TREC-3: Experience With Conceptual Relations in Information Retrieval

David Gardiner, John Riedl, and James Slagle
University of Minnesota
Computer Science Department
(gardiner | riedl | slagle)@cs.umn.edu

Abstract

This report describes an experiment evaluating the performance gains that can be achieved by using high-level conceptual relations in information retrieval. The objective of the experiment is to determine if conceptual relations can improve overall retrieval performance and, if so, under what conditions using relations is likely to be justified. We represent five TREC topics each as two concepts linked by a single relation, where a concept corresponds to a noun phrase and a relation corresponds to a verb phrase or a noun phrase that describes an action, activity, or relationship. A Boolean search (with proximity) is associated with each concept and a parameterized search with each relation. We then compare the performance of the expanded concept-relation-concept representation with the searches for the two concepts linked by each of several proximity operators. Our results show that use of relations can provide significant performance improvements but that the improvements are dependent on the nature of the two concepts and the relation with respect to the text collection being searched.

1. Introduction

The primary objective of the field of information retrieval is to provide technology allowing people to find sufficient information to answer questions. Historically, the problem of information retrieval was finding *any* relevant information. Now, with the wide-spread availability of enormous electronic databases, being overwhelmed with *too much* information is common. The challenge of information retrieval is therefore to provide relevant and *only* relevant information. Unfortunately, providing more relevant information tends to also result in more non-relevant information, just as eliminating non-relevant information tends to eliminate information that is relevant.

A key to improving information retrieval performance is for the retrieval system to have a better representation of the user's information need. Information need is often expressed as a set of concepts. The focus of this research project is to determine whether specifying and searching for specific relations between concepts — one way to more completely specify the information need — can improve retrieval performance. For the purpose of this project, a concept corresponds to a noun phrase and a relation corresponds to a verb phrase or a noun phrase that describes an action, activity, or relationship.

Ideas that form information needs are comprised of both concepts and relations. If there are multiple relations that can occur between two concepts, specification of the relation becomes critical to properly specifying the query. Cimino & Barnett note the importance of relations in their analysis of questions generated by physicians (Cimino & Barnett 1992). They cite an example of the concepts *methotrexate* and *psoriasis* and several questions that might be posed by these two concepts: *"Does methotrexate cause psoriasis?", "Does methotrexate treat psoriasis?",* and *"Is the use of methotrexate contraindicated in*

someone with psoriasis?" They advocate development of a vocabulary of "Relation Concepts" to allow more precise specification of medical questions.

While use of relations can allow better representation of information need than concepts alone, there has been limited research evaluating the impact that using relations can have on retrieval performance. The two research questions addressed in this project are:

1. Can retrieval performance be improved by explicitly specifying and searching for relationships between concepts?

2. Under what conditions is the improvement likely to be justified?

Systems such as RUBRIC (McCune et al., 1985) search for concepts. Relationships between concepts are specified implicitly, using Boolean operators and proximity. The idea of using explicit relations to improve retrieval performance has been explored in the context of document indexing. (Farradane, 1981) describes a notation for indexing relations between concepts. Concepts and relations can be specified in a search. Searches then use the index to identify matches. (Farradane & Thompson 1980) describes preliminary experiments indicating that relational indexing can improve retrieval performance. Relational indexing requires manual recognition and indexing of relations in each document in the collection. While manual indexing of concepts has been widely used, it is labor-intensive. Our approach builds on (Farradane, 1981) in that we also explore using relations to improve retrieval performance. However, instead of indexing each relation, we construct a search expression for the existence of the relation in the collection. The manual effort required is in developing the recognition expression. The only indexing required is automated word-based indexing as used in most current information retrieval systems.

2. Experimental Design

To answer these research questions, our experiment compares the performance of searches using a relation between two concepts to searches using a simple proximity operator between the same two concepts. We use concept-relation-concept queries because doing so reduces the issues associated with representing a topic. In many cases, an information need cannot be represented as a single concept-relation-concept (Liddy & Myaeng 1994). A complete representation of an information need often involves multiple concepts and relations, nesting, and functional dependencies. Evaluating the effect of a single relation in a complex representation would be difficult. Furthermore, it is not clear what a complex representation would be compared against. We chose proximity operators as our baseline for comparison because they are commonly used to specify that some relationship exists between two terms without specifying the relationship explicitly.

Since this experiment is done as a part of the Third Text REtrieval Conference (TREC-3), we use the TREC document collection, topics, and relevance judgments. Most TREC topics require more than two concepts and a relation for a complete representation (Liddy & Myaeng 1994). Of the five topics that we use, only two were adequately represented by concept-relation-concept. This means that, for three of the topics used, there is a significant discrepancy between the topic on which the relevance judgments are based and the representation of the topic that we use for searching. We use failure analysis to analyze the effect of this discrepancy on our results.

Figure 1 shows the experimental process used to collect performance results. An information need specification is manually translated into both a concept-relation-concept representation and a concept-

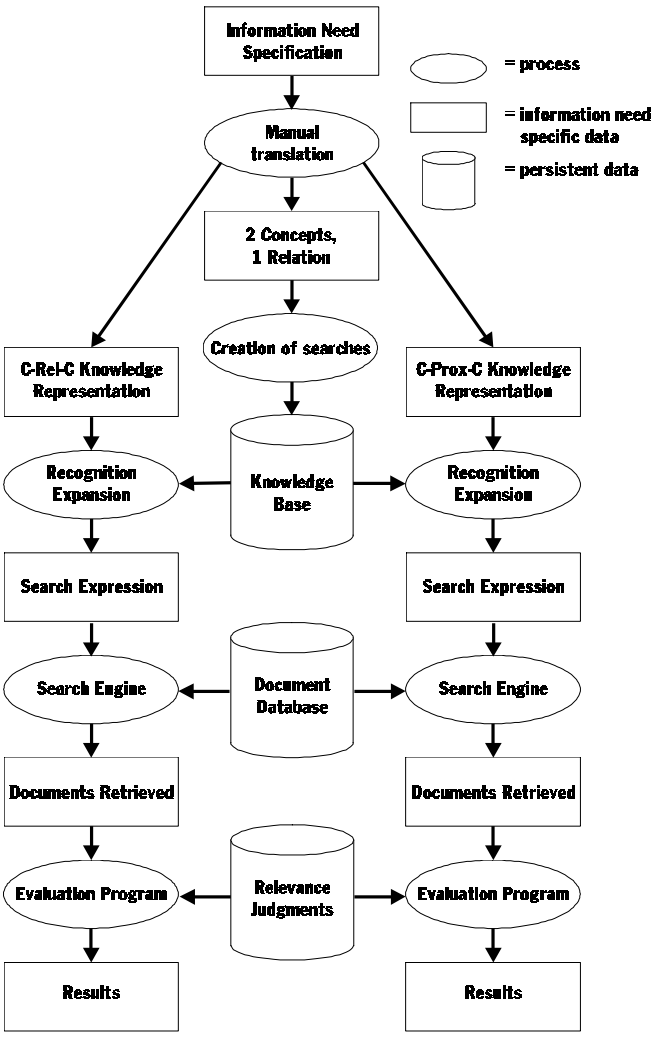

Figure 1. Experimental Process for a Single Information Need. For the experiment, information needs are manually translated into Concept-Relation-Concept representation. The same two concepts are also linked with a proximity operator to produce a Concept-Proximity-Concept representation. The two representations go through the succeeding steps in parallel. First, each is expanded into a search expression using the knowledge base. Next, documents are retrieved by a search engine using the search expression. Then the retrieved documents are evaluated using the relevance judgments to produce performance results.

proximity-concept representation. A search expression is created for each of the concepts and the relation and put into the knowledge base. At this point there are two parallel, identical paths: one for the concept-relation-concept representation and one for the concept-proximity-concept representation. Each representation is expanded using the searches stored in the knowledge base and the expanded search run against a large document collection. The documents retrieved are compared to the pre-existing relevance judgments and an evaluation program outputs performance results. The results for the concept-relation-concept and concept-proximity-concept representations are then compared.

