Phrasier: a System for Interactive Document Retrieval Using Keyphrases

Steve Jones Mark S. Staveley
Department of Computer Science
University of Waikato
Private Bag 3105
Hamilton
New Zealand
email : {stevej, mss2}@cs.waikato.ac.nz

ABSTRACT

Users' information needs are often too complex to be effectively expressed in standard query interfaces to full-text retrieval systems. A typical need is to find documents that are similar to a given source document, yet describing the content of a document in a few terms is a difficult task. We describe Phrasier, an interactive system for browsing, querying and relating documents within a digital library. Phrasier exploits keyphrases that have been automatically extracted from source documents to create links to similar documents and to suggest appropriate query phrases to users. Phrasier's keyphrase-based retrieval engine returns ranked lists of documents that are similar to a given source text. Evaluation indicates that Phrasier's keyphrase-based retrieval performs as well as full-text retrieval when recall and relevance scores assigned by human assessors are considered.

Keywords: keyphrase-based retrieval, query interface, interactive retrieval interface, evaluation

1. INTRODUCTION

The World Wide Web (WWW or Web), digital libraries and similar electronic document repositories have become important tools in the research process. However, as increasing numbers of documents become available through such media, the task of finding useful publications via standard keyword query interfaces becomes more time-consuming [6].

Although keyword query interfaces are widely used in information retrieval systems, they require users to distill potentially complex information needs (such as "I want more papers like this one") into a few terms. Usage analysis of Web based retrieval systems shows that user queries tend to be short and simple [13, 14], yet documents often address a range of topics with varying emphasis and are thus difficult to characterise effectively in a short keyword query. Also, the standard iterative *form query — submit query — assess results — refine query* process required by such interfaces fails to support natural movement from one related document to another.

This paper reports on the design and evaluation of Phrasier, an interactive system that addressess these issues within the context of a digital library. Phrasier uses keyphrases as a semantic feature of documents which can help to identify relationships with related documents. These keyphrases are automatically extracted from document text and are used to provide navigable topic based links, to suggest related phrases which can be used to form queries, to characterise documents, and to reveal topic coverage within a document. At its heart is a novel keyphrase-based retrieval engine that returns ranked lists of related documents. Specifically

1. given a users' source document, Phrasier dynamically and automatically introduces navigable links to related items in a document collection;

2. given a document within a collection, Phrasier introduces links to related documents within the collection;

3. given a source text Phrasier suggests phrases which can be used as query components (either automatically or under user control) to retrieve related documents from the collection.

The paper is organised as follows: first we present a brief overview of the use of phrases in information retrieval interfaces. The features of Phrasier are then outlined, followed by a description of how keyphrases are automatically identified within documents and used to create keyphrase indexes. We then describe a study which compared a full-text retrieval mechanism and the keyphrase-based retrieval of Phrasier. We present and discuss the results and describe avenues for future work.

2. PHRASE-BASED INFORMATION RETRIEVAL INTERFACES

As long ago as 1977, the THOMAS system [16] illustrated how keywords or phrases could be used to guide users in uncovering useful reference documents. In this system users would view sets of keywords and phrases associated with documents and then select subsets of them to focus or broaden their search. More recently, the *Paraphrase* system [3] has utilised phrases in a common World Wide Web search tool. Again users are provided with suggested

