Evaluation of Document Retrieval Technologies to Support Access to Presidential Electronic Records

PERPOS Technical Report ITTL/CISTD 02-3
December 2002

William E. Underwood
Matthew G. Underwood

Information Technology and Telecommunications Laboratory
Computer Science and Information Technology Division
347 Ferst St.
Atlanta, GA 30332-0832 USA

The sponsors of this research are the Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA). The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

Abstract

Archivists need to respond to Freedom of Information Act (FOIA) requests for electronic records that have not yet been systematically processed. This paper describes three text-based retrieval technologies—Boolean, statistical, and natural language-based — that can be used to find electronic records that are relevant to FOIA requests. Document retrieval experiments were conducted with document retrieval systems corresponding to each of the document retrieval technologies. WebGlimpse was used as an example of a Boolean text-based retrieval technology; Oracle Text with word queries as an example of statistical search technology; and Sun's NOVA Precision Content passage retrieval system as an example of natural language-based search technology. Queries used in the experiments were derived from actual FOIA requests submitted to the Bush Presidential Library. The experiments were conducted using the Bush Public Papers as a sample collection.

The results of the experiments are analyzed to explain the difference in performance for different topics. Oracle Text with word queries had the best performance with regard to average precision, and especially for broad general queries with many alternatives. NOVA's Precision Content Retrieval, while not performing as well overall, out performed Oracle Text on topics where the request was for specific information, and the query involved just a few words. NOVA's performance would have been better if the user interface allowed a larger number of passages to be retrieved and relevancy feedback had been used to refine the NOVA queries. WebGlimpse, using a Boolean search technology without relevance ranking, did not perform as well as the other search technologies.

The average precision of the systems evaluated in this paper is significantly greater than the average precision of the document retrieval systems evaluated in the Ad Hoc Query Track of the Eighth Text Retrieval Conference (TREC-8). Precision is dependent on the size of the document set searched and is typically lower for larger document sets. The experiments described in this paper were conducted on a document set of about 5000 documents. There were more than 500,000 documents in the TREC-8 document set. Hence, the results of the experiments reported in this paper are not conclusive. The experiments should be conducted again using a larger corpus, for instance, the Bush Administration personal computer or e-mail records.

Keywords: precision content retrieval, Boolean retrieval, natural language processing, FOIA requests, document retrieval, passage retrieval

Table of Contents

Introduction

To facilitate access to documents in the Bush Presidential Library, the staff makes available to researchers the finding aids that were used by the White House Office of Records Management (WHORM) during the presidential administration. These include the White House Office of Records Management Subject (WHORMS) Filing Categories and the Alphabetic Name (ALPHA) files list. The staff creates additional finding aids for documents that were not managed by the WHORM, for example, folder title lists for Staff Member and Office Files (SMOFs). A researcher can used these finding aids to identify collections and folders that might contain documents relevant to their interests.

A researcher can only immediately see or obtain copies of documents that have been opened to the public. That is to say, the archival staff has reviewed those documents to ensure that they do not contain information that is exempt from release under provisions of the Freedom of Information Act (FOIA) or that are restricted from release under provisions of the Presidential Records Act (PRA). If a researcher wants to obtain copies of documents that have not been reviewed, he may submit a FOIA request, that is, request documents under a provision of the Freedom of Information Act.

To respond to FOIA requests, the Bush Presidential Library uses the finding aids, and estimates the volume (number of pages) that might be relevant to the request. However, the Bush Presidential electronic records, including electronic mail files and personal computer (PC) files, do not have finding aids. Finding aids can be created for the PC files systems by using the Archival Processing Tool (APT), a prototype system that supports archivists in processing electronic records [Underwood, et al, 2002]. However, archivists may need to respond to FOIA requests for records in the unprocessed email files or the unprocessed electronic file systems of staff members and offices. There is an opportunity to use text-based document retrieval technologies to index, search and retrieve Presidential electronic records from the unprocessed PC file systems and email files.

The purpose of this paper is to describe an evaluation of text-based retrieval technologies in support of responding to FOIA requests. The primary question that we are seeking to answer is: Are there significant differences in the performance of document retrieval technologies in supporting search of the unprocessed digital collections in response to FOIA requests?

The next section reviews document retrieval technologies and describes systems representative of these technologies that were used in document retrieval experiments. The third section describes the collection of Bush Presidential documents used in the experiment. The fourth section describes the FOIA requests and queries used in the experiments. Section 5 describes the evaluation measures and experimental methodology. Section 6 presents and analyzes the results of the experiment. The final section summarizes our conclusions and discusses issues needing further investigation.

Document Retrieval Technologies

Boolean Document Retrieval Technology

Boolean techniques for document retrieval involve word stemming, indexing, and Boolean queries with wildcards. *WebGlimpse* is an example of a commercially available document retrieval system that uses Boolean techniques [Manber and Wu, 1993; Manber et al, 1998]. The Bush Presidential Library currently uses it to provide access to the Bush Presidential Public Papers.[1]

For Boolean queries, WebGlimpse uses a semicolon (;) for AND, a comma (,) for OR and a tilde (~) for NOT. Logical operators are evaluated left to right, for instance,

POW,MIA;Vietnam,Laos

is equivalent to

((POW,MIA);VietNam),Laos

Curly braces {} are used to group words, phrases and operators, for instance, the query

{POW, MIA};{Vietnam, Laos}

will find documents containing POW or MIA and containing Vietnam or Laos.

If words separated by blanks are entered in a query, WebGlimpse will search for the phrase containing the words, for instance

Radio marti,TV marti

will find documents containing the phrase *Radio marti* OR the phase *TV marti* or containing both.

The symbol '#' is used to denote a sequence of any number (including 0) of arbitrary characters. For instance, the query *#lang#* will find documents containing *slang*, *language*, and *clangs*. The symbol % is used to denote one or zero arbitrary characters. For instance, U%S% will match the abbreviation for United States, U.S. or US.

The HTML versions of the Bush Public papers contain markup such as Y:89, M:01, and D:20 indicating the year, month and day the document was created. If such expressions are included in a WebGlimpse query, they will match documents produced in the corresponding years and months.

[1] http://bushlibrary.tamu.edu/papers/

Statistical Document Retrieval Technology

Statistical document retrieval techniques extend the capabilities of Boolean systems by using frequency of occurrence and relevance ranking. Oracle Text with Word Queries is a statistical document retrieval technology [McGregor, 1999, 2002].[2]

Oracle Text with Word Queries assumes that the document set to be searched has been indexed and the document collection is located in a datastore. Our experiments use a CONTEXT indextype because it offers the most latitude for searching on large, cohesive documents stored as rows of a single database column. There are other indexing possibilities for documents that have more complex structure (e.g., multiple columns or XML structures).

There are several datastore options depending on where the documents actually reside. URL_DATASTORE provides for documents stored on intranets or on the internet, but for our experiments we used a DIRECT_DATASTORE whereby the documents were input into a local database table. By using a FILE_DATATORE or URL_DATASTORE, only a reference to the document location would need to be stored in the database.

The SQL language is used to search the database table using a CONTAINS operator in the WHERE clause of a SELECT statement. CONTAINS queries are used on collections that have been indexed as a CONTEXT index. The CONTEXT/CONTAINS method of indexing and searching provides for theme indexing the document set. Theme indexing allows for the use of the ABOUT operator and a knowledge base for query expansion (Discussed in the next section on Natural Language-based Document Retrieval).