Our planned experimental methodology is shown in Figure 2. We plan to collect results for approximately five information needs. The searches will then be developed and refined using a training collection. When the searches are as good as we can make them, we will collect "after refinement" results. We then plan to load two other collections in turn, collecting "blind" results, refining the searches, and then collecting "after refinement" results on each collection. Finally, we plan to load a final test collection and collect "blind" results on that collection.

Our methodology involves manual effort used to create the concept-relation-concept representations and the concept and relation searches. Since this experiment is intended as a qualitative exploration of the value of relations in information retrieval, the source of the relations and their linked concepts is not significant. Their significance is in what we learn from evaluating their performance. Similarly, the source of the concept and relation search expressions is unimportant. The quality of the concept search expressions has little effect on the results of the experiment as the same concept searches are used in both the concept-relation-concept and concept-proximity-concept searches. For relation searches, our objective is to find whether good relation searches can be defined. If the relation

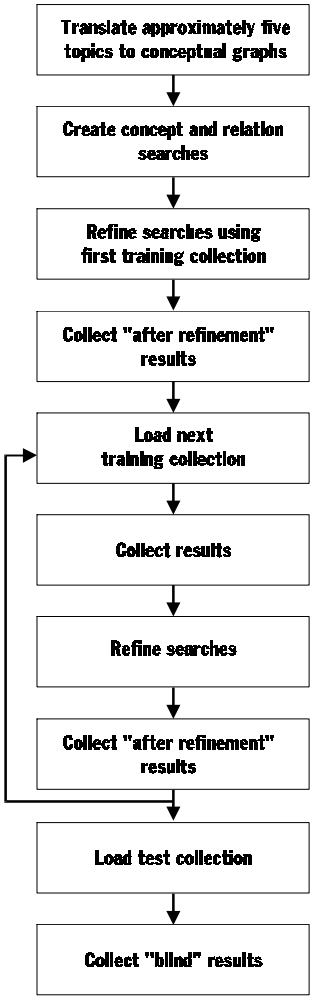

Figure 2. Planned Methodology. TREC topics are translated to concept-relation-concept graphs. Searches for the concepts and relation are created and refined using training collections. A test collection is loaded and final "blind" results collected.

searches are poorly defined, then this experiment might make relations look less effective than they really are.

A danger in using manually created search expressions is that they might be *tuned* to a particular collection so that changes that improve performance in one collection make performance worse in another. We contrast tuning with *refinement*, which improves performance in a new collection without negatively impacting the performance in a previous collection. The use of multiple training collections allows us to determine how much tuning is occurring. The collections used are drawn from two different sources, with all documents in each collection drawn from a single source. This diversity between collections helps ensure that we do not tune for a single source. If tuning occurs, this implies that a relation search may work only with collections very similar to the training collection. If tuning does not occur, it suggests that relations may be defined that can work on a broader set of collections.

The following sections describe the components of the dataflow diagram in detail.

2.1. Information Need Specification and Knowledge Representation

The start of any information retrieval process is some specification of an information need. For this experiment, we use topics and relevance judgments provided as part of the Third Text REtrieval Conference (TREC-3). The narratives of the topics used in this experiment are shown in Figure 3.

We need a notation for the concept-relation-concept representation of an information need. Since our representational need is simple, virtually any knowledge representation has the representational power required. The notation we use is Sowa's conceptual graphs (Sowa 1984). Examples of other representations that would have worked are frames or any Lisp-like notation.

Conceptual graphs are directed graphs containing concepts (denoted with brackets) and relations (denoted with parentheses) connected with arrows. Concept types are organized in a type lattice stored in the knowledge base. While Sowa's notation for concepts allows specification of both a concept type name and a referent, we replace the referent with a *type scope* indicator that is either null, >=, or >. The type scope indicators specify that the concept refers to the specified concept type *only*, to the specified concept type

and all narrower types, or to narrower types *only*, respectively. For example, the concept [politician] represents *only* the specific concept of a "politician" while [politician: >=] represents the concept of a politician or any kind of politician such as a senator, congressperson, or supreme court judge. [politician: >] refers to kinds of politicians, not to the concept of "politician" itself. Use of type scope indicators simplifies representation of information need and appears to simplify knowledge base creation and maintenance.

The relations that we use in this project are high-level, complex relations. Examples are *develops, spies on,* and *finances.* These relations could be defined by conceptual graphs containing concepts and more fundamental relations such as *agent* and *object.* We use complex relations because they tend to be represented in text by words rather than sentence structure and are therefore easier to recognize.

In this project, we use conceptual graphs only as a convenient notation for the concept-relation-concept form of an information need. Unlike previous work in information retrieval that has used conceptual graphs (Dick 1991; Liddy & Myaeng 1994), we do not do direct matching between a conceptual graph and the document collection. Instead, a conceptual graph is converted into a search expression and that search is run on the collection.

2.2. Knowledge Base and Recognition Expansion

Our knowledge base is our repository for concept and relation recognition knowledge. That is, it contains the knowledge used to translate a conceptual graph representation of information need into a form that can be evaluated by the search engine.

Both concept and relation types have *recognition expressions* associated with them. These recognition expressions are written in the Personal Librarian full-text search language with one extension: a concept, which may contain a conceptual graph, may be inserted anywhere in a recognition expression. This capability allows recognition expressions to refer to other concept and relation types rather than repeating the search expression for those types. For example, the concept type *cancer drug* has the recognition expression

[cancer: >=] NEAR/20 [drug]

The recognition expression for a relation type is similar to that of a concept type but must contain two parameters, represented as %1 and %2. These parameters correspond to the recognition expressions of the relation input and output concepts, respectively (we limited our implementation to binary relations). The following is the recognition expression for the relation type *finances:*

%1 NEAR/15 (financ* or fund* or (pay for) or contrib* or donation*) NEAR/15 %2

Recognition expansion is the process of transforming a conceptual graph into a full-text search language expression by replacing each concept and relation with its fully expanded recognition expression.

2.3. Document Collection

As in any information retrieval experiment, we need a large document collection. For this experiment, we need four large document collections: one for testing, two for the knowledge base refinement iterations, and one for the final test. TREC provided three CD-ROMs full of compressed text files for information retrieval research. The data is organized by source. Each of the collections we used was one source from one disk. Our initial training collection was Disk 1 Wall Street Journal (WSJ) data. The other collections,

in the order used, were: Disk 2 WSJ, Disk 3 Associated Press (AP), Disk 1 AP. The Disk 3 AP data was used to produce our official TREC-3 conference results. Due to technical problems, we were able to load only four of twelve months of this data.

2.4. Search Engine

The search engine is the final component in the process of identifying documents that are relevant to the information need. The search engine takes a search expression and returns a set of documents from the document collection. There are two basic kinds of search engines readily available: intelligent search engines that take natural language input and Boolean search engines that require a specific, Boolean-based search language. We chose to use Boolean searching for this experiment because Boolean provides us with greater control over the searching. We did not want the effects of an intelligent search mechanism confounding our results.