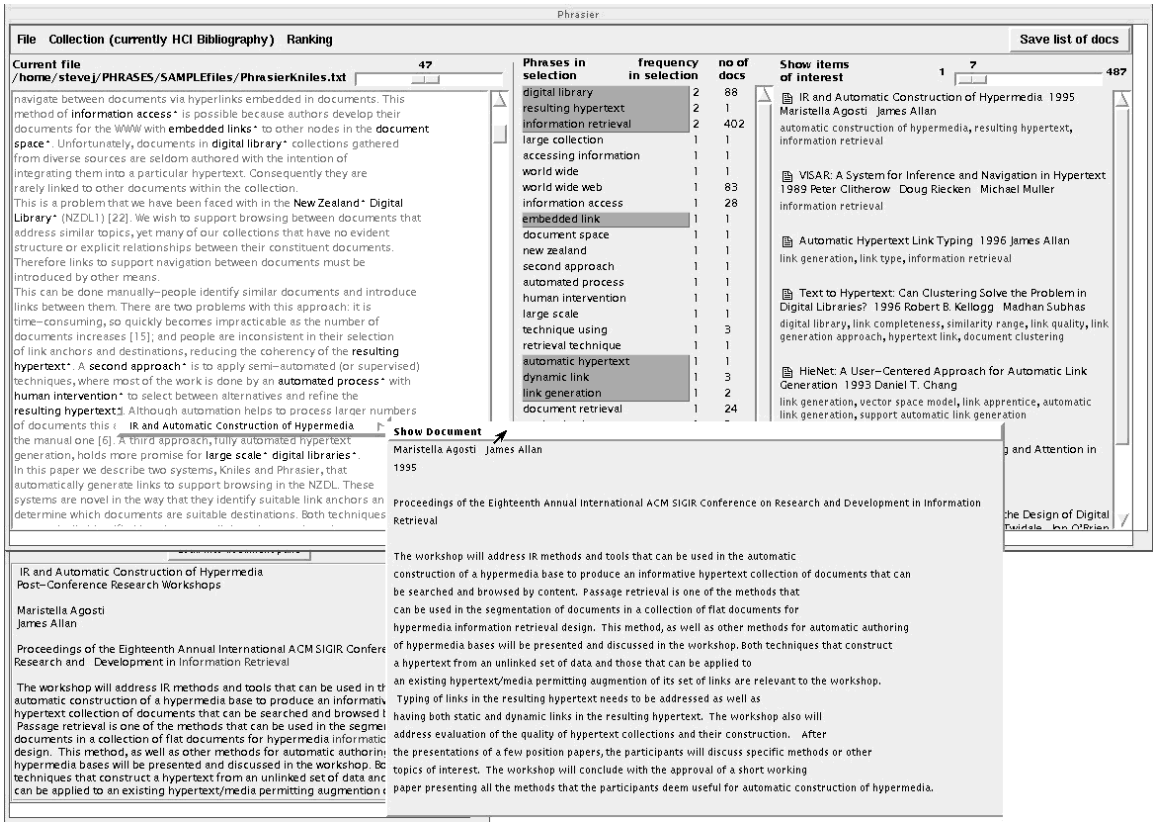

Figure 1: The Phrasier user interface showing the document pane to the left—a keyphrase anchor has been selected and the document preview displayed. The keyphrase pane is in the middle—several keyphrases have been selected and the resulting list of related documents is shown in the document summary pane to the right. To the bottom left is a window showing the text of a selected document.

phrases which they can select and combine to augment or replace their current query. Snippet Search [17] also provides phrases for query extension. It extends the technique used in THOMAS and *Paraphrase* by displaying useful phrases in the context in which they appear within documents.

We follow the approach exemplified by these systems (supporting user selections from phrase lists) because it ameliorates the difficulties faced by users in identifying suitable query terms with which to refine their search. Although users are constrained to a restricted vocabulary of query components (phrases), those components are stronger semantic features of documents than individual terms. In fact, in Phrasier we exploit keyphrases that are strongly associated with documents (see Section 4), and not merely statistically 'important' within a document collection.

Phrasier also supports browsing through the introduction of links into viewed documents as the user moves from document to document. This is somewhat similar to the approach adopted by Golovchinsky in the VOIR system [11]. In VOIR 'important' terms are identified within documents using a blend of heuristic and statistical methods which determine those that discriminate well between documents. These terms become anchors, which when selected are combined with surrounding context (other terms in the sentence) to form a query. The documents resulting from the query become link destinations. Although Phrasier also blends hypertext linking and information retrieval techniques it differs from VOIR in that it exploits automatically extracted keyphrases as inter-document discriminators. These form good candidate link anchors within documents and can be more descriptive than the single term anchors used by VOIR. Unlike VOIR, Phrasier allows users to preview target documents when link anchors are selected, and to make selections across ranges of the document to more flexibly define a context for link-based queries.

3. PHRASIER

Phrasier can be considered as a system for automated hypertext creation [1]. It exploits Information Retrieval techniques (IR) [18] to determine similarity between documents. However, it differs from other systems [2, 4, 7, 19, 20] in that it uses keyphrases rather than the full text of documents to determine similarities. Also, it creates links "on the fly" at the time of browsing rather than off-line prior to browsing. This reduces the overhead in maintaining the hypertext whenever the underlying document collection changes.

3.1 Linking Documents "On The Fly"

The *document pane* (to the left of Figure 1) enables users to enter the text of a document that they are authoring, or to load a previously created document for editing or viewing. Existing documents may also be sourced from the document collection with which the user is interacting. As a user enters text, or it is read from a file, keyphrases are identified and highlighted. These are

phrases which have been allocated as keyphrases to documents within the target collection. Each phrase is issued as a query to a keyphrase-based retrieval engine which ranks the resulting documents, returning the list to Phrasier. A link anchor is then inserted into the text for the keyphrase, pointing to the set of related documents.