The CONTAINS query returns a relevance ranking for each document by referencing the SCORE operator in the SQL SELECT statement. The following example returns the identifier of all documents (docname) that score positively on a search matching the 'query' against the document column.

```
SELECT SCORE(1), docname FROM HtmlCorpus
     WHERE CONTAINS(document, 'query', 1) > 0 ORDER BY SCORE(1) DESC;
```

The SCORE is returned in descending (DESC) order from the most relevant to the least relevant document. The document can be returned as plain text or with the query terms highlighted. The CONTAINS operator should be followed by the threshhold operator symbol (>), which specifies that the score returned must be greater than some threshold value for the document to be considered relevant.

To calculate the relevance score for a document that matches a query, Oracle Text uses an inverse frequency algorithm based on Salton's formula.[3] Inverse frequency scoring assumes that for a document to score high, the query term must occur frequently in the document, but infrequently in the entire document set.

[2] Oracle Text was formerly known as ConText and interMedia Text.

[3] $3f(1+\log(N/n)$ where f is the frequency of the work in the document, N is the number of documents, and n is the number of documents containing the query term.

Oracle Text uses the basic Boolean operators AND (&), OR (|) and NOT (~). Parentheses can be used for grouping expressions.

A root word prefixed with a dollar sign ($), e.g., $broadcast, will find all documents containing its root word (stem) or derivatives, e.g., broadcasts, broadcasting, or broadcaster. The EQUIV operator (=) can be used to indicate that two or more words are equivalent, for instance (91=1991).

Using the ACCUM(ulate) (,) and weight (*) operators, one can increase the score for documents that match a query by weighting terms differently. For instance, in searching for documents related to the *Clarence Thomas nomination to the Supreme Court*, the expression

(justice, judge, Supreme Court*5, Clarence Thomas *10)

will increase the score of the term Supreme Court by 5 times and the term Clarence Thomas by 10 times. This signifies that documents related to *Clarence Thomas* and *Supreme Court* are most relevant to the query. The ACCUM operator gives the highest scores to documents that contain the terms within the scope of the operator; for instance, ACCUM (dog, pet, Millie) will give the highest score to documents that contain all three terms.

One can search for terms that are in close proximity with the NEAR operator. For example, to find all documents where Soviet is within 6 words of Revolution, the following query would be issued.

NEAR((Soviet, revolution), 6)

The default and maximum value for the NEAR operator is to search for terms separated by no more than 100 words.

In conjunction with Boolean operators, the NEAR operator constrains the scope of a query. Used with the section searching operator WITHIN, the NEAR operator can constrain the search to predefined zones (sentence, paragraph, HTML sections).

For the Bush Public papers, a document identifier is stored in the database column DOCNAME and the associated HTML document is stored in the DOCUMENT column. The document identifier is of the form YYMMDDNN, where YY is the last two digits of the year of the document, MM and DD the month and day of the month of the document, and NN is a sequential number of the document for that day. It is possible to restrict the search to documents prior or subsequent to a date. For instance, the SQL expression

DOCNAME < 89122100

enables one to restrict the search to only those documents prior to December 21, 1989.

Natural Language-Based Document Retrieval Technology

Document retrieval systems utilizing natural language technologies extend the capabilities of the statistical systems through inclusion of morphological, syntactic and semantic knowledge and techniques such as parsing and query expansion. There are a large number of commercially available systems that use natural language technologies for instance, Thunderstone's Webinator,[4] Readware's ConSearch,[5] Convera's RetrievalWare,[6] Autonomy's Conceptual Search,[7] and Oracle Text with about queries. Oracle Text with about queries uses a thesaurus (which Oracle calls a knowledge-base) to expand the query. For example, the query 'ABOUT (politics)' finds all documents that are about the subject politics, not just the documents that contain the word politics.

A difficulty with natural language-based document retrieval systems based on query expansion is that while they increase the recall of relevant documents, they do this at the expense of retrieval of many more irrelevant documents than would be retrieved with statistical document retrieval systems.

Precision content retrieval is a Natural Language-based passage retrieval technology being developed by Bill Woods and his colleagues at the Sun Microsystems Research Laboratory [Ambroziak and Woods, 1998]. The experimental system that uses this technology is called NOVA. In addition to morphological and syntactic knowledge, NOVA uses lexical subsumption to construct a conceptual index to passages in collections of textual documents. Lexical subsumption is the relationship of generality between concepts in which a term *X* subsumes a term *Y*, if *X* is more general than *Y*, or equivalently, if *Y* is more specific than *X*.

Lexical knowledge and subsumption are used to automatically construct a conceptual index of the passages in a set of documents from words and phrases extracted from the documents. Suppose that a set of documents contains the phrases *classification method* and *subsumption method*. Those phrases can be added to the conceptual index as indicators of the contents of the documents. Furthermore, since *subsumption* is a kind of *classification*, *classification method* subsumes *subsumption method*. Now, if someone requests a document or passage about *classification methods*, the passages containing references to *classification method* and *subsumption method* will both be retrieved. Suppose that someone requests a document about an *indexing technique*, and there is no document that contains that phrase. The passage about *classification method* will be retrieved, because *indexing* is a kind of *classification*, and *technique* is a kind of *method*.

NOVA uses a relaxation ranking method to rank the passages most relevant to a query. It penalizes each passage by an amount that reflects lack of confidence that the passage is relevant, and orders the passages with the smallest penalties first. Penalty-based scoring prefers passages in which all of the terms of the query occur exactly as they were typed,

[4] www.thundersone.com/texis/site/pages/webinator.html
[5] www.readware.com/prod_consearch.asp
[6] www.convera.com/Products/index.asp
[7] www.autonomy.com/Content/Products/IDOL/f/Conceptual_Search (Also a probabilistic retrieval model)

with no intervening words. Such passages receive no penalty. NOVA searches for passages in which the terms of the request occur with inflected forms or in which the terms of the query are matched with semantically related terms. For each departure from the exact query, a slight penalty is assessed. There is also a penalty for extra words occurring between the terms of the query, for the words of the query being in a different order, or for one or more words of the query not occurring in the passage.

For example, in the use of NOVA to search a database of the Bush Public Papers, the best match found in response to the request for "Radio TV Marti broadcast Cuba" was displayed as follows:

1. **99.695** **radio** tv tv marti broadcast cuba
   May 20, 1991

   White House and was released by the Office of the Press Secretary on May 20. In his message, President Bush referred to President Fidel Castro Ruz of Cuba. The message was **broadcast into Cuba with a Spanish translation on Radio and TV Marti**.

This says that the passage is ranked number 1 and had a penalty of 0.305 deducted from 100, which would have been a perfect match. The terms *radio* and *tv marti* and *broadcast* and *Cuba* all occurred. The penalty was because the terms were not in the exact order. The date, May 10, 1991, is actually a hypertext link to the document containing the passage. The content of the passage in which the terms were found is displayed to provide information about the match. The passage where the terms occurred is highlighted in red. This enables the requestor to decide whether they want to go to see the passage in the document by clicking on the hypertext link.

The fourth best match was the following:

4. **89.829** (missing: radio) tv tv marti television broadcasting cuba
   August 27, 1990

   Broadcasting to Cuba August 27, 1990 The President signed on Sunday, August 26, a Presidential determination that the tests of **TV Marti have demonstrated that television broadcasting to Cuba** is feasible and will not cause objectionable interference with the broadcasts of domestic television licensees. Our international telecommunications commitments have been observed

This passage had a penalty of 10 because it was missing an important word, *radio*. The additional penalty of 0.171 was due to the text having a more specific (lexically subsumed) phrase of the term broadcast and intervening words between TV Marti and broadcasting. It is nevertheless, a passage relevant to the query.