The search system used for this experiment is Personal Librarian, created and sold by Personal Library Software. Personal Librarian provides a robust Boolean search mechanism as well as an intelligent search mechanism that we did not use. We stress that we do not test the performance of Personal Librarian's natural language-based search mechanism in this experiment but only its robust Boolean and proximity search capabilities.

The Personal Librarian search language includes the standard Boolean operators AND, OR, and NOT, as well as unidirectional and bidirectional proximity operators. A unidirectional operator requires that the operands be in the specified order and proximity. A bidirectional operator requires only that the proximity criterion be met. In Personal Librarian, the unidirectional and bidirectional operators are W/*n* and NEAR/*n*, respectively, where *n* is the proximity in words. For example, the search "A W/10 B" will return only documents where word A occurs ten or fewer words *before* word B. "A NEAR/10 B" will return all documents where word A occurs within ten words of word B, regardless of the order.

In the experiment, we compare the performance of each relation search expression to ten alternatives: both unidirectional and bidirectional proximity operators for 5, 10, 15, 20, and 30 words. We had hoped to also use *within sentence* and *within paragraph* proximities but these operators were not available in the search engine used. Figure 1 is simplified to show the process for the relation and a single proximity.

Personal Librarian takes the search expression produced by recognition expansion and returns a list of documents. We use a utility program to change this list into a form that can be processed by the TREC evaluation program, discussed below. The results returned by Personal Librarian are ordered by the Personal Librarian search engine's proprietary algorithm.

2.5. Relevance Judgments

Evaluating the performance of a retrieval method requires relevance judgments specifying which documents in the collection are relevant for an information need and therefore should have been retrieved by the system. Relevance judgments were supplied by TREC. Ideally, relevance judgments would exist for all combinations of topics and documents. Since TREC now has 200 topics, the collection contains hundreds of thousands of documents, and relevance judgments must be made by people, only a small fraction of the possible relevance judgments exist. These judgments were created for the top ranked documents returned by systems used in previous TREC conferences.

The limited relevance judgments are issue for our experiment, as for all TREC participants. The TREC evaluation program treats all unjudged documents as not relevant. If a significant portion of the unjudged documents are relevant the results may be skewed.

2.6. Evaluation Program and Metrics

Once we have the list of documents retrieved and the relevance judgments, we can perform evaluation to produce our experimental results. TREC provides an evaluation program that outputs several different measures. The metric that we use in our experiment is non-interpolated average precision. Average precision corresponds to the area under an ideal (non-interpolated) recall/precision curve and is computed as the average of the precision at each relevant document retrieved. We use average precision as our primary metric for three reasons:

1. **It is a single measure of overall system performance.** Any retrieval method will provide some mixture of precision and recall. Since improving precision tends to reduce recall, and vice versa, those cannot be used for comparison. We needed one metric that would provide a measure of *overall* system performance, and average precision is the best metric of that type at this time.
2. **It is a standard metric that is widely used within the information retrieval research community.** Since retrieval performance is subjective, depending on the exact information need of each user, no metric can provide an indisputable measurement of system performance. Average precision is currently the most commonly used metric of overall system performance.
3. **The TREC evaluation program computes it.** This ensures that our computations of average precision would be standard.

Measurement of time for this experiment was "wall clock" time. Wall clock time measures the interval between when a query is submitted and results are displayed, which is the key measure of time relevant to retrieval system users. The standard Windows clock was displayed in digital mode. The start time was recorded when the Personal Librarian "Search" button was pressed, and the finish time was recorded when the search results *began* displaying. The clock display was in one second increments, so the precision of the time results is plus or minus one second.

2.7. System

We developed a system named *Teknos* that is used to build and maintain the knowledge base, maintain a list of conceptual graph queries, and perform recognition expansion on the graphs. The resulting search expression is copied (via the Windows clipboard) to Personal Librarian where the search is run. The results are output to a file and run through a utility program to provide a results file that can be processed by the TREC evaluation program.

Teknos is implemented in Borland C++ using the POET object-oriented database system and runs under Microsoft Windows 3.11. The computer used for this project is a Gateway 2000 486DX2 66 MHz PC with 12 MB of RAM and two hard drives: one has 424 MB of space, the other has 540 MB compressed using Stacker 4.0 to provide over 1 GB of space.

Topic 111

Topic Narrative: *A relevant document will report on efforts by the UN's International Atomic Energy Agency to monitor compliance with the Nuclear Non-proliferation Treaty, or, report on efforts by the United States, Britain, France, USSR, India, or China to control the transfer of technology, equipment, materials, or delivery systems to nations suspected of nuclear weapons development programs, or, a relevant document will report any nuclear activities by Argentina, Brazil, Iraq, Israel, North Korea, Pakistan, South Africa, or Iran (all suspected proliferators).*

Conceptual Graph: [nuclear wannabe: >] → (develops) → [nuclear weapon].

Graph Narrative: *A relevant document will report on nuclear weapons development programs in Argentina, Brazil, Iran, Iraq, Israel, North Korea, Pakistan, or South Africa (all suspected proliferators).*

Topic 122

Topic Narrative: *A relevant document will report on any phase in the worldwide process of bringing new cancer fighting drugs to market, from conceptualization to government marketing approval. The laboratory or company responsible for the drug project, the specific type of cancer(s) which the drug is designed to counter, and the chemical/medical properties of the drug must be identified.*

Conceptual Graph: [company] → (tests) → [cancer drug].

Graph Narrative: *A relevant document will report on any company testing a new cancer fighting drug.*

Topic 129

Topic Narrative: *A relevant document will discuss reported espionage by entities of the Soviet government - KGB, GRU, etc. - conducted within the territory of the United States of America, or against U.S. diplomatic or military facilities overseas. Reported entrapment or involvement of U.S. citizens, residents, or employees in Soviet spying, be it overseas or within U.S. territory, is also relevant. However, espionage cases involving states linked to the USSR - Czechoslovakia, Bulgaria, Cuba, etc. - are NOT relevant, unless linkage to Soviet intelligence can be demonstrated.*

Conceptual Graph: [USSR] → (SpiesOn) → [USA].

Graph Narrative: *A relevant document will discuss any spying conducted by the USSR against the USA.*

Topic 131

Topic Narrative: *A relevant document will provide specified data on contracts awarded to McDonnell Douglas for the production of military aircraft for any nation, any military service, or of any aircraft type (interceptor, fighter-bomber, helicopter, etc.). The contract must be for completed airframes, NOT contracts for aircraft development, factory tooling, components, spare parts, services, etc. To be relevant, the document also must specify the number of aircraft to be delivered, the dollar size of the contract, and the aircraft type sought.*

Conceptual Graph: [McDonnell Douglas] → (WonContractFor) → [military aircraft].

Graph Narrative: *A relevant document will discuss McDonnell Douglas winning a contract for some kind of military aircraft. The contract must be for completed airframes or aircraft development, not factory tooling, components, spare parts, services, etc.*

Topic 150

Topic Narrative: *A relevant document will show how U.S. politicians (federal, state, or local — individually or as a group) pay for their election campaigns, the role played by "special interests" and contributors in the electoral process, allegations or evidence of campaign contributions buying political favors, and/or proposals to limit the cost of campaigns or "reform" electoral finance practices.*

Conceptual Graph: [politician: >=] → (Finances) → [campaign].

