategies is poorer than that obtained by *tf+idf* in either collection. However the best performance is achieved by the combination of *idf, theme* and another characteristic.

The combination of evidence experiments in this section treated all queries and all documents in the same way; they used the same combination of characteristics to rank all documents for all queries. However, as we have suggested certain characteristics, or combinations of characteristics, may be better suited to certain queries. We investigate this in the next section.

7. Experiment Three - relevance feedback

7. 1 Methodology

In these experiments we performed a series of relevance feedback experiments, selecting which characteristics to use based on the differences between the relevant and non-relevant documents.

Our methodology was as follows:

- take the 10 top documents from the initial *idf* ranking
- calculate for each term the average score for each characteristic in the relevant and non-relevant set, e.g. the average *tf* for term 1 in relevant documents, the average *tf* for term 1 in non-relevant documents.
- select criteria based on the relative averages. Various selection methods were tried, each will be discussed separately in sections 7.3-7.5.
- re-rank the remaining retrieved documents
- calculate recall-precision values using a residual ranking scheme (Chang, Cirillo, and Razon, 1971), to ensure that we are only comparing the effect of each technique on the unretrieved, relevant documents.
- compare the results given, over the same set of documents, by doing no relevance feedback, the results obtained from the best combination of criteria (section 6) and an alternative relevance feedback algorithm, the $F_{4.5}$ method (section 7.2).

This set of experiments was designed to test the hypothesis that some queries or documents will be more suited to certain combinations of characteristics. For example some queries will do better if we take into account, e.g. *tf* or *theme* rather than *context*.

7.2 $F_{4.5}$

We need to compare our technique for relevance feedback against another relevance feedback algorithm. For this we have chosen the $F_{4.5}$ weighting algorithm (Robertson and Sparck Jones, 1976), equation 3 which assigns a new weight to a term based on relevance information. This modified version of the original F_4 technique for reweighting query terms was chosen partly because it has been shown to give good results but also because it does not add any new terms to the query. As our technique also does not add any new terms to the query, we feel this is a fair comparison with which to test our techniques.

Equation 3: $F_{4.5}$ function, which assigns a weight to term *t* for a given query. *r*= the number of relevant documents containing the term *t*, *n* = the number of documents containing *t*, *R* = the number of relevant documents for query *q*, and *N* = number of documents in the collection

$$w_q(t) = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)} \qquad (3)$$

7.3 Feedback 1 - characteristic selection by query

In this experiment we selected for each query which characteristics to use for each query term. The average values were used to decide these characteristics. For example, if the average *context* value for a term was greater in the relevant documents than in the non-relevant documents, then the *context* of that term was taken to be a better indicator of relevance than non-relevance and so contributed to the document score.

In section 6, we mentioned the difficulty of scaling the importance of each characteristic. For example, when combining *context* and *theme*, we may obtain better results by treating *context* as only half as important as *theme* information, but when combining *context* and *tf*, it may be better to treat *context* information as twice as important as *tf* information. It is then not just a matter of how to select which characteristics to combine but also how much evidence from each characteristic to use in calculating the document scores.

There are two solutions to this: either test various scaling factors for each separate combination of characteristics to determine the optimum scaling factors, or try to find an optimum for each characteristic independent of what other characteristics it is combined with. We have chosen the latter approach as it is less complex and we believe that the scaling factor may be derived in relevance feedback. An investigation into this is reported in section 7.5.

Consequently, we tested a variety of scaling factors (e.g. halving the *tf* value, doubling the *context* value, etc.) for each characteristic to produce an optimum performance in relevance feedback. We also used these scaling factors when testing the best combination (*tf* + *idf*), to ensure that we were comparing the optimum relevance feedback performance with the optimum combination performance for these experiments.