3.2 Retrieving Similar Documents

Link anchors within the source text support navigation to documents retrieved with respect to a single keyphrase. Phrasier also supports retrieval of documents that are similar to the source text as a whole, or which relate to sets of topics addressed within the source text.

As keyphrases are identified in the source text they are inserted into the *keyphrases pane* (shown in the middle of Figure 1). Two items of information are displayed for each keyphrase: the frequency with which each keyphrase occurs in the text, and the number of the documents within the collection to which each keyphrase has been allocated.

Keyphrases in the list can be selected individually, in contiguous blocks, and in multiple disjoint blocks. A selected list of keyphrases can be issued as a query to the retrieval engine, which then returns a ranked list of documents which are displayed in the document summary pane (to the right of Figure 1). Issuing the complete list as a query can be thought of as retrieving documents that are similar to the source document as a whole. The user can set this behaviour as the default. Issuing single keyphrases as a query returns the same documents that are associated with the corresponding phrase anchor in the document pane. Issuing a group of keyphrases allows the user to focus on a subset of the topics addressed in the document. The keyphrase pane essentially provides users with a controlled vocabulary of query phrases which the source document has in common with the target collection.

3.3 Document Display

The *document summary pane* displays summary detail of documents. When bibliographic information is available, some items (such as title, author, year, and keywords) are displayed in the summary pane. Summary information can be saved to disk for later consideration. The text of a document can be retrieved and viewed by selecting its icon in this pane. Text of this document is displayed in a separate window so that the source document text can be viewed at the same time, and the keyphrases that formed the query are highlighted within the document text. The text of the document can then be transferred into the document pane to continue the browsing process.

3.4 Collections

Phrasier currently works with two collections of the New Zealand Digital Library (NZDL, http://www.nzdl.org) [21]. The first is a mirror of the Human Computer Interaction (HCI) Bibliography (http://www.hcibib.org) which contains the bibliographic details of more than 15,000 documents. The second is the Computer Science Technical Report (CSTR) collection which contains more than 40,000 documents collected from electronic repositories around the world. The Phrasier interface is a client which connects to a collection server across a network—there is nothing which binds the interface to particular collections.

4. KEYPHRASE EXTRACTION AND INDEXING

Phrases which characterise a document are often inserted into the document by the author (as is the case with this paper). Often, when there are no author-specified keywords, experts will assign keywords and phrases to documents to support classification and retrieval. Many of the documents in the NZDL collections do not contain keywords or phrases—less than a third of the entries in the HCI Bibliography, for example. The original format of CSTR documents is postscript from which it is difficult to extract such metadata. Manual keyword/phrase assignment is prohibitively resource intensive.

The NZDL project has developed Kea [10,] a system that applies machine learning techniques to automated keyphrase extraction. Kea is trained using a small subset of a collection's documents for which phrases have already been specified (or manually assigned). A model is built for determining whether any given candidate phrase from the collection is a keyphrase or not. In the extraction phrase new documents without known keyphrases are passed through the model. Candidate phrases from within the document are accepted or rejected by the model resulting in a list of accepted keyphrases for each document. The maximum number of keyphrases to be accepted for each document is a parameter to the process. Typically this number will be between 10 and 15. Evaluation of Kea indicates that its performance is on a par with the current state-of-the-art [10].

We use the output from Kea to create keyphrase indexes [12] for a collection. A *keyphrase list* associates an identifier with each keyphrase. A *keyphrase to document* index lists every extracted keyphrase and indicates for each the documents from which it was extracted. A *document to keyphrase index* lists every document in the collection and the keyphrases that were extracted from each. Other tables store precomputed frequency statistics for use in document similarity measurements.

5. SIMILARITY MEASURES USING KEYPHRASES

To establish relationships between keyphrase queries and documents we are interested in the degree of overlap between sets of keyphrases for given documents—the closer the overlap, the more similar the documents are. We have used standard similarity measures to compare sets of document keyphrases. Specifically we have adopted two vector space model cosine measures that have been modified to use keyphrase frequencies rather than the term frequencies of full-text retrieval systems. Here, query and document vectors contain keyphrases rather than terms and so we have replaced references to terms (t) with reference to phrases (p) accordingly. The first measure is taken from Witten et al [24]:

$$\cos ine(q, D_d) = \frac{\sum_{p \in q \cap d} W_{q,p} W_{d,p}}{W_q W_d} = \frac{\sum_{p \in q \cap d} (f_{q,p}.\log \frac{N}{f_p})(f_{d,p}.\log \frac{N}{f_p})}{(\sum_{p \in q} (f_{q,p}.\log \frac{N}{f_p})^2)^{\frac{1}{2}} (\sum_{p \in d} (f_{d,p}.\log \frac{N}{f_p})^2)^{\frac{1}{2}}} \quad (1)$$