Graph Narrative: *A relevant document will discuss a politician financing his or her campaign.*

Figure 3. Narratives for TREC Topics Used in Experiment

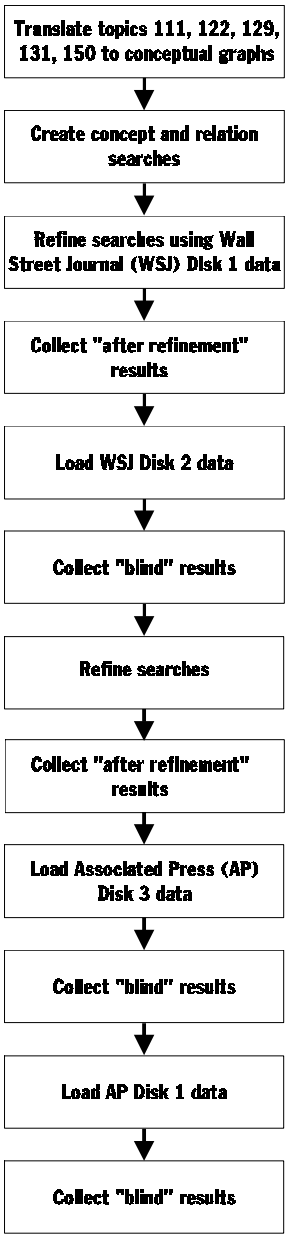

Figure 4. Actual Methodology

3. Experimental Process

This section discusses the details of our data collection process as shown in Figure 4.

To isolate the effects of single relation, a requirement of this experiment is that each query had to be represented by a single concept-relation-concept graph. Most TREC topics require several such graphs for a complete representation (Liddy & Myaeng 1994). We identified five topics that we could "bend" for our use: 111, 122, 129, 131, and 150. Topics 129 and 150 were reasonably well represented by a simple conceptual graph. The substance of Topic 131 was well represented but the topic contained a requirement for more detailed information than the conceptual graph could represent, hence the conceptual graph was more general than the topic. To get two additional topics, we allowed the topic to contain two primary concepts that were linked via multiple relations (Topic 122) or to have a major aspect of the topic representable with a simple conceptual graph (Topic 111).

The topic-graph mismatch for Topics 111, 122, and 131 means that the relevance judgments for these topics are only partially valid. This issue is discussed in more detail in Section 5. The topic narratives, conceptual graphs, and conceptual graph narratives are presented in Figure 3.

Our training collection was Wall Street Journal Disk 1, which was loaded into Personal Librarian. To develop the searches, we reviewed a few of the relevant documents in the test data and used them as guides to develop preliminary concept and relation searches. We then used an iterative process of searching, analysis using the TREC evaluation program, and refinement to improve the individual concept searches. When the concept searches worked reasonably well individually, we began testing the relation using the whole conceptual graph search. All of the relation searches are of the form

%1 *proximity* (*relation search*) *proximity* %2

where %1 and %2 are substituted with the concept searches linked by the in and out arrows, respectively. The two *proximity* operators are the same and the *relation search* is a search for words indicating the relation independent of the concepts. The first focus was refining the relation search using techniques similar to those used for the concepts. Next, we determined the most effective proximity to use. We experimented with same *proximity* operators used later in the experiment —unidirectional and bidirectional operators for 5, 10, 15, 20, and 30 words — collecting average precision results for each. The proximity

operator that produced the best results on the test data was used in the final version of the relation search. One of the benefits of using relations is that recognition criteria such as proximity can be experimentally determined during knowledge engineering.

Once the searches were complete, we collected final results for WSJ Disk 1, which were "after refinement" results. We then loaded the WSJ Disk 2 data and collected "blind" results. Our methodology called for refinement of all searches and subsequent collection of "after refinement" results. We determined that a search needed refinement if the results for a new collection were significantly worse than they had been for an earlier collection and failure analysis indicated that the cause was relation words that occurred in the new collection but not in the earlier one. We actually needed to refine the searches and collect "after refinement" results for Topic 111 only. The performance of the other four topics did not indicate a need for further refinement of the searches, though the performance of Topic 122 indicated other issues. The results from the relation-based search for topic 122 were much worse than for the proximity operators and preliminary analysis indicated that this was largely due to the topic-graph mismatch and the limited number of relevance judgments.

We decided to create custom relevance judgments for Topic 122 and the WSJ Disk 2 data to eliminate these problems. The set of documents judged was the set of 208 documents returned by the broadest proximity operator used in the experiment, NEAR/30. Due to time and funding constraints, judgments were created by one of the authors of this paper rather than an independent judge. We used the conceptual graph narrative for topic 122 as the guide to relevance and focused on being as objective as possible.

Because of the small amount of search refinement that we had to do with the WSJ Disk 2 data, we modified our experimental methodology to use both of the remaining collections for blind results only.

Our third collection was the Associated Press (AP) Disk 3 data used for the official TREC-3 runs. This was the first collection used for blind results only. Our plan was to load all of the AP Disk 3 data. Unfortunately, technical difficulties in indexing caused us to load only one third of that data (four of twelve months). Two of our five relation-based searches — topics 122 and 131 — returned no documents on this collection. We submitted our results to TREC for the three searches that did return documents. Since we needed to have results for eleven searches for each topic (one relation-based search and ten proximity-based searches), our TREC-3 results appear as eleven runs. Within each run are the results for the three topics that worked.

The fourth and final collection used was the AP Disk 1 data. This was the second collection used for blind results only. The relation-based search for topic 131 did not return any documents on this collection. There was one relevant document in the collection, which *was* returned by all of the proximity-based searches. As with the official TREC-3 run, all results from this collection were "blind" and based on the standard TREC relevance judgments (the judgments resulting from TREC-2, not the new TREC-3 judgments).

4. Results

Table 1 is a summary of the results of this experiment, showing the relative improvement of the relation-based search over the best proximity alternative and the relative increase in time required. The detailed performance results are presented graphically in Figure 5 through Figure 9. Performance results are reported as *precision ratios* which allow easy comparison between a relation-based search and its proximity-based alternatives. A precision ratio is computed as

$$\text{precision ratio (operator)} = \frac{\text{average precision (operator)}}{\text{average precision (relation)}}$$

so that the value for the relation-based search is 1.000. Any operator whose performance is greater than 1.000 did better than the relation-based search; any operator whose performance is less than 1.000 did worse.

A few notes on the results:

1. The WSJ Disk 2 results for Topic 111 are the "after refinement" results
2. The WSJ Disk 2 results for Topic 122 for both the standard TREC relevance judgments and our custom judgments are shown.
3. Due to technical difficulties and time constraints, we did not collect time search times for the AP Disk 3 data

Topic	WSJ Disk 1		WSJ Disk 2		AP Disk 3		AP Disk 1	
	Ave Prec	**Time**	**Ave Prec**	**Time**	**Ave Prec**	**Time**	**Ave Prec**	**Time**
111	+18.7%	+328%	-5.1%	+222%	-13%		+9.7%	+109%
122	-7.3%	+29%	***+55%***	***+26.7%***			+18.6%	+33%
129	***+2073%***	***0%***	***+1200%***	***+5.7%***	+252%		***+676%***	***+3%***
131	***+82.3%***	***+77%***	***+227%***	***+71%***				
150	+8.8%	+164%	+22.2%	+182%	+176%		***+114%***	***+73%***

Table 1: Relative Improvement of Relation vs. Best Alternative
Italics indicate where the percent performance increase is greater than the percent time increase