One of the useful features of NOVA is that it supports browsing of the conceptual index of a collection to find more general or specific concepts that might be better query terms

for indicating the researcher's interests. The following response to a query illustrates that feature. The desire was to find passages (documents) that were "Records relating to Japan (Trade and Economic Policy). The request to NOVA was "Japan Trade and Economic Policy." The response below was ranked 5th in relevance.

5. **98.223** japan various trade and economic economic policies January 9, 1992

that we together work to further promote the building of the new world order, the new world. And it is important that the United States continues to exercise leadership. And **Japan wishes to actively support those efforts by the United States. I believe that the meetings that I had with the President would mark a concrete first step towards the building of a Japan-U.S. global partnership. I had a candid exchange of views on various trade and economic issues as well. And in addition to steadily implementing our economic policies** as reflected in the joint statement issued yesterday, I believe we were able to engage in substantive discussions on various measures related to the automobiles and

The underlined terms in the first row match terms in the passage. They are also hyperlinks into the conceptual taxonomy. If one selects the phrase economic policies, one will see a display something like the following:

```
                _TOP_  (0)
           policy  (1269)
      economic policy  (25)
economic policies  (7) Show Instances
       administration's economic policies  (2)
       coordinate economic policies  (1)
       czechoslovakia's economic policies  (1)
       economic policies, coordination of  (1)
       economic policies, development of  (1)
       economic policies, realignment of  (1)
       fact good economic policies  (1)
       government economic policies  (1)
       have economic policies  (1)
       implementing economic adjustment policies  (1)
       implementing economic reform policies  (1)
       international economic policies  (1)
       market-based economic policies  (1)
       market-oriented economic policies  (2)
       new economic policies  (1)
       president chiluba's economic policies  (1)
       pursue economic policies  (1)
       reforming broad economic policies  (2)
       republican economic policies  (1)
       responsible economic policies  (1)
       sound economic policies  (9)
       statist economic policies  (1)
       uruguay's economic policies  (1)
       venezuela's economic policies  (1)
```

This part of the conceptual index, which looks something like a back-of-the-book index, was created from a base lexicon indicating semantic subsumption relationships of more than 15,000 words. To this were added more than 2,000,000 phrases that occur in the Bush public papers. This part of the conceptual index indicates that economic policy is a more general concept than economic policies, and policy is a more general term than economic policy, and policy is a root concept. If one selects economic policies or any of the terms below it in the conceptual index, one will be shown the number of documents indicated to the right of the term and the passages that contain the term. If one selects the more general terms economic policy or policy one will be shown the more specific terms subsumed by that term. For instance, if one selected policy, one would see something like the following.[8]

Under Foreign Policy, one finds a document on Japan's Foreign Policy. It also becomes clear than NOVA does not index on conjunctions of terms, so separate queries for "Trade Policy" and "Economic Policy" might be appropriate.

```
   _TOP_ (0)
policy  (1269)  Show Instances
      ...
      economic policy  (25)
            coherent economic policy  (1)
                  have coherent economic policy  (1)
            domestic economic policy  (2)
            economic policies  (7)
                  administration's economic policies  (2)
                  coordinate economic policies  (1)
                  czechoslovakia's economic policies  (1)
                  economic policies, coordination of  (1)
                  economic policies, development of  (1)
                  economic policies, realignment of  (1)
                  fact good economic policies  (1)
                  government economic policies  (1)
                  have economic policies  (1)
                  implementing economic adjustment policies  (1)
                  implementing economic reform policies  (1)
                  international economic policies  (1)
                  market-based economic policies  (1)
                  market-oriented economic policies  (2)
                  new economic policies  (1)
                  president chiluba's economic policies  (1)
                  pursue economic policies  (1)
                  reforming broad economic policies  (2)
                  republican economic policies  (1)
                  responsible economic policies  (1)
                  sound economic policies  (9)
                  statist economic policies  (1)
                  uruguay's economic policies  (1)
                  venezuela's economic policies  (1)
            economic policy, division of  (1)
```

[8] The ellipses (…) indicates that portions of the conceptual index are not shown in this example. The portion of the conceptual index for policy is more than 40 pages if printed. It names many specific policies and is a good index to Bush Administration foreign and domestic policies.

```
     economic policy, first rule of  (1)
     economic policy, implementation of  (1)
     sound economic policies, implementation of  (1)
     economic policy, worst form of  (1)
     economic policy, represents worst form of  (1)
     forget economic policy  (1)
     global economic policy  (1)
     governor clinton's economic policy  (1)
          governor clinton's economic policy, parts of  (1)
     international economic policy  (4)
          international economic policies  (1)
          international economic policy, professor of  (1)
     president's economic policy  (1)
          president's economic policy, member of  (1)
     regional economic policy  (1)
          regional economic policy, office of  (1)
     responsible economic policy  (1)
          responsible economic policies  (1)
          responsible economic policy, part of  (1)
     sound economic policy  (2)
          be sound economic policy  (1)
          enhance sound economic policy  (1)
          sound economic policies  (9)
          sound economic policy, framework of  (1)
          sound economic policy, key aspect of  (1)
          sound economic policy, underpinning of  (1)
     tying economic policy  (1)
...
foreign policy  (219)
     active foreign policy  (2)
     administration's foreign policy  (2)
     ambitious foreign aid policy  (1)
     america's foreign policy  (1)
          american foreign policy  (6)
     articulate foreign policy  (1)
     base foreign policy  (1)
     bipartisan foreign policy  (5)
     build foreign policy  (1)
     conduct foreign policy  (5)
     dictate foreign policy  (1)
     different foreign policy  (1)
     divide foreign policy  (1)
     effective foreign policy  (1)
     enjoy foreign policy  (1)
     enlightened foreign policy  (1)
     establish foreign policy  (1)
     execute foreign policy  (3)
     expressing foreign policy  (1)
     failed foreign policy  (1)
     farsighted foreign policy  (1)
     fine foreign policy  (1)
     foreign policies  (7)
     foreign policy of country  (8)
     foreign policy of great power  (1)
     foreign policy of united  (5)
     foreign policy, complex realms of  (1)
     foreign policy, conduct of  (1)
```

foreign policy, congressional micromanagement of (1)
foreign policy, kind of (1)
foreign policy, master stroke of (1)
foreign policy, review of (1)
foreign policy, serious students of (1)
foreign policy, specific areas of (1)
foreign policy, subject of (1)
foreign policy, tool of (1)
foreign policy, whole area of (1)
formulate foreign policy (1)
good foreign policy (2)
governor clinton's foreign policy (1)
have foreign policy (1)
ignores foreign policy (1)
imaginative foreign policy (1)
implemented foreign policy (1)
important foreign policy (1)
it's bad foreign policy (1)
japan's foreign policy (1)
kept foreign policy (1)
legislate foreign policy (1)
managing foreign policy (1)
mastermind foreign policy (1)
micromanaging foreign policy (1)
nation's foreign policy (4)
one's foreign policy (1)
overall foreign policy (1)
pretends foreign policy (1)
prudent foreign policy (1)
pursuing foreign policy (1)
run foreign policy (2)
shape foreign policy (1)
shift foreign policy (1)
soviet foreign policy (4)
states foreign policy (4)
strong foreign policy (2)
that's foreign policy (2)
think foreign policy (1)
today's world foreign policy (1)

…

trade policy (9)
administration's trade policy (1)
administration's trade policy, goal of (1)
aggressive progrowth trade policy (1)
have aggressive progrowth trade policy (1)
agricultural trade policy (1)
american free trade policy (1)
american free trade policy, result of (1)
fair trade policy (1)
fair trade policies (2)
forward-looking trade policy (1)
proposed forward-looking trade policy (1)
international trade policy (1)
protectionist trade policy (1)
produce protectionist trade policy (1)
strategic trade policy (2)
that's sensible trade policy (1)

```
tough trade policy  (1)
need tough trade policy  (1)
trade policies  (4)
      fair trade policies  (2)
      open trade policies  (1)
      pursue trade policies  (1)
      sensible trade policies  (1)
```

Figure 1. A Portion of the Conceptual Index for the Term Policy.