It was used because it is utilised in MG [24], the retrieval engine used by the NZDL, against which we wished to compare keyphrase-based retrieval. A second method (from [9]) was used to test whether any performance differences were independent of a particular retrieval algorithm. Also the second algorithm provides the benefit that $W_{d,p}$ is independent of the size of the document collection, requiring less index maintenance to be carried out when the collection is updated:

$$\cos ine(q, D_d) = \frac{\sum_{p \in q \cap d} (W_{q,p}.W_{d,p})}{\sqrt{\sum_{p \in q} W_{q,p}^2 . \sum_{p \in d} W_{d,p}^2}} = \frac{\sum_{p \in q \cap d} \log(f_{q,p}).\log(\frac{N}{f_p}+1).\log(f_{d,p}+1)}{\sqrt{\sum_{p \in q} (\log(f_{q,p}).\log(\frac{N}{f_p}+1))^2 . \sum_{p \in d} \log(f_{d,p}+1)^2}} \quad (2)$$

A question that arises in the transition from term-based to phrase-based measures concerns the f_p value. In term based measures f_t represents the number of documents that contain term t, and so f_p might represent the number of documents that contain keyphrase p. However, keyphrases are associated with only a subset of the documents in which they occur, and so we might consider f_p to be the number of documents which have p as a keyphrase. Therefore we implemented each variation of the two measures:

Method	Description
P1	equation 1 where f_p represents the number of documents that contain p
P2	equation 1 where f_p represents the number of documents to which p is allocated
P3	equation 2 where f_p represents the number of documents that contain p
P4	equation 2 where f_p represents the number of documents to which p is allocated

6. EXPERIMENT

We carried out a study to determine the utility of the documents returned by Phrasier's keyphrase-based similarity measures. We were interested in several aspects of such an approach to document retrieval. First we wished to determine the level of similarity (or from a user's perspective, relatedness) perceived by users between a document under consideration and documents retrieved using the similarity measures. Additionally we wished to investigate if there was any difference between the similarity measures in terms of user perceptions of the related documents returned by each. We also wished to investigate how keyphrase-based measures perform in comparison to standard full-text term-based ranked retrieval mechanism.

Subject	Paper	References in paper	Occurring within the collection	%
S1	A	9	2	22
	B	17	7	41
S2	A	15	9	60
	B	20	12	60
S3	A	23	3	13
	B	19	1	5
S4	A	20	0	0
	B	13	2	15
S5	A	25	13	53
	B	5	1	20
S6	A	21	1	5
	B	0	0	0
		Total	51	

Table 1: Overlap of source document reference lists with target collection

Mean number of returned documents common to lists				
	P1	P2	P3	P4
T1	1.08	1.08	1.08	1.08
P1		7.25	7.42	7.58
P2			8.50	9.00
P3				9.25

Table 2: mean number of returned documents common to top ten returned documents

We studied a task that users of a digital library might realistically undertake when carrying out a literature search—finding documents related to a given source document. The keyphrase extraction process is collection (domain) focussed rather than general and so we required query documents that were related to a particular domain (in this case HCI). For this reason we chose not to use a controlled but variable-domain set of query and target documents such as those available from TREC. Consequently we did not have a well-defined set of queries, documents and relevance judgements against which to assess performance.

The question therefore arose as to how we might measure recall of relevant documents, and hence how relevance might be established. Importantly we were interested in variations in relevance score rather than simple relevant/irrelevant classifications. Therefore, as with TREC, human assessors considered the relevance of documents to queries. One difference was that the assessors were the authors of the query documents and by necessity formed a focussed group who had published in areas covered by the supported document collections. Additionally we could not determine exactly how many documents in the entire collection were relevant.

To measure recall we considered documents to be relevant by two criteria. First, documents that occurred both within the reference list of the query document and the collection were considered relevant. Second, returned documents that scored more than 4 on a 1 to 7 relevance scale during human assessment were deemed to be relevant.