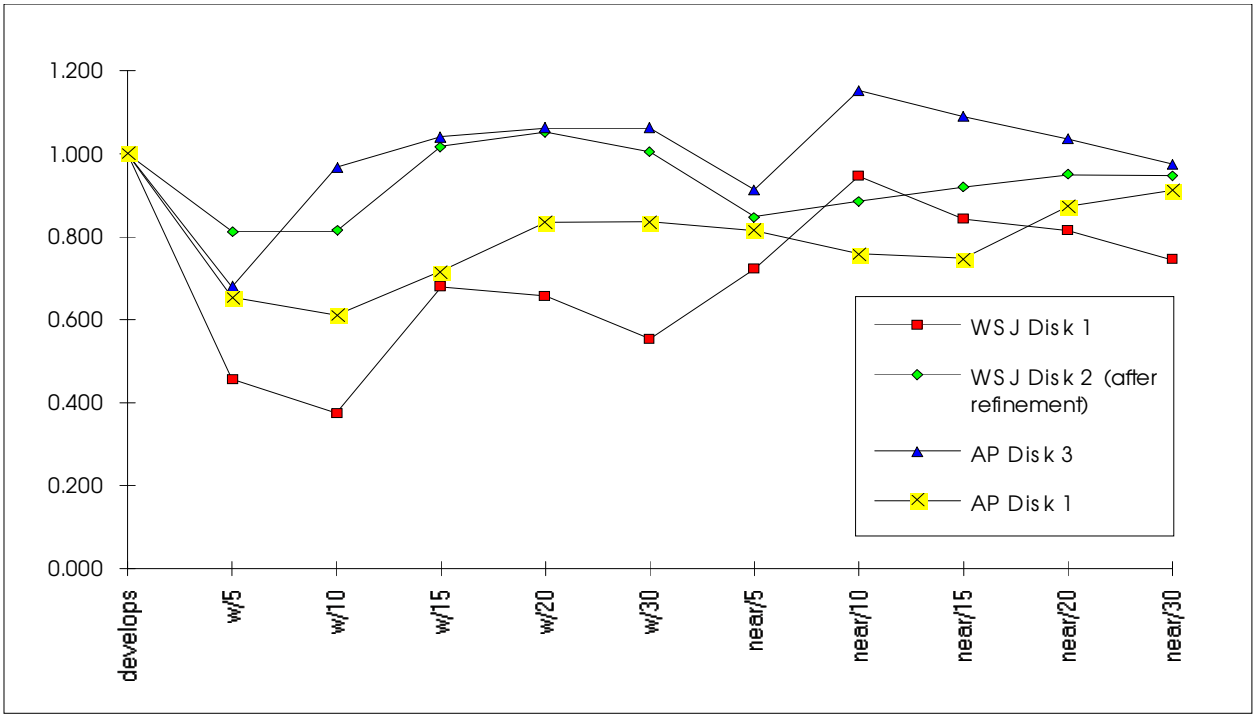

Figure 5. Topic 111 Precision Ratio

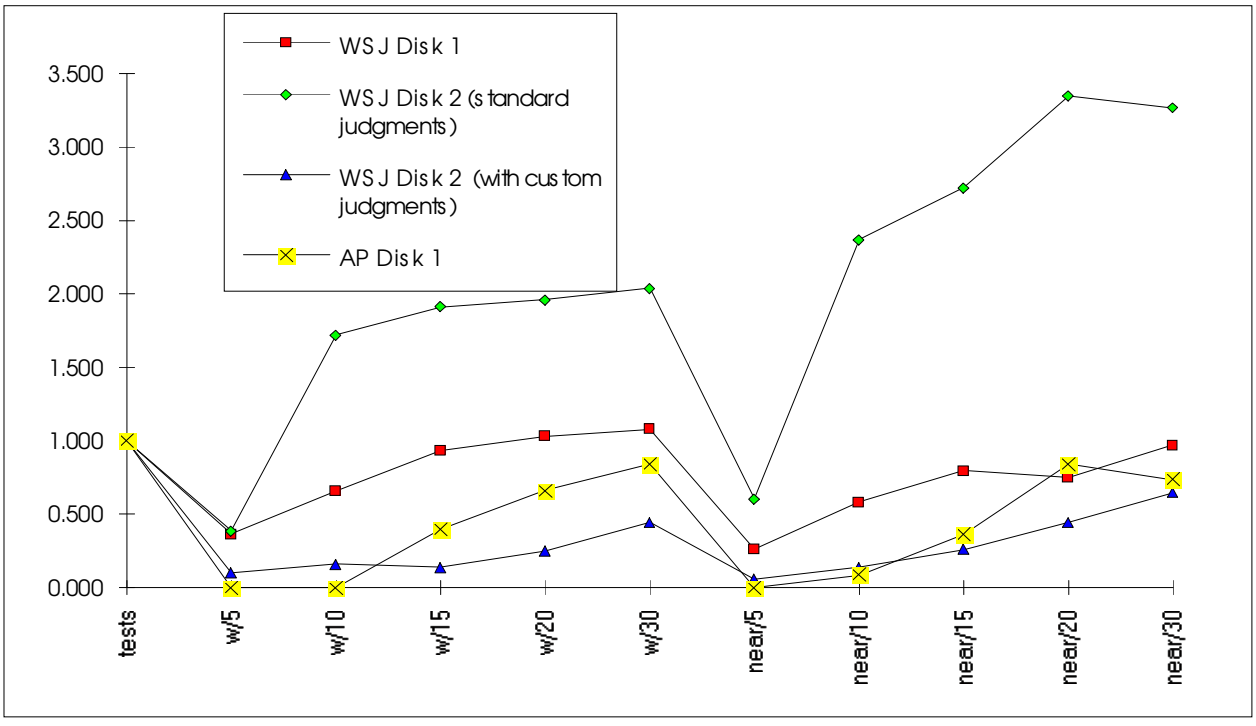

Figure 6: Topic 122 Precision Ratio

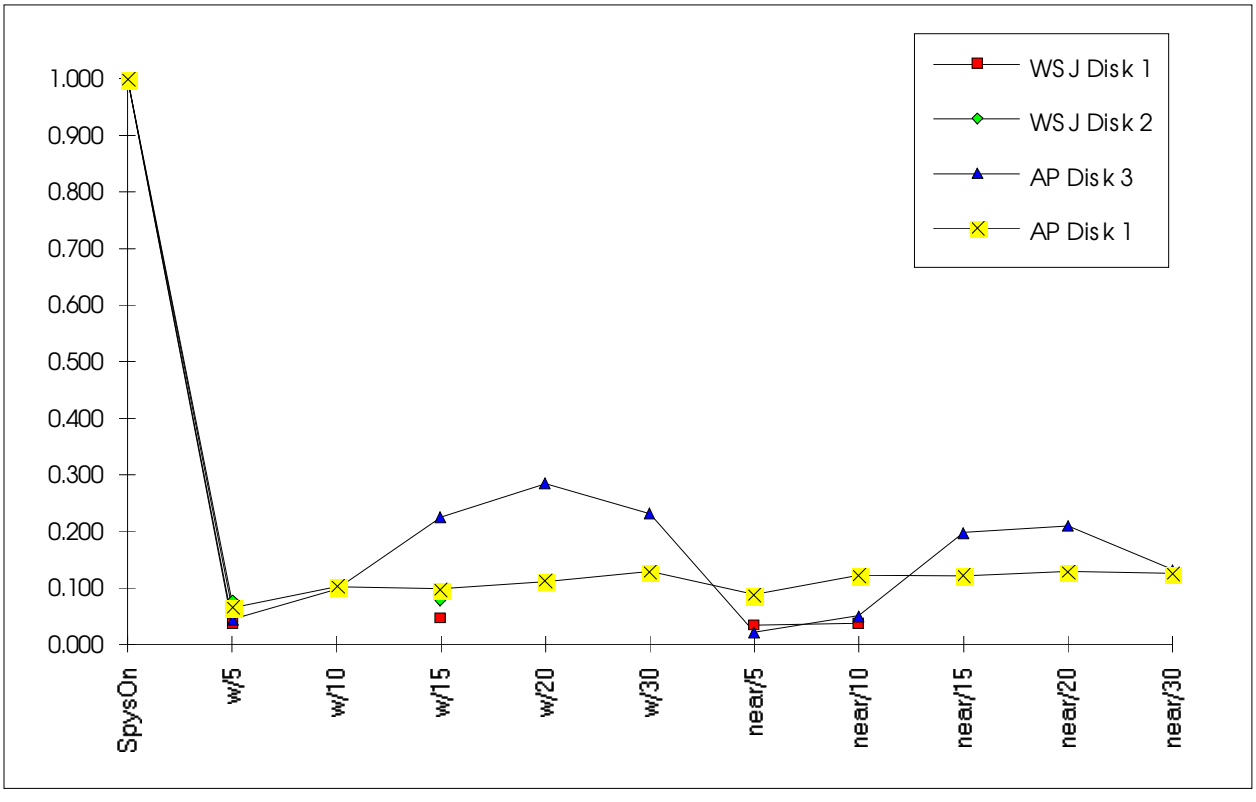

Figure 7: Topic 129 Precision Ratio

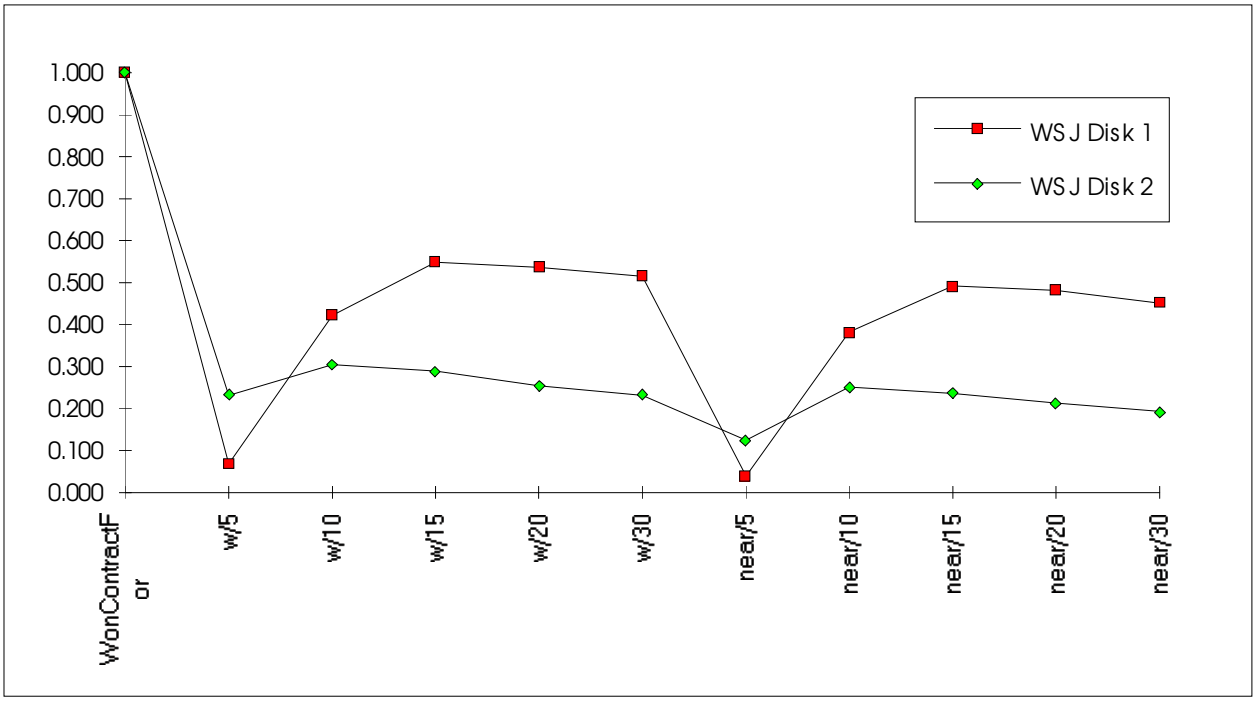

Figure 8: Topic 131 Precision Ratio

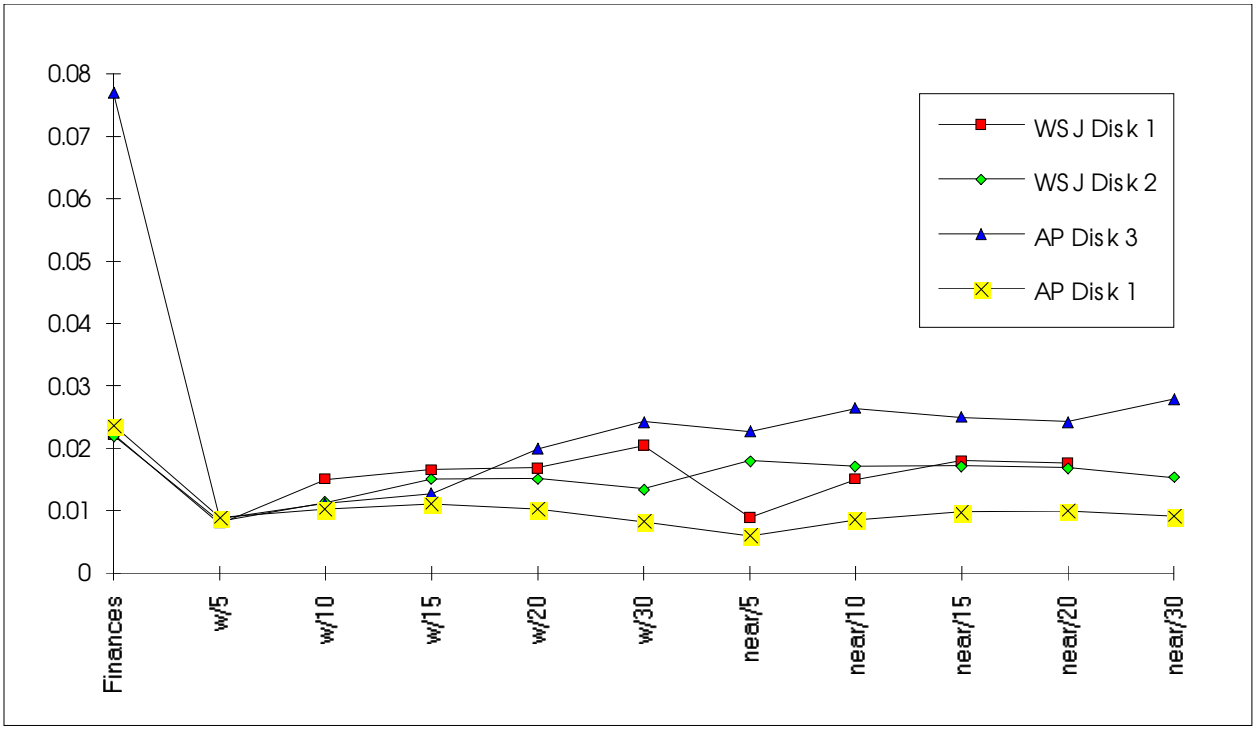

Figure 9. Topic 150 Precision Ratio

5. Analysis of Results

The methodology that we use in this experiment lends itself to more of a qualitative than quantitative analysis. Therefore, we discuss what we learned from each topic separately. The topic narratives, conceptual graphs, and conceptual graph narratives for the topics used in the experiment are shown in Figure 3. Prior to discussing the topics, we discuss our failure analysis process.