Sample Collection

The Bush Presidential Public Papers (1989-1993) were selected as the sample collection for evaluation of document retrieval technologies.[9] This is a collection of 5173 documents consisting of statements by the President, addresses, communications to Congress and Federal agencies, interviews with the news media, joint statements, and various other public papers. These digital documents will be used to create a publication to be included in Public Papers of the US Presidents.[10]

Each document is stored as an HTML file with file name of the form YYMMDDNN.htm, where YY is the last two digits of the year of the document, MM and DD the month and day of the month of the document, and NN is a sequential number of the document for that day. The size of the collection of HTML documents is 32 megabytes.

Record paper copies of these documents are in collections of the Bush Presidential Library. They do not need to be reviewed for FOIA exemptions or PRA restrictions because they are already in the public domain. They are not actually representative of the types of documents that might be identified as relevant to a FOIA request, because they are already public. However, they can be searched for topics that occur in FOIA requests. The Bush Public Papers are, however, representative of types of open documents that might contain information sought by a researcher.

FOIA Requests and Queries

The Bush Presidential Library began to respond to FOIA requests in 1998. Processed Freedom of Information Act Requests are listed as a finding aid at the Bush Presidential Library web site.[11] Researchers also request documents that have been opened to the public using a Reference Request Form. They can formulate their request using the libraries finding aids. Fifty queries constructed from the FOIA Requests are listed in Appendix A.

[9] http://bushlibrary.tamu.edu/papers/
[10] http://www.gpo.gov/nara/pubpaps/srchpaps.html
[11] http://bushlibrary.tamu.edu/papers/

The information that a researcher needs to find may not match the subjects, names, or folder titles in the finding aids. A reference librarian is often a useful intermediary in finding relevant documents because of their knowledge of the Presidential administration and the holdings of the library that are not necessarily reflected in the subjects, names and titles of the finding aids.

WebGlimpse has already been used by the Bush Presidential Library to index the Bush Public Papers, and was used through an interface at the Bush Presidential Library web site. A researcher learned to use the system and performed retrieval experiment over a week or so. Queries were constructed from the FOIA requests in Appendix A for use with WebGlimpse. They were refined to include additional relevant terms discovered in retrieved documents. The resultant WebGlimpse queries are shown in Appendix D.

The Bush Public Papers were downloaded from that Bush Presidential Library web site to a server at Georgia Tech and indexed using Oracle Text. A researcher learned to use the system and performed the retrieval experiment over several weeks. Queries were refined to include additional relevant terms discovered in retrieved documents. The queries were also refined to use some of the best features of Oracle Text with word queries. Queries formulated from the FOIA requests and used with Oracle Word Search are shown in Appendix C.

The experiment with NOVA was performed in a laboratory at Sun Microsystems Research Laboratory. A staff member of the laboratory used NOVA to create a conceptual index for the Bush Public Papers. The experiment was performed in about 8 hours over a two-day period in which a researcher from Georgia Tech learned to use the system, formulated queries, and collected results. There was inadequate time to use relevancy feedback to improve the NOVA queries. Queries formulated from the FOIA requests and used with NOVA are shown in Appendix B.

The NOVA interface shown in Figure 2 was especially configured for this experiment.

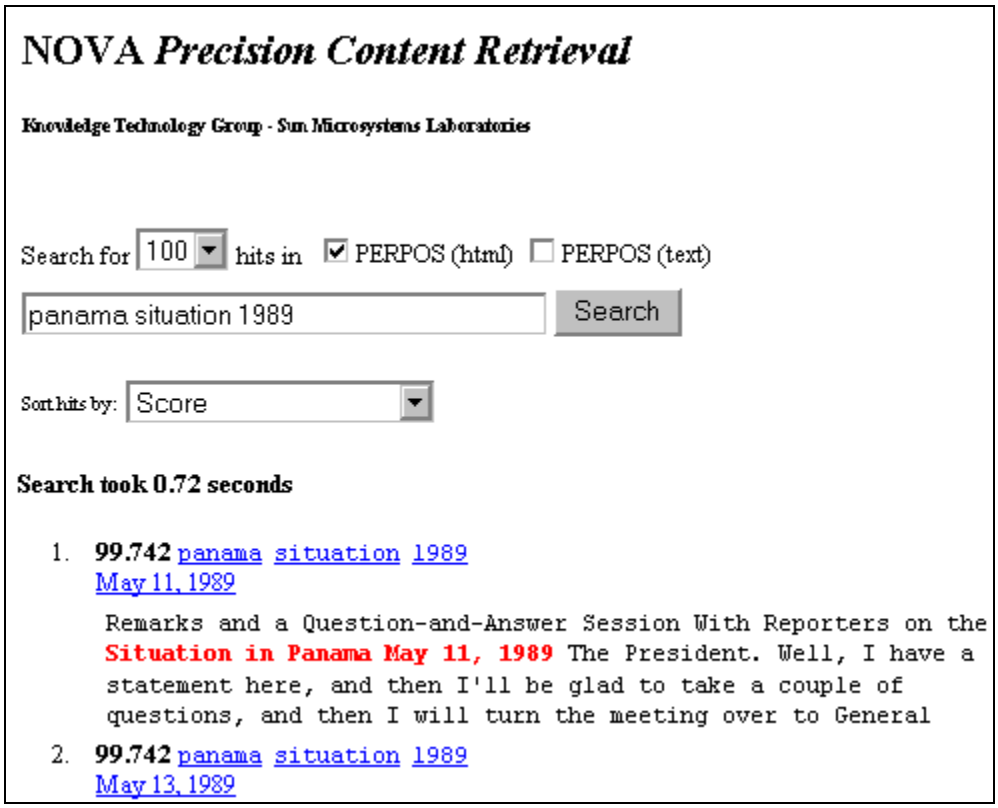


Figure 2. NOVA Precision Content Retrieval Interface.

The researcher had to select whether a maximum of 20, 50 or 100 passages would be retrieved for each query. Since passages were retrieved, not documents, there were often fewer documents represented in the results than the maximum number of passages to be retrieved. Even when a maximum of 100 passages was selected, it was clear that other relevant passages, and possibly additional documents, would have been retrieved had one been able to select a larger maximum number of passages. Consequently, the results with regard to retrieved relevant documents do not accurately represent the performance of the system with regard to the number of relevant passages (documents) that it might have retrieved, but not displayed.

On the other hand, one might select a maximum of 100 documents to be retrieved for a query, have a query of n terms, m of which had to occur. The NOVA interface did not provide the capability to indicate which terms or how many terms had to occur. However, for each term in a query that is missing, the penalty score drops by 10. We established a cutoff value determined by how many terms had to appear in the passage. For instance, with a query with 5 terms, if three had to occur, the cutoff would be 70.