6.1 Subjects

Nine subjects took part in the study as a whole. Three subjects were graduate students and six were faculty members in a

Subject	Paper	Number of references contained in top ten list				
		T1	P1	P2	P3	P4
S1	A	0	0	0	0	0
	B	0	0	0	0	1
S2	A	3	3	3	3	3
	B	3	4	4	4	4
S3	A	0	0	0	0	0
	B	1	0	0	0	0
S4	A	X	X	X	X	X
	B	1	1	1	1	2
S5	A	3	4	4	4	4
	B	1	0	0	0	0
S6	A	0	1	1	1	0
	B	X	X	X	X	X
Total		12	13	13	13	14
Percent of maximum		23.5	25.5	25.5	25.5	27.5

Table 3: recall at ten documents. X indicates that the document contained no references from the collection

Subject	Paper	Number returned in full list		Number of references contained in full list returned by	
		P1-4	T1	T1	P1-4
S1	A	261	500	1	0
	B	1058	500	1	2
S2	A	244	500	5	5
	B	353	500	8	5
S3	A	1084	500	2	2
	B	1107	500	1	1
S4	A	281	500	X	X
	B	760	500	2	2
S5	A	674	500	9	8
	B	161	500	1	1
S6	A	611	500	0	1
	B	493	500	X	X
Total				30	27
Percent of maximum				58.8	52.9

Table 4: recall for full returned lists

Computer Science department of a university. Two of the subjects took part in a pilot study and the seven remaining subjects (referred to as S1 to S7) took part in the study proper. One subject (S7) did not fully respond and their data was discarded.

6.2 Methodology

In the study proper, a total of twelve recently published research papers were collected—2 papers (referred to as A and B) authored by each subject. The subjects' research topics were related to Human-Computer Interaction (HCI) which is covered by the HCI Bibliography—the document collection used in the study. For each paper the process was as follows: formatting markup which did not contribute to the content of the paper was removed; the reference list was removed; the paper was read into Phrasier; the full set of keyphrases was identified within the document, and a ranked set of similar documents was retrieved using each of the four similarity measures in turn. Also, after being cleaned-up, the entirety of each paper was submitted as a ranked, full-text query to the New Zealand Digital Library HCI Bibliography collection. The ranked result list (labelled T1) was stored.

Subject	Paper	Mean relevance assessment of each top ten list				
		T1	P1	P2	P3	P4
S1	A	4.7	5.3	5.3	5.9	5.8
	B	3.2	2.3	2.3	2.3	2.7
S2	A	4.4	4.8	5.4	5.1	5.2
	B	5.7	5.8	5.8	5.8	5.8
S3	A	3.7	3.5	4.1	4.3	4.0
	B	3.6	3.1	4.0	3.6	3.6
S4	A	1.5	1.7	1.8	1.8	1.8
	B	2.0	5.6	5.7	5.7	5.7
S5	A	3.7	6.1	5.6	6.0	6.0
	B	2.2	6.2	6.4	6.4	6.4
S6	A	2.7	2.6	3.0	3.1	3.0
	B	2.5	2.5	2.3	2.4	2.3
mean across subjects		3.33	4.13	4.31	4.37	4.36

Table 5: mean relevance from combined individual document scores.

Therefore five lists of related documents were generated for each source document. The top ten documents from each list were then combined and duplicates removed to create a sixth composite list. The top ten from each list were combined because of evidence that users infrequently investigate beyond the first few documents returned in a result set [13, 14]. The mean length of these lists was just over 22 documents, the longest containing 25 documents, the shortest 17. An item in each list contained the title, author list, and source (the name of the conference proceedings, journal etc) of the returned document.

Each subject was required to carry out two tasks for each paper that they submitted. In the first task subjects were presented with the composite list and asked to judge how related each item in the list was to their original source document. In the second task subjects were presented with each of the six lists in turn and asked to judge the relevance of each list as a whole. In both cases judgements were on a scale from 1 (not at all relevant) to 7 (highly relevant). Subjects could provide comments on the documents or lists if they desired.

6.3 Relationship Of Papers To The Collection

One indicator of how strongly related the papers were to the target collection was the overlap between the reference list of each paper and the documents within the collection. This is shown in Table 1. For three papers more than half of the references were drawn from the HCI Bibliography. For a further three papers between a fifth and a half of the references came from the collection, and for the remaining six papers there were fewer than a fifth from the collection. In total 51 references were found to be in the collection.

6.4 Results

A coarse indicator of the commonality between the similarity measures is the mean number of returned documents common to pairs of top ten lists. Table 2 shows that there was little overlap (a mean of just over one in ten documents) between the lists returned by retrieval on the full text of the document (T1) and those returned by phrase-based retrieval. The lowest mean overlap between phrase-based retrieval measures (P1 and P2) is just more than seven documents, with a highest mean of just over nine documents being common to P3 and P4.

Score	Mean number of document s allocated score in each list				
	T1	P1	P2	P3	P4
1	2.50	2.00	1.58	1.50	1.67
2	1.67	1.33	1.42	1.42	1.25
3	1.25	0.83	0.75	0.75	0.75
4	1.33	0.58	0.83	0.75	0.75
5	1.50	1.50	1.25	1.33	1.25
6	1.42	2.00	2.33	2.33	2.42
7	0.33	1.75	1.83	1.92	1.92

Table 7: mean number of documents allocated each relevance score.