5.1. Failure Analysis

Two key issues that affect the validity of our results are the topic-graph mismatch, discussed in Section 2.1, and the limited number of relevance judgments, discussed in Section 2.5. We perform failure analysis on the WSJ Disk 2 results to determine how these issues affect the validity of our results. For each relation-based search we create three lists:

- **Hits.** A hit is a document returned by the relation-based search that is relevant in the TREC judgments.
- **False hits.** A false hit is a document returned by the relation-based search that is not relevant in the TREC relevance judgments.
- **Misses.** A miss is a document that is relevant in the TREC relevance judgments but was not returned by the relation-based search.

For each of these lists, we manually evaluate a random sample of the documents to determine if those documents are relevant to the conceptual graph representation of the topic. The list is sorted by date and every *n*th document is evaluated. Relevance of a document with respect to the conceptual graph is determined using the conceptual graph narrative as a guide. These results are organized in tables. For example, Table 2 shows the failure analysis results for Topic 111. The rightmost column, "% Deviation from TREC Judgments" shows the strength of the evidence that the TREC judgments are not valid for the conceptual graph. The equations for the deviations are as follows:

$$\text{Hits deviation} = 1 - \frac{\#\ \text{Hits Relevant to Graph}}{\#\ \text{Hits Analyzed}}$$

$$\text{False Hits deviation} = \frac{\#\ \text{False Hits Relevant to Graph}}{\#\ \text{False Hits Analyzed}}$$

$$\text{Misses deviation} = 1 - \frac{\#\ \text{Misses Relevant to Graph}}{\#\ \text{Misses Analyzed}}$$

Hits deviation is high when the documents returned by the relation-based search that are relevant to the TREC topic are not relevant to the conceptual graph. Hits deviation is often high when the conceptual graph is more specific than the topic. High Hits deviation indicates that the actual performance of the relation may be worse than computed using the TREC judgments. False Hits deviation is high when many of the false hits are relevant to the conceptual graph. There are two causes of False Hits deviation being high: the conceptual graph being more general than the topic and missing relevance judgments. Misses deviation is high when few of the misses are relevant to the conceptual graph. Like Hits deviation, this is primarily an issue when the conceptual graph is more specific than the topic. High False Hits deviation or Misses deviation indicates that the actual performance of the relation may be better than the value computed using the TREC judgments.

Failure analysis was performed on the WSJ Disk 2 collection. Results of failure analysis are presented in the individual topic discussions.

One limitation of our failure analysis process is that it indicates how the results of the relation-based searches might improve with better relevance judgments but gives no information on how the proximity-based alternatives would be affected.

5.2. Topic 111

The conceptual graph representing Topic 111 is more specific than the topic in that there are several concepts described in the TREC topic narrative that are not subsumed by the conceptual graph: "International Atomic Energy Agency", "Nuclear Non-proliferation Treaty", and "nuclear nations controlling nuclear technology transfer". Documents mentioning these concepts are relevant for the TREC topic but are not relevant to our conceptual graph.

The results for Topic 111 indicate that the relation performed marginally better than the proximity alternatives, although there were situations where proximity was better. For this topic, co-occurrence of concepts in proximity appears to be a good predictor of relevance in itself. Including specific relation words in the search does not appear to have significant benefit for this topic. Discussion of a nuclear wannabe near discussion of nuclear weapons generally indicates a connection between the two.

Failure analysis results for Topic 111 are shown in Table 2. This table shows that 10 items in the Hits list were analyzed, and all 10 of them are relevant to the graph. This means that our judgments and those of TREC agree completely for the Hits list, a deviation of 0%. Of the 10 False Hits analyzed, 7 are relevant to the graph, giving a 70% deviation. To have complete agreement with the TREC judgments, none of the False Hits would be relevant to the graph. Of the 12 Misses analyzed, 4 are relevant, giving a 67% deviation. For the Misses to have complete agreement with the TREC judgments, all of them would be relevant to the graph.

The deviations indicate that the real performance of the relation-based search should not be any worse than reported and may be significantly better. That is, the relation-based search results would likely improve significantly with relevance judgments that are more complete and based on the conceptual graph rather than the TREC topic. Due to time and resource constraints, we were unable to create custom relevance judgments for this topic as we did for Topic 122.

List	# Analyzed	# Relevant to Graph	% Deviation from TREC Judgments
Hits	10	10	0%
False Hits	10	7	70%
Misses	12	4	67%

Table 2. Topic 111 Failure Analysis Results

5.3. Topic 122

The conceptual graph for Topic 122 is more specific than the topic in some regards and more general in others. The conceptual graph is more specific in that it does not cover any phase of development other than testing, such as research, evaluation, or marketing. We chose to focus on "testing" for this topic because it is a clearly defined, specific relation. The conceptual graph is more general than the topic in that it does not require as complete information about the specific types of cancers the drug is designed to counter and the chemical/medical properties of the drug. Searching for these details would required more than a single concept-relation-concept graph.

Using the standard judgments, the performance of the relation-based search ranges from slightly worse to much worse than the best proximity alternative.

Failure analysis results for Topic 122 are shown in Table 3. The Hits deviation indicates that the real performance of the relation-based search might be significantly worse than reported. The False Hits and Misses deviations indicate that the performance might be significantly better. These failure analysis results strongly indicate that the conceptual graph is a poor match for the relevance judgments. To obtain more valid performance results, we created custom relevance judgments for this topic.

The graph in Figure 6 shows the dramatic difference in the performance reported using the standard judgments versus using the custom judgments. Using the custom judgment results, the relation-based search performance is significantly better than the proximity-based alternatives.

List	# Analyzed	# Relevant to Graph	% Deviation from TREC Judgments
Hits	7	3	57%
False Hits	10	7	70%
Misses	10	0	100%

Table 3. Topic 122 Failure Analysis Results

5.4. Topic 129

Topic 129 was easily representable as a simple concept-relation-concept graph. The failure analysis results shown in Table 4 show no deviation of any kind, suggesting that the results are valid.

The relation-based search performed much better than the proximity operator alternatives. The proximity operators, especially the broad ones, returned thousands of documents. Writing these results to a file took as much as an hour for each operator tested. Early results were so dramatic and collecting data for each operator took so long that results for some operators were not collected.

Topic 129 is an ideal topic for using relations. The two concepts USA and USSR co-occur frequently in the collections with a very large number of relations. Specifying the relation led to large performance improvements.

List	# Analyzed	# Relevant to Graph	% Deviation from TREC Judgments
Hits	4	4	0%
False Hits	10	0	0%
Misses	3	3	0%

Table 4. Topic 129 Failure Analysis Results

5.5. Topic 131

While the conceptual graph representing Topic 131 effectively captures the primary meaning of the topic, it does not require *"the number of aircraft to be delivered, the dollar size of the contract, and the aircraft type sought"*. Hence, the conceptual graph is more general than the topic: there are documents that are relevant to the conceptual graph but not relevant to the topic.

The performance results for this topic show that the relation-based search performed significantly better than the proximity-based alternatives. Topic 131 is a good topic for using relations. The two concepts co-

occur frequently with many different relationships. One of the most common relations is simply the implicit indication that McDonnell Douglas is the builder of the aircraft mentioned: *". . . a McDonnell Douglas F-15 . . ."*

The failure analysis results shown in Table 5 suggest that the real performance of the relation-based search was no worse than reported and is likely to be considerably better. None of the Hits are not relevant to the graph but many of the False Hits are relevant to the graph and many of the Misses are not relevant to the graph.