The NOVA Precision Content Retrieval technology is designed to perform best for short queries of just a few words [Woods, 1998]. Since the system treats inflected and derived forms of words as if they were more specific than the base form of the word, queries using the base form are more effective than the inflected or derived forms. The researcher

formulating the queries from the FOIA requests failed to follow this guideline and often formulated queries with two or more phrases in them. Since the system limited the number of passages that were retrieved, this produced worse results than would have been achieved by submitting multiple queries made up of single noun phrases and then combining the results. This failure to formulate the query properly became more evident during the analysis of the results for each query, which is discussed later in this paper.

Methodology

The number of items in the collection relevant to a query was determined first by pooling the document identifiers retrieved using the three retrieval systems. Since NOVA retrieves passages rather than documents, just the unique document identifiers identified by NOVA were included in the pool. Secondly, a person judged for each document whether it was relevant to the FOIA request. While two of the retrieval systems ranked the documents by relevance to the query, the person judging the relevance of the document to the FOIA request made no attempt to establish a ranking of documents judged to be relevant.

Figure 3 shows the number of relevant documents per topic.

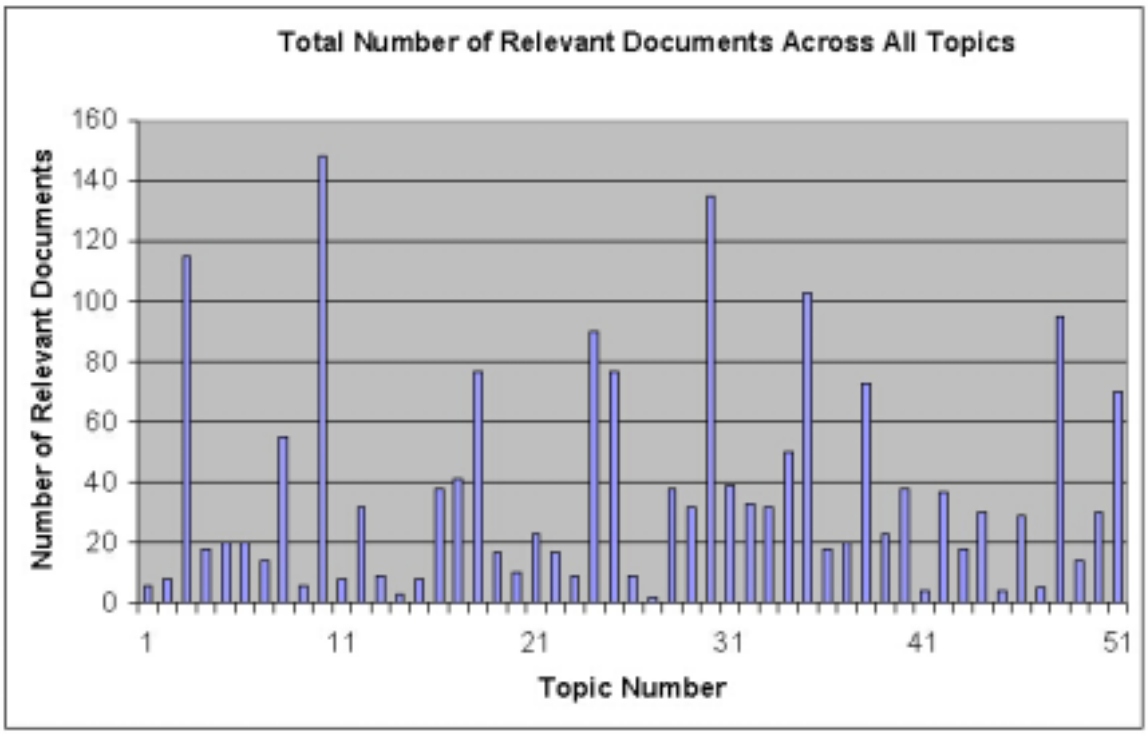


Figure 3. Number of relevant documents per topic.

Recall is a measure of the ability of a system to present *all* relevant documents.

$$\text{Recall for a topic} = \frac{\text{number items retrieved that are relevant to a topic}}{\text{number of items in a collection relevant to a topic}}$$

It is easy to achieve perfect recall. Just return every document in the collection for every query. Hence, recall alone is not a good measure of the quality of a document retrieval system.

Precision is a measure of the ability of a system to present *only* relevant documents.

$$\text{Precision for a topic} = \frac{\text{number of items retrieved that are relevant to a topic}}{\text{total number of items retrieved for a topic}}$$

For response to FOIA requests, a document retrieval system must have high recall. To reduce the number of documents that have to be reviewed the retrieval system should have high precision, without sacrificing recall.

Average precision is a good measure of the utility of a document retrieval system. Average precision combines precision, relevance ranking and overall recall. Average precision is the sum of the precision at each relevant hit in the hit list divided by the total number of relevant documents in the collection.

Evaluation of the results of the experiments was performed using the metrics used in the Text Retrieval Conference [Voorhees and Harman, 2001] and the `trec_eval` program.[12]

Results

To facilitate computing average performance over a set of topics, each with a different number of relevant documents, individual topic precision values are interpolated to a set of standard recall levels (0 to 1 in increments of .1). The interpolated precision at standard recall level *i* is the maximum precision obtained for the topic for any actual recall level greater than or equal to *i*. These values are plotted in Recall-Precision graphs. The Interpolated recall-precision graphs over each topic for the three retrieval systems are shown in Appendix E. Figure 4 shows the interpolated recall-precision averages over all topics for the three retrieval systems.

[12] Trec_eval was written by Chris Buckley and can be obtained by anonymous ftp at ftp.cs.cornell.edu in the directory pub/smart.