Table 6, columns 2 -5 (*Feedback 1*) show the results of this technique compared to the alternative methods outline in section 7.1 (no feedback, $F_{4.5}$ and best combination). The best combination method (column 3) does very well against performing no relevance feedback (column 2). The $F_{4.5}$ method (column 4), although it improves performance over no feedback, does not perform as well as the best combination. This may be because we are using quite small samples for this method. However our characteristic selection method (Feedback1 column 5) outperforms all three (none, *tf*+*idf*, $F_{4.5}$) with an overall average precision increase of 153% on the WSJ collection and 89% for the FT collection. All differences between the RP figures for these four methods are statistically significant for the WSJ collection and, with the exception of no feedback against $F_{4.5}$, also on the FT collection.

Table 6: RP figures for the WSJ and FT collections comparing *idf* (*no feedback)* ranking, best combination (*tf+idf)*, $F_{4.5}$ method and three feedback methods based on term characteristics (*Feedback1*, *Feedback2* and *Feedback3*)

Recall	**WSJ**					
	No Feedback	*tf+idf*	$F_{4.5}$	Feedback 1	Feedback 2	Feedback 3
0.100	0.253	0.366	0.274	0.608	0.456	0.448
0.200	0.223	0.347	0.244	0.580	0.441	0.419
0.300	0.215	0.292	0.246	0.492	0.409	0.414
0.400	0.174	0.294	0.204	0.472	0.365	0.328
0.500	0.171	0.270	0.202	0.463	0.365	0.333
0.600	0.154	0.237	0.179	0.408	0.296	0.314
0.700	0.145	0.173	0.170	0.349	0.293	0.280
0.800	0.140	0.155	0.166	0.327	0.286	0.265
0.900	0.115	0.139	0.142	0.310	0.284	0.262
1.000	0.114	0.137	0.141	0.304	0.278	0.262
Average	0.170	0.241	0.197	**0.431**	0.347	0.333
% Change	0.00	+41.44	+15.44	**+153.11**	+103.75	+95.14

FT						
Recall	No Feedback	*tf+idf*	$F_{4.5}$	Feedback 1	Feedback 2	Feedback 3
0.100	0.339	0.636	0.337	0.734	0.561	0.526
0.200	0.318	0.571	0.314	0.699	0.564	0.512
0.300	0.278	0.492	0.276	0.565	0.392	0.412
0.400	0.271	0.462	0.269	0.546	0.391	0.392
0.500	0.277	0.423	0.274	0.532	0.384	0.362
0.600	0.248	0.337	0.258	0.429	0.320	0.327
0.700	0.239	0.316	0.250	0.417	0.311	0.323
0.800	0.221	0.283	0.230	0.368	0.263	0.285
0.900	0.223	0.273	0.231	0.356	0.266	0.279
1.000	0.222	0.273	0.231	0.349	0.268	0.278
Average	0.264	0.407	0.267	**0.499**	0.372	0.370
% Change	0.00	+54.20	+1.29	**+89.40**	+41.01	+40.15

7.4 Feedback 2 - characteristic selection by document

The previous experiment selectively combined evidence on a query-to-query basis, ranking all documents based on the same set of characteristics for a query. This experiment varies the characteristics also on a document-to-document basis. The intuition behind this is: if a characteristic is indicated as a good indicator of relevance then we should not only bias retrieval of documents which demonstrate this characteristic but suppress retrieval of documents which do not.

We use the same averaging technique as in the previous experiment then for each document compare the characteristic score of each query term against the average. If the characteristic score is greater than the average then we count the score as part of the document score, if not we ignore the evidence. This experiment is, then, a more strict case of Feedback 1. Feedback 1 selected criteria with which to rank all documents, whereas this experiment selects characteristics for a query and then uses them selectively across documents.

Table 6, columns 2 -4, column 6 (*Feedback 2*) show the results of this technique. The method works significantly better than no feedback and F4.5 and significantly worse that the first feedback method in both collections. In the FT collection it works worse than *tf+idf* but the reverse happens in the WSJ. Comparing both feedback methods it maybe the case that this method is too strict and that we should not want to eliminate weak information.

7.5 Feedback 3 - scaling by importance

This final experiment eliminates the scaling factors we introduced in the first two experiments. Instead we use the ratio of the average characteristic value in the relevant to the non-relevant documents, e.g. average *context* score for a term in the relevant documents divided by the average in the non-relevant documents. The intuition behind this is that if a characteristic does not discriminate well over the relevant and non-relevant set then we should not prioritise this information. If the ratio is high then the characteristic may be a good indicator of relevance and should be prioritised.