Subject	Paper	Relevance assessments of top ten lists				
		T1	P1	P2	P4	P3
S1	A	6	4	5	6	3
	B	4	3	2	2	2
S2	A	5	6	7	6	7
	B	6	6	5	5	7
S3	A	5	5	6	6	6
	B	3	2	3	4	4
S4	A	1	1	3	5	1
	B	1	6	7	7	6
S5	A	4	7	5	5	6
	B	3	6	6	5	5
S6	A	3	5	3	3	4
	B	4	5	3	1	2
	mean across subjects	3.75	4.67	4.58	4.58	4.42

Table 8: relevance assessments for top ten lists viewed as a whole.

The recall at ten documents is shown for each measure in Table 3. Relevant documents are defined as those that occur within the reference list of a paper and also in the document collection. The recall across all documents for all measures is approximately a quarter of documents common to the reference lists and the collection. T1 recalled 12 documents, P1-P3 recalled 13 documents and 14 were recalled by P4.

When we consider the complete lists of documents returned by each measure we see recall improvements for both types of measure. For T1, the maximum number of documents that can be returned through the query interface is 500. The maximum number of returned documents is the same for each of P1-P4 and is dependent upon the keyphrases that constitute the query. Table 4 shows that when full returned lists are considered T1 returns 30 of the 51 references, giving a recall of 58.8% across all papers. In comparison P1-4 gives recall of 52.9% (27 references).

The relevance scores that were assigned to documents in the composite list were collated back into the individual top ten lists for each source document. The mean subject-allocated relevance score was then calculated for each list and is shown in Table 5. The lowest mean relevance across all subjects (3.33) was achieved by the full-text retrieval measure (T1) and indicates that the results returned by T1 were judged to be slightly but not completely unrelated to the source document. Each of the phrase-based measures achieved mean scores greater than four across all subjects. Using Friedman two-way analysis of variance by ranks we find no evidence that there is a difference between the scores allocated by subjects to the measures (p=0.005). We consider a score of 5 or more to indicate that a subject assessed a document to be relevant. The mean number of documents assessed to be relevant (across all subjects) is shown in Table 6. On average, for each of P1-P4 more than half of the ten documents were judged to have some relevance, and almost 2 documents of the ten were judged to be highly relevant.

Document position	Mean relevance assessment of documents in each of the locations in the top ten lists				
	T1	P1	P2	P4	P3
1	4.33	5.25	4.92	5.17	5.17
2	4.00	4.58	4.58	4.67	4.50
3	4.25	4.92	4.75	4.83	4.75
4	3.75	3.92	5.17	4.83	4.83
5	3.25	4.08	4.75	4.75	4.83
6	2.75	4.58	4.08	4.83	4.00
7	3.33	4.50	4.50	4.33	4.25
8	2.50	3.42	3.67	3.42	4.08
9	2.50	3.08	3.17	3.92	3.25
10	2.58	2.92	3.50	2.83	4.00

Table 9: mean relevance assessments for each location in the top ten lists.

The mean number of documents in each top ten list (across subjects) to which each score (range 1 to 7) was allocated is shown in Table 7. For each of P1-P4 subjects had a tendency to assign either high or low rather than median scores to documents. Fewer high relevance scores were assigned to T1, although Friedman two-way analysis of variance by ranks indicates that there is no significant evidence of differences between the distributions of mean scores (p=0.005).

Table 8 shows the relevancy assessment by each subject for each top ten list. This was allocated by subjects when viewing the list as a whole. The differences between the methods are not significant (Friedman two-way analysis of variance by ranks, p=0.005) and there is no evident pattern of subjects assigning higher scores to the whole lists than result from collation of individual document scores.

The mean scores assigned by subjects to documents in each position in the top ten lists are shown in Table 9. Friedman two-way analysis of variance by ranks provides evidence that there is a difference between the scores allocated by subjects to the measures (p=0.005, tie adjusted). Multiple comparisons (α=0.1) indicate a significant difference between the scores for T1 paired with each other measure. No other pair-wise differences are significant.

6.5 Discussion

The sets of top ten documents returned by full-text and keyphrase retrieval mechanisms are quite distinct. On average, only one document was common to T1 and P1-P4, whereas there is a noticeable degree of overlap between each of P1-P4 (at least 7 common documents). However, recall of relevant documents is almost identical at 10 documents for T1 and P1-P4. The two approaches achieve the same recall but produce different sets of relevant documents. Recall scores at a range of thresholds are very similar for both approaches. T1 begins to perform better at result set sizes over 300, although the difference is slight.