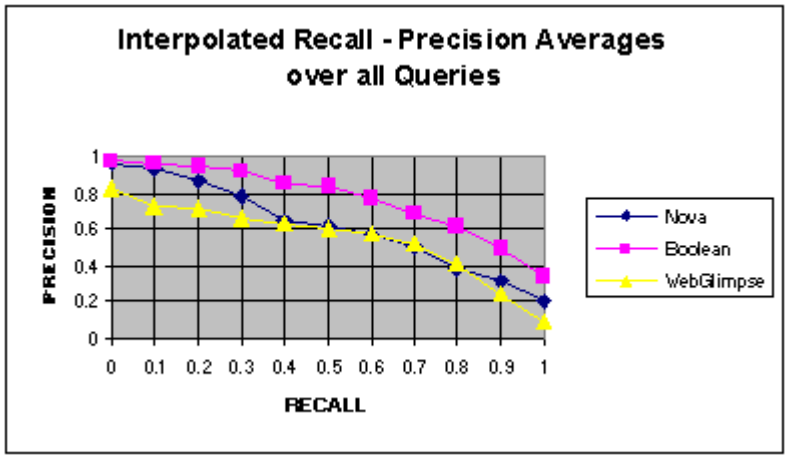
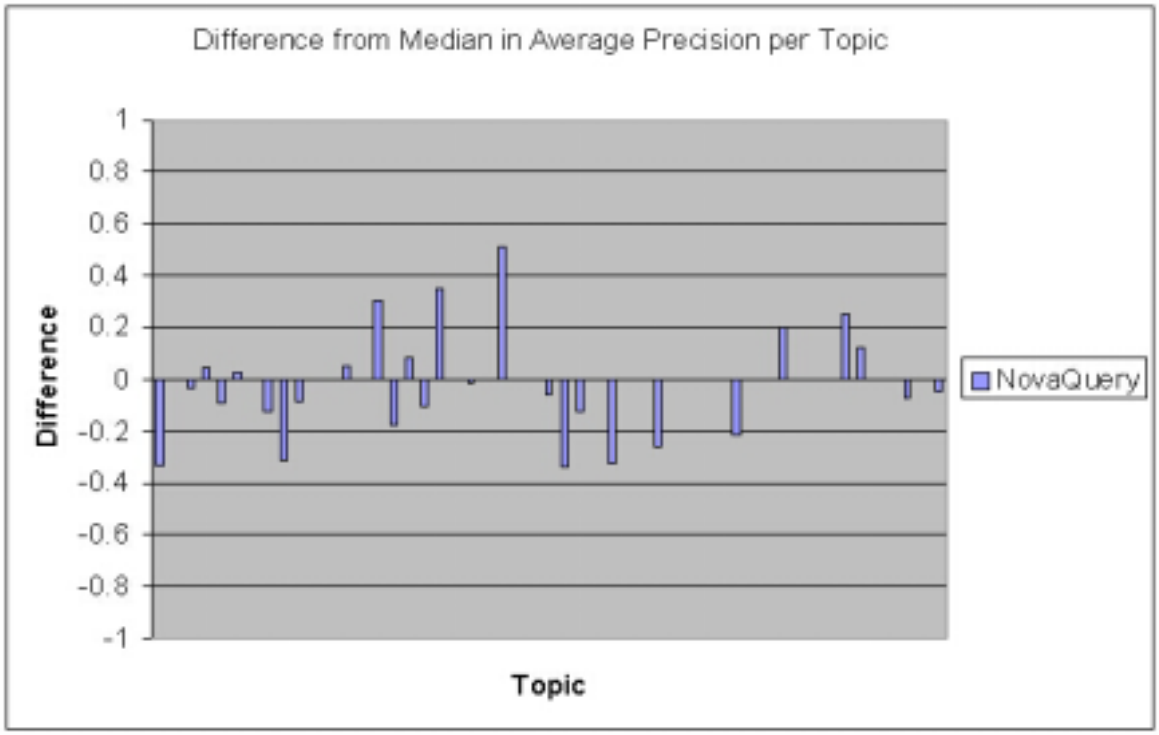


Figure 4. Interpolated Recall-Precision Averages

The area under each of the curves corresponds to the average precision for that retrieval system. The average precision of Oracle Text with word queries was .7620, NOVA .6165, and WebGlimpse .5436.

The average precision histogram measures the average precision of a run on each topic against the median precision of all corresponding runs on that topic. This graph provides insight into the performance of individual systems and the types of topics that they handle well. Figure 5 shows the difference from the median in average precision per topic for each of the three systems.



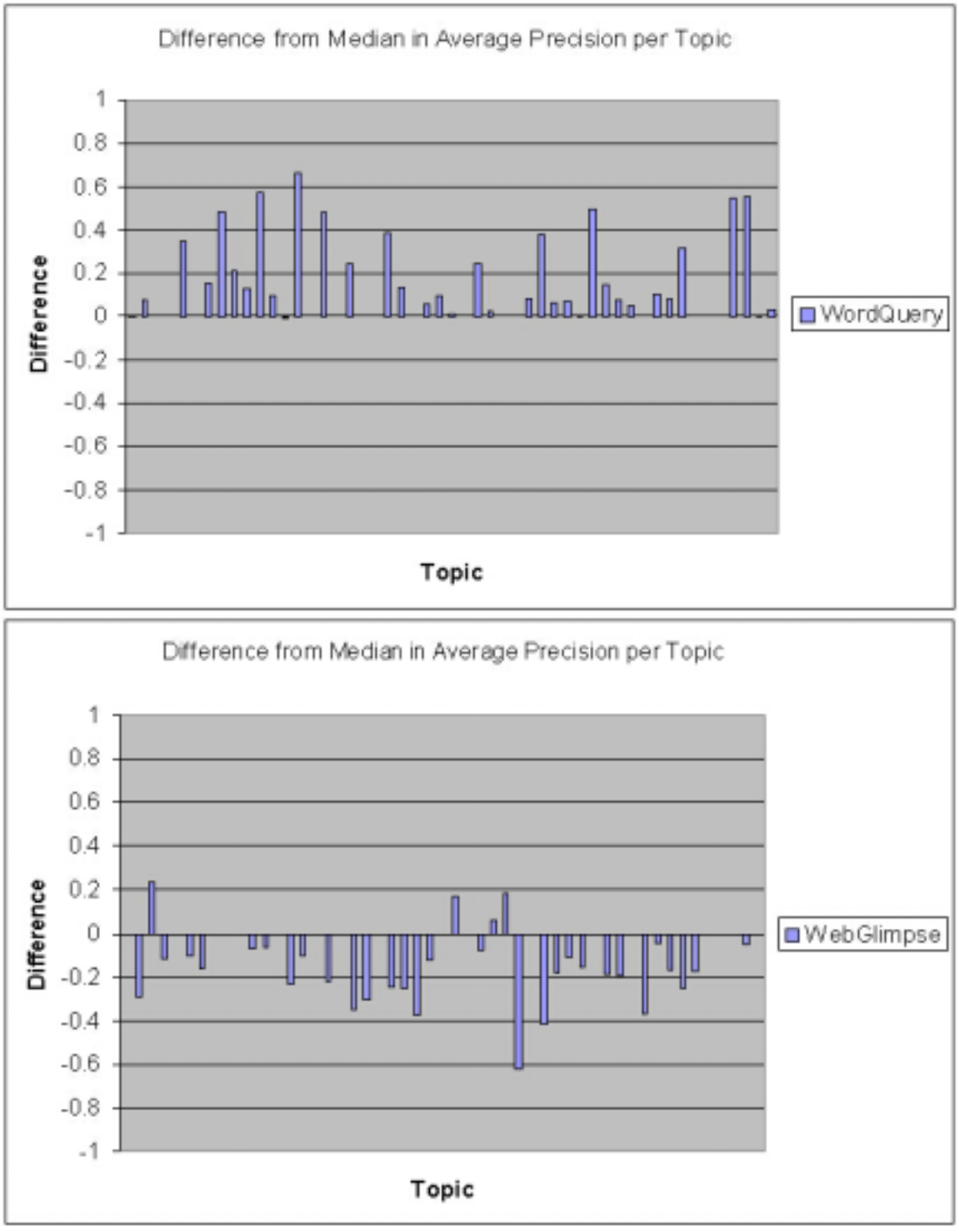


Figure 5. Average Precision Histograms.

All three systems exhibited the same performance on FOIA request 47, "Records relating to SUPER 301 / Omnibus Trade and Competitiveness Act of 1988" The query to NOVA and WebGlimpse was "Super 301". The query to Oracle Text with word queries was "(super 301) OR (omnibus trade OR (competitiveness act AND 88=1988))." Adding the alternatives did not increase the recall.

Analysis of WebGlimpse's Performance

WebGlimpse has a low computational and storage overhead. It indexes only single words and the storage requirements for the index are typically 5% of the size of the collection indexed.