Table 6, columns 2 -4, column 7 (*Feedback 3*) show the results of this technique. It works in a very similar manner to the second feedback technique: significantly better than no feedback and F4.5 and significantly worse that the Feedback 1 method in both collections. In the FT collection it works worse than *tf+idf* and the Feedback 2 method but in the WSJ it is worse than Feedback2 but better than *tf+idf*. In both collections, Feedback 1 works better than Feedback 2 which, in turn, works better than the Feedback 3 method.

8. Discussion

Our overall research goal is not only to make retrieval more effective but to make a user’s interaction with an IR system more meaningful. In part this may be achieved by increasing the range of ways a user can express or indicate his or her information need. This paper investigates a very particular means of achieving this: by using information on how terms are used within a document to direct relevance feedback. We have shown that *selecting* which term characteristics of use on a query-to-query basis can significantly improve retrieval effectiveness.

We should stress that this is a very initial investigation. The limitations of our experiments are fairly obvious: we only use part of the document collection, our algorithms are very simple and certain techniques such as averaging and scaling are fairly ad-hoc. To fully exploit our intuitions behind this work, we believe that a formal theory will be necessary, partly to better explore the behaviour of term characteristics, and also to eliminate some of the ad-hoc nature of this current work. Nevertheless, the simplicity of our techniques demonstrate that we can achieve significant results without having to consider elaborate indexing or representation techniques.

We have demonstrated that incorporating information on how terms are used within documents, in a feedback situation, can lead to dramatic improvements in retrieval effectiveness, across collections. We have also supported, by the success of the Feedback 1 strategy, our belief that certain combinations of characteristics will be more suitable for certain queries. That is, relevance is better described by different sets of characteristics for different queries.

In future work, we intend to investigate how users can influence this process, by selecting for themselves those characteristics which best describe their information need. In a digital library context, this provides more information upon which to decide those documents to retrieve and also allows the user more flexibility in describing their information need.

Acknowledgements

We would like to thank the IR groups at the University of Glasgow and the University of Dortmund for their support and comments on this work. We would also like to acknowledge the anonymous referees for their helpful remarks.

References

Barry, C.L., and L. Schamber. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information, Processing and Management*. 34. 219-237.

Chang, Y.K., C. Cirillo, and J. Razon. (1971). Evaluation of feedback retrieval using modified freezing, residual collection & test and control groups. In: *The SMART retrieval system: experiments in automatic document processing.*(G. Salton, ed.). Ch. 17, pp 355-370. Prentice-Hall.

Denos, N., C. Berrut and M. Mechkour. (1997). An image system based on the visualization of system relevance via documents. In: *Database and Expert Systems Applications (DEXA '97).* pp 214-224. Springer.

Harman, D. (1992). Ranking algorithms. In: *Information retrieval : data structures & algorithms* . (W. B. Frakes and R. Baeza-Yates, ed.). Ch. 14. pp 363 - 392.

Hearst, M.A.and C. Plaunt, (1993). Subtopic structuring for full-length document access. In: *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, USA. ACM Press.

Paradis, F and C. Berrut. (1996). Experiments with theme extraction in explanatory texts. In: *Second International Conference on Conceptions of Library and Information Science, CoLIS 2.* pp 433-437.

Robertson, S. E. and K. Sparck Jones. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science*. 27. p129-146.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 28. 11-20.

Voorhees, E. M. and D. Harman, (1996). Overview of the Fifth Text REtrieval Conference (TREC-5). In: *Proceedings of the 5th Text Retrieval Conference.* Gaitherburg, MD. pp 1-29. Nist Special Publication 500-238.

Voorhees, E. M. and D. Harman, (1997). Overview of the Sixth Text REtrieval Conference (TREC-5). In: *Proceedings of the 6th Text Retrieval Conference.* Gaitherburg, MD. pp 1-25. Nist Special Publication 500-240.