There is little difference between the number of query document references returned in full result lists (500 documents for T1,

variable size for P1-P4). Again the sets of relevant documents differ. Where relevant documents are common to the returned lists for T1 and P1-P4 they tend to have been allocated different rankings by the two approaches.

Using a result set size of 500 (the maximum for T1), neither approach returned the full set of 51 relevant reference documents from all papers (30 for T1 and 27 for P1-P4). There is little difference between the approaches when we consider the point in the result list at which they achieved their best recall level. The mean value was just under 40% of the result list for P1-P4 and just over 40% for T1. Although P1-P4 tended to return more documents than T1 (mean of 574), all but one relevant document was found within the first 500. For documents that were references within source documents we believe that the two approaches give highly similar recall performance overall.

When combined document scores, or scores for top ten lists as a whole are considered it subjects perceive P1-P4 to return documents that are more relevant than those returned by T1. However, there is no significant difference between the mean number of documents assigned each relevance score (1 to 7) for each measure.

There is only one measure by which the methods vary significantly—the mean relevance scores assigned to documents in each of the first ten locations in each list. This is an important measure because it reflects precision and also how well the documents are ranked within the list. For P1-P4, across all subjects, the first 6 documents were judged relevant (score greater than 4) with only one exception (position 4 for P1). For T1, only the first 3 documents were judged relevant. Although recall (based on reference list documents) is almost identical for the each approach, it seems that subjects perceived the phrase-based approach to provide slightly higher precision within the first ten result documents. Given that for all methods perceived relevance tends to decrease from beginning to end of the top ten list it appears that they are effectively ranked in relevance order.

Overall we see little difference in the performance of the two methods, either in recall of identifiably relevant documents or perceived relevance of returned documents. Within the phrase-based method we see no evidence of differences between the two algorithms or their variations.

Previous studies vary in their conclusions regarding the comparative effectiveness of phrase-based and full-text retrieval. For example, Croft et al [8] report that phrases extracted from natural language queries improve retrieval when used in structured queries. Although Mitra et al [15] observe that phrases are effective in particular contexts (such as determining the relative ranks of low-ranked documents) they also note that as overall retrieval effectives increases, the additional benefit of using phrases is decreasing. Strzalkowski and Carballo [22] report substantial performance improvements from the use of phrasal terms, but indicate that the approach is suited to longer queries in particular. Smeaton [21] suggests that the use of phrases in query formulation (as in Phrasier) benefits experienced users, but not naïve users.

It is difficult to generalise about the effectiveness of phrases from these and other studies and to therefore offer a direct comparison with our results. The studies mentioned above report improved performance with phrases in particular contexts. We observed little difference between the phrase-based and full-text approach using a range of measure, except in the case of user perceptions of precision. It is worth emphasising that the keyphrases used in Phrasier are identified using machine learning techniques. This novel approach differs from the standard syntactic and statistical techniques which are widely established and used. We expect retrieval performance to improve as the keyphrase extraction software is refined.

7. CONCLUSIONS

We have described Phrasier, a system which supports browsing and querying of a document collection using automatically extracted keyphrases rather than combinations of single terms to form queries. One novel component of Phrasier is a keyphrase-based retrieval system that carries out ranked retrieval on given sets of keyphrases. We have carried out a study to determine, from a user perspective, the effectiveness of keyphrase-based retrieval compared to full-text retrieval. The study shows that keyphrase-based retrieval can be as effective as full-text retrieval, both in terms of recall of identifiably relevant documents and relevance scores assigned to returned result sets by human assessors.

This result is encouraging because indexes consisting solely of keyphrases can be substantially smaller than conventional full-text indexes [12]. We believe that the results are not an anomaly of a single ranking algorithm modified to use keyphrases, given that P1-P2 and P3-P4 perform with no significant differences. This is also encouraging because P3 and P4 allow us further flexibility in collection maintenance as document weights are not dependent on collection size. The fact that there is no noticeable difference between the variations on the two phrase-based algorithms is also useful. The use of f_p as the number of documents containing p would require more storage for indexes than the alternative, but now appears unnecessary.

Although our study was small scale, we focussed on a particular task associated with digital library usage, and consider the results to be supportive of further investigation. We will follow three main avenues of future work. The first will involve evaluation of the Phrasier user interface by standard usability evaluation techniques. Second, we will evaluate the suitability of the links generated by Phrasier, deploying a variety of methods [5]. Third, we will undertake standard retrieval performance evaluation of keyphrase-based retrieval using standard corpora.