WebGlimpse's overall performance was below that of the Oracle and NOVA systems. However, its average precision was better on queries 3, 27, 30, 31 than that of the other systems. In the cases of queries 3 and 30, NOVA would probably have outperformed WebGlimpse had enough passages been retrieved (see analysis of NOVA's performance below). In the cases of queries 27 and 31, WebGlimpse had better queries than those provided Oracle and NOVA.

FOIA request 31 was 'Records related to 1992 UNCED ("Earth Summit") in Rio'. WebGlimpse's query was "UNCED,{Earth Summit},{Rio;{Y92}}" Oracle Word's query was "({UNCED}|earth summit) OR (Rio NEAR 92=1992). NOVA's query was "UNCED Earth Summit Rio 1992." The WebGlimpse Query required that if the terms Earth and Summit occurred they had to occur together. The NOVA query did not require this and retrieved documents about summits that were not the Earth Summit. Similarly, 1992 could occur in a passage retrieved by a NOVA query and not include Rio. The Oracle Text query required that Rio be near 1992, and did not retrieve all relevant documents.

Analysis of Oracle Text with Word Queries Performance

Oracle Text's average precision on all queries was significantly better than that of the other two systems. Its average precision per topic was greatest on 34 of the 51 topics.

FOIA request 5 was "documents which concern POW/MIAa in Southeast Asia, including Vietnam and Laos." The Oracle Word query for this request was "(southeast asia | vietnam |laos) AND (({POW}|prisoner of war) OR ({MIA}|missing in action))." The capability to use logical OR combined with logical AND gave Oracle an advantage over NOVA, and Oracle Text's statistical ranking of results gave it an advantage over WebGlimpse.

FOIA request 11 was "Records relating to US military intervention in Somalia (i.e., Operation Restore Hope)." The Oracle Word Query was "(Somalia NEAR military) OR (operation restore hope)." The capability to express in the query that the entire phrase "operation restore hope" had to occur gave Oracle an advantage over NOVA, which also retrieved documents for "operation", "restore hope", "restore" and "hope."

FOIA request 21 was " Documents related to the creation of a solution to the Savings and Loan crisis between the date of the Bush Administration coming into office until August 9, 1989 (the latter date being the passage of the Financial Institutions Reform, Recovery and Enforcement Act)." The Oracle Word query was "($saving {and} loan) and DOCNAME < 89090900." The capability to specify that documents had to be created

prior to a certain date gave Oracle an advantage over NOVA queries, which did not have that capability. A similar situation occurred for other FOIA requests, e.g., requests 33, 37 and 38.

FOIA request 48 was "Records relating to Japan - Trade and Economic Policy." The Oracle Word Query was "($Japan=Japanese NEAR $trade = $economic)." The NOVA query was "Japan Trade and Economic Policy." Oracle had the advantage because the terms Japan or Japanese had to occur as well as the terms trade or economic. NOVA found many documents relating to trade policy and economic policies that were not related to Japan, because NOVA provides no mechanism to ensure that the concept Japan occurs.

FOIA request 49 was:

> Records relating to Bilaterals (Ministerial Meetings)
> 3/2-3/90 Bush - Kaifu Ministerial Meetings (Palm Springs, CA)
> 4/4/91 Bush - Kaifu Ministerial Meetings (Newport Beach, CA)
> 7/11/91 Bush - Kaifu Ministerial Meeting (Kennebunkport, Maine)
> 9/1/89 Bush - Kaifu Ministerial Meetings (Washington)

The Oracle Text with word query was"

> '(Kaifu, Bush, $minister, ministerial, $meeting, $discuss, $discussion, (March NEAR 90=1990), (April NEAR 91=1991), (July NEAR 91=1991), (September NEAR 89)) AND (Kaifu)'

The capability to require that the term *Kaifu* occurred in the document gave Oracle Text an advantage over the NOVA query "bush kaifu meeting." The NOVA query retrieved many passages regarding a *Bush meeting* that were not related to meetings with Prime Minister *Kaifu*.

Analysis of NOVA's Performance

NOVA requires a high computational and storage overhead to construct and store the conceptual index. Typically the conceptual index requires as much or more storage than the original collection. The response time for NOVA queries in this experiment ranged from .03 to 3.37 seconds and averaged .36 seconds.

NOVA's average precision was marginally better than WebGlimpse's and significantly less than Oracle Text's. However, NOVA's average precision on 10 of the topics was better than the other two systems. In each of these cases there were few passages/documents relevant to the query and the request was for a very narrow topic.

NOVA's performance on FOIA request 19 exemplifies a good expression of the request and how the conceptual index supports finding relevant documents. The FOIA request

was "A request for materials relating to or reflecting President Bush's concern for ethical conduct in the Federal executive, specifically, the creation, operations, and staffing of the Office of Government Ethics as a separate agency in October 1989." The query to NOVA was "Office of Government Ethics". The conceptual map created for the concepts looks something like this:

Ethics
    Government Ethics
        Government Ethics, Office of

NOVA first finds all documents that have exactly "Office of Government Ethics", and then documents that have "Government Ethics". These are exactly the documents requested.

FOIA request 23 is also a good example of cases in which NOVA will perform well. The FOIA request was: "Documents concerned with nuclear proliferation with respect to India or Pakistan." The Query to NOVA was "India Pakistan nuclear proliferation". NOVA first tries to find passages that have terms exactly in this order, possibly with intervening words. Then it will try to find passages in which one of the terms is missing, working left to right in the query. So it will look for "Pakistan nuclear proliferation", then "India nuclear proliferation."

For many of the FOIA requests in which NOVA's average precision was less than the median average precision, the NOVA queries were incompletely or poorly expressed. For instance, the eighth FOIA request was "Materials pertaining to Human Immunosuppressant Virus or HIV, and Acquired Immune Deficiency Syndrome or AIDS. [Note: HIV is more often the abbreviation of Human Immunodeficiency Virus]." Query 8 to NOVA was "HIV Human immunodeficiency virus AIDS." While NOVA retrieved passages corresponding to 29 of 55 relevant documents in the collection, separate queries should have been issued for "Acquired Immune Deficiency Syndrome" and for "Human Immunosuppressant Virus" and the results combined.

The NOVA query corresponding to FOIA request 9 was "Refugee Asylum Central America". NOVA hasn't the knowledge that *Costa Rica*, *Belize*, *El Salvador*, *Guatemala*, *Nicaragua* and *Honduras* are parts (meronyms) of *Central America*. However, it can have semantic knowledge that *Central American* subsumes *Nicaraguan*, *Guatemalan*, etc. Hence, this query would have been better expressed as "Central American Refugee". Then it might have retrieved passages containing "Nicaraguan Refugee", "Guatemalan Refugee", etc.

For FOIA request 17, the NOVA query was "nomination Clarence Thomas Supreme Court Justice." Because nominee is not a kind of nomination, and nominee occurred in passages when nomination didn't, a better query would have included both terms.

For FOIA request 21, the documents had to be prior to August 9, 1989 in order to be relevant, but NOVA had no way to include this restriction. If it did, its precision would

have been much better. Also, NOVA would have performed better, if it had a rule that the phrase "Savings and Loan" subsumed its abbreviations "S&L" and "S and L".

FOIA request 24 was for "Information on U.S. efforts to convene an Arab-Israeli peace conference in 1991. The conference ultimately occurred in Madrid, Spain, and opened on October 10, 1991." "I am requesting information on the efforts to convene the conference and convince all relevant parties to attend. I believe this primarily took place between February and October, 1991 (Though U.S. discussions might have begun as early as late 1990)." NOVA did not relate the terms *Israel* and *Israeli*, because *Israeli* is not a kind of *Israel*. To find one or the other, both would have to be included in a query. Furthermore, most of the relevant documents were missing the terms *Arab* and *Israeli*. Instead they had the terms *Mideast* or *Middle East*.