List	# Analyzed	# Relevant to Graph	% Deviation from TREC Judgments
Hits	4	4	0%
False Hits	11	6	55%
Misses	8	3	63%

Table 5. Topic 131 Failure Analysis Results

5.6. Topic 150

Like Topic 129, Topic 150 was easily and effectively representable as a simple conceptual graph. This topic showed interesting behavior in that the performance of the relation-based search was only marginally better than the alternatives for the first two collections. This surprised us, as the concepts politician and campaign seem to be ones that would frequently co-occur with many different relations: finances, starts, makes vows in, discusses issues in, etc. After some consideration, we realized that the first two collections were both Wall Street Journal, which is fundamentally focused on financial issues. We hypothesized that performance of the relation-based search would be much better on a collection with a broader focus. The tests run on the Associated Press data support this hypothesis: the relation-based search performs much better than the alternatives.

The failure analysis results shown in Table 6 indicate that the performance results for this topic are valid. The one Hit document that was judged not relevant to the conceptual graph was marginally relevant to both the topic and the conceptual graph. That TREC judged it as relevant and we judged it as not relevant appears to be more a variation between judges than a graph-topic mismatch.

List	# Analyzed	# Relevant to Graph	% Deviation from TREC Judgments
Hits	10	9	10%
False Hits	11	0	0%
Misses	10	10	0%

Table 6. Topic 150 Failure Analysis Results

6. Conclusions

Can retrieval performance be improved by explicitly specifying and searching for relationships between concepts?

The data we collected provides evidence that, in some cases, retrieval performance can be improved by explicitly searching for relationships. In all but two of our tests, relations provided improved retrieval performance over all of the proximity-based alternatives. In the best case, performance was twenty times

better. In the worst case, performance was only slightly worse than the best alternative. In these worst cases, failure analysis suggests that the relation would have performed significantly better than the alternatives if we had used complete relevance judgments based on the conceptual graph rather than the topic. Hence, we believe that relations will often improve performance.

Under what conditions is the improvement likely to be justified?

Our experience suggests that using a relation for retrieval provides the greatest performance improvement when the two concepts it links frequently occur in proximity to each other with several different relationships. The effectiveness of relations is dependent on the concepts, the relation, and the data being searched. Some pairs of concepts, like *nuclear wannabe* and *nuclear weapon* (Topic 111), inherently have a small number of relationships between them. For these concepts, using relations will generally provide a small increase in performance. Other pairs of concepts inherently have a large number of relations, such as *U.S.A.* and *U.S.S.R.* (Topic 129). With these concepts, specifying a relation provides a large increase in performance and, for many purposes, may be essential to getting usable results. For some pairs of concepts, such as *politician* and *campaign* (Topic 150), the number of relations depends on the nature of the document collection. In this case, using a relation that is very common (such as *Finances* in the WSJ collection) provides relatively little performance gain but using a less common relation provides significant benefit.

The cost and value of using relations are both subjective measures that depend on the system and the user's needs. There are at least three costs associated with using relations: reduced recall, increased retrieval time, and the cost of building the supporting knowledge base.

Like any method of improving precision, relations do so at the expense of recall. The difference is that, for a given improvement in precision, relations may cause less degradation of recall than other methods, such as tightening proximity. Use of relations may provide significant benefit to a user whose focus is on precision or a balance of precision and recall. For a user who requires high recall, the degradation in recall caused by relations may not be acceptable.

Explicitly searching for relations increases the complexity of the search expression and therefore increases the processing time cost associated with the search. In Table 1 we show the percentage increase in performance provided by the relation-based search compared to the best proximity alternative. We have italicized the cases where the relative increase in performance exceeds the relative increase in time. This cutoff point is not necessarily the point where the time cost of using relations exceeds the value. A 300% increase in time is probably cheap if the current response time is 0.1 second but prohibitive if the current response time is 30 seconds. Similarly, a 30% improvement in performance is more significant when there are 1000 documents returned than when there are 10. In Table 1 note that the greatest performance improvements have the lowest time cost while the smallest performance improvements tend to have high time cost.

The approach to recognizing concepts and relations that we used in this experiment requires a knowledge base that takes substantial time and expertise to construct. The number of concepts and relations required in such a knowledge base depends on the domain to be searched but could easily be in the thousands or tens of thousands of concepts and hundreds of relations. If each concept and relation takes one hour to define, a knowledge base of 10,000 concepts and 200 relations would require 4.6 person years to construct.

7. Research Questions For Future Work

While the results of this experiment indicate that relations provide significant performance advantages, there are several related questions that could affect the ability to use this approach in a retrieval system.

Will our relation definitions work with other concepts? We defined five relations, each of which was used for a single pair of concepts. Are the searches for these relations general enough that they will work properly with a wide variety of concepts? If not, can they be redefined so that they do?

Can human users consistently identify the same relation for the same information need? One use of relations is to have a list of concepts and relations and allow the user to build a conceptual graph representing his or her information need. Given an information need, two specified concepts, and a list of relations, would users consistently identify the same relation between the concepts? What information or mechanisms could be provided to aid a user in selecting the correct relation?

How well can directionality be detected? Conceptual graphs are, by definition, directional. This is essential, as there is a major difference between the U.S.S.R. spying on the U.S. and the U.S. spying on the U.S.S.R. The relation-based searches in this experiment all used bidirectional search operators and therefore ignored directionality. What can be done to better represent directionality in relation-based searches?

Can relations be identified more effectively using a different (non-Boolean and proximity-based) search mechanism? In this experiment we used Boolean searching. How might other search mechanisms be used, and how do those results compare to ours?

Does explicitly specifying relations help users better frame their information need? That is, does the process of selecting a relation from a list or otherwise explicitly specifying the relation help the user to better understand what he or she is looking for?

8. Acknowledgments

The authors would like to thank Donna Harman for her patience and flexibility with our non-standard TREC experiment, Personal Library Systems for the use of Personal Librarian for this experiment, and Howard Turtle for his comments on our project plan and an earlier draft of this paper.

9. References

Brooks, B.C. (1986). Jason Farradane and relational indexing. *Journal of Information Science, 12,* 15-18.

Cimino, C. & Barnett, G.O. (1992). Analysis of physician questions in an ambulatory care setting. *Computers and Biomedical Research, 25*, 366-373.

Dick, Judith P. (1991). On the usefulness of conceptual graphs in representing knowledge for intelligent retrieval. *Proceedings of the Sixth Annual Workshop on Conceptual Graphs. Binghamton, NY, July 11-13, 1991.* 153-167.

Farradane, J. (1980). Relational indexing. *Journal of Information Science*, *1,* 267-276, 313-324.

Farradane, J. & Thompson, D. (1980). Testing relation indexing by diagnostic computer program. *Journal of Information Science, 2*, 285-297.

Liddy, E.D. & Myaeng, S.H. (1994). DR-LINK: A system update for TREC-2. In Harmon, D.K. *Proceedings of the Second Text REtrieval Conference (TREC-2),* NIST Special Publication 500-215, 85-99.

McCune, Brian P., Tong, Richard M., Dean, Jeffrey S., & Shapiro, Daniel G. (1980). RUBRIC: A system for rule-based information retrieval. *IEEE Transactions on Software Engineering, SE-11(9),* 939-945.

Sowa, J.F. (1984). *Conceptual Structures: Information Processing In Mind and Machine.* Reading, MA:Addison-Wesley.