8. ACKNOWLEDGEMENTS

The authors are grateful to Gordon Paynter, Gene Golovchinsky and Richard Littin for a range of helpful comments on this work. Thank you to Carl Gutwin who collaborated on the initial design and implementation of Phrasier.

9. REFERENCES

1 Agosti, M., Crestani, F., and Melucci, M. On the use of information retrieval techniques for the automatic construction of hypertext. Information Processing and Management, 33, 2 (1997), 133-144.

2 Allan, J. Building hypertext using information retrieval. Information Processing and Management, 33, 2 (1997), 145-159.

3 Anick, P.G. and Vaithyanathan, S. Exploiting clustering and phrases for context-based information retrieval. In: Proceedings of SIGIR 97, (Philadelphia, USA, 1997), ACM Press, 314-322.

4 Bernstein, M. An apprentice that discovers hypertext links. In Proceedings of the European Conference on Hypertext (ECHT'90), (1990), Cambridge University Press, 212-223.

5 Blustein, J., Webber, R.E., and Tague-Sutcliffe, J. Methods for evaluating the quality of hypertext links. Information Processing and Management, 33, 2, (1997), 255-271.

6 Bollacker, K.D, Lawrence, S. and Giles, C.L. CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In: Proceedings of the Second International Conference on Autonomous Agents, (Minneapolis, St.Paul, May 9-13, 1998). ACM New York, 116-123.

7 Chua, T-S. and Choo, C-H. Automatic generation and refinement of hypertext links. The New Review of Hypermedia and Multimedia, 1, (1995), 41-66.

8 Croft, B.C., Turtle, H.R. and Lewis, D.D. The use of phrases and structured queries in information retrieval. In: Proceedings of SIGIR 91, ACM Press, 32-45

9 de Kretser, O., Moffat, A., Shimmin, T., and Zobel, J. Methodologies for distributed information retrieval, In Proceedings of 18th International Conference on Distributed Computing Systems, (Amsterdam, May 1998), 66-73.

10 Frank, E., Paynter, G.W, Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. Domain-specific keyphrase extraction. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers, San Francisco, CA. (1999), In Press.

11 Golovchinsky, G. What the query told the link: the integration of hypertext and information retrieval. In Proceedings of Hypertext '97, 1997, ACM Press, 67-74.

12 Gutwin, C., Paynter, G., Witten, I.H., Nevill-Manning, C.G., and Frank, E. Improving browsing in digital libraries with keyphrase indexes. Technical Report, Department of Computer Science, University of Saskatchewan, Canada (1999)..

13 Jansen, B.J., Spink, A., Bateman, J., and Saracevic, T. Real life information retrieval: a study of user queries on the web. SIGIR Forum, 32, 1, (1998), 5-17

14 Jones S., Cunningham S.J. and McNab R.J. An analysis of usage of a digital library. In Proceedings of ECDL'98 Second European Conference on Digital Libraries, 1998. Springer, 261-277.

15 Mitra, M., Buckley, C., Singhal, A. and Cardie, C. An analysis of statistical and syntactic phrases. In: Proceedings of RIAO 97 (Montreal, Canada, June, 1997), 200-214.

16 Oddy, R.N. Information retrieval through man-machine dialogue. Journal of Documentation, 33, 1 (1977), 1-13.

17 Pedersen, J., Cutting, D. and Tukey, J. Snippet Search: a single phrase approach to text access. Xerox PARC Technical Report SSL-91-08.

18 Salton, G. Automatic text processing—the transformation, analysis and retrieval of information by computer. Addison-Wesley Publishing Co, Reading, MA. 1989.

19 Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41, 4, (1990), 288-297,.

20 Salton, G., Singhal, A., Mitra, M., and Buckley, C. Automatic text structuring and summarization. Information Processing and Management, 33, 2, (1997), 193-207.

21 Smeaton, A.F. and Kelledy, F. User-Chosen phrases in interactive query formulation for information retrieval. In: Proceedings of the 21st BCS IRSG Colloquium, (Autrans, France, March 1998).

22 Strzalkowski, T. and Carballo, J.P. Natural language information retrieval. In: The Fourth Text Retrieval Conference (TREC-4) (Gaithersburg, Maryland, November 1-3, 1995), 245.

23 Witten, I.H., McNab, R., Jones, S., Cunningham, S.J., Bainbridge, D., and Apperley, M. Managing multiple collection, multiple languages, and multiple media in a distributed digital library. IEEE Computer, 32, 2, (1999), 74-79.

24 Witten, I.H., Moffat, A. and Bell, T.C. Managing gigabytes: compressing and indexing documents and images. Van Nostrand Reinhold, 1994.