FOIA request 33 specifies that documents after 12/20/89 are not required. The NOVA interface did not allow us to specify that document date should be less than or equal to that date, so NOVA retrieved documents relevant to the query after that date, but not relevant to the FOIA request.

FOIA request 37 was for "All materials pertaining to breakdown of the Soviet Union 12/91". The NOVA query was: "revolution Soviet Union December 1991". The results would have been better if the query had included the terms *breakdown* or *collapse* and *Russia*.

FOIA request 39 was "Documents and materials held by the Bush Presidential Library that pertain to Human Rights as a part of foreign developmental aid, particularly for the years post-1989." "Assistance" would have been a better query term than "aid". In a few cases, NOVA found AIDS (acquired immune deficiency syndrome), because it does not distinguish between upper and lower case letters. The term "aid" subsumes "aids", but should not subsume "AIDS".

The NOVA query for FOIA request 49 was "Kaifu Bush meeting." The query response would be improved if one could ensure that a particular term such as Kaifu occurred. The query results would have been better had the query simply be "Kaifu." Some documents that described meetings between Kaifu and Bush were not retrieved, e.g., March 2-4, 1990 meeting in Palm Springs. The reason they were missed is that the term "President" was used in the document, not the term "Bush". The results would have been better had the term "President Bush" been in the query. Specific dates were mentioned in the FOIA request. If they had been included in the query, the performance would not have improved because they would have to occur in the passage, not in the file name. Many of the relevant papers did not include the term *meeting*, but did include the term *discussion*.

The user interface to NOVA required one to select 20, 50, or 100 relevant hits (passages). There could be fewer passages retrieved than the selected number, for instance, Query 12, *Bosnia*, was set to return 100 hits, but there were only 87 passages containing the term *Bosnia*, and 29 unique documents containing the term. However, if there were more documents relevant to a query that the number of hits selected, the number retrieved were

stopped at the number selected. For instance, on query 10, "Operation Desert Shield Storm Persian Gulf War IRAQ Kuwait", *Search for 100 hits* was selected and 100 passages were retrieved. 66 passages were relevant from 59 documents. However, there were 150 documents judged as relevant to this query. If it had been possible to set the cutoff to a higher number, say 500 hits, the average precision (and recall) on this topic would probably have been higher. As shown in the Figure 6, there were at least ten NOVA queries in which the recall and average precision would have been significantly higher had the retrieved passages not been cutoff too soon.

Query	Search for *n* Passages	Relevant Passages Retrieved	Documents Corresponding to Passages	Relevant Documents in Collection
3	50	37	32	115
10	100	66	59	150
16	20	8	8	40
18	100	30	22	77
28	100	92	26	38
30	100	75	35	135
35	100	63	63	102
38	100	98	47	75
44	50	25	20	30
48	100	50	33	95

Figure 6. NOVA Queries in which the Limit on Passages Retrieved Affected Average Precision.

Comparison with the Results of the TREC-8 Ad Hoc Retrieval Task

Document retrieval technologies have been extensively evaluated in the Ad Hoc Query Track of the annual Text Retrieval Conferences (TREC) [Voorhees and Harman, 1999]. In TREC-8, the Ad Hoc Query Task collections (document set) consisted of 1.904 Gigabytes of documents. There were 528,155 documents from the *Financial Times* (1991-1994), the *Federal Register* (1994), the Foreign Broadcast Information Service and the *LA Times*. The documents in the collections were tagged using SGML to allow easy parsing. Depending on document type, e.g., newspaper article, Federal Register Notice, the documents have different tags.

Fifty natural language topic statements were used in the Ad Hoc Query Task. Figure 7 shows a sample TREC-8 topic.

<num> Number: 409
<title> legal, Pan Am, 103
<desc> Description:
What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988
<Narr> Narrative:
Documents describing any charges, claims, for fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.

Figure 7. A sample TREC-8 Topic.

Figure 8 shows the Recall/Precision graph for the top five ad hoc runs. Each of these experiments used relevancy feedback to refine the queries.

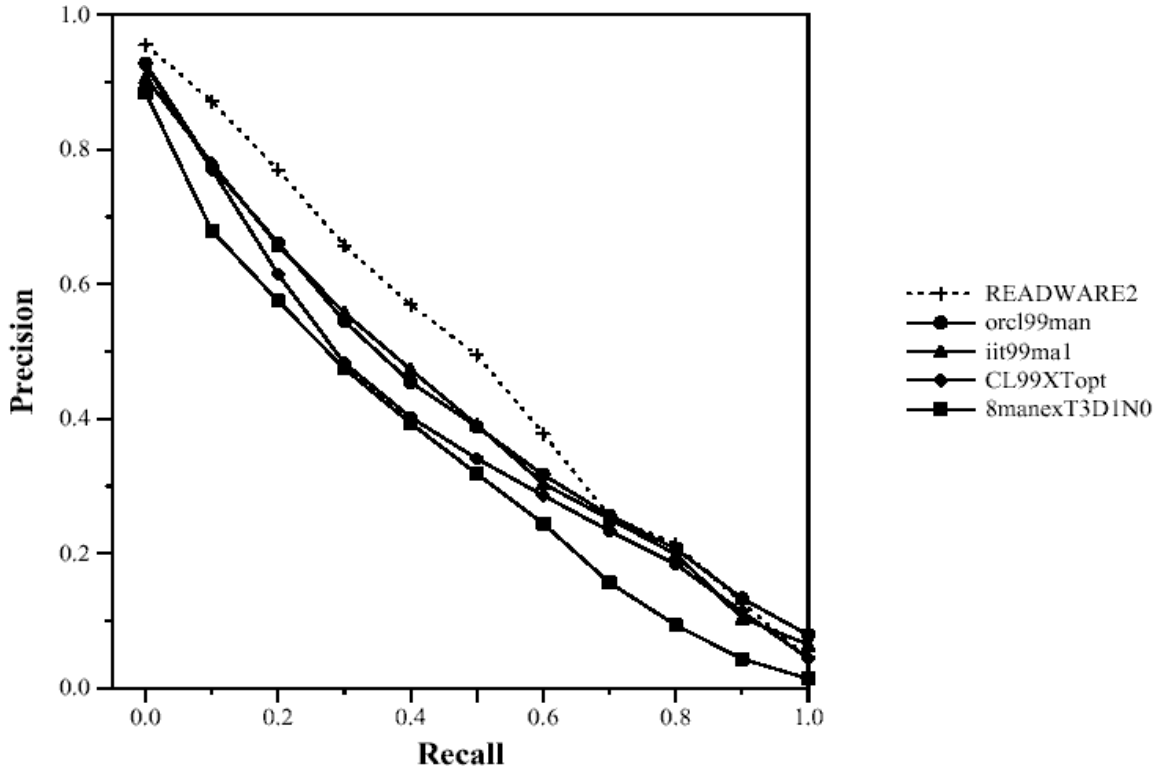


Figure 8. Recall/Precision graph for the top five ad hoc runs in TREC-8.

READWARE2 -- Management Information Technologies, Inc. READWARE is a knowledge-based natural language document retrieval system. Multiple queries were used for each topic and the submitted results were the union of the output of the different queries for each topic. READWARE had an average precision of about .47 and had the best average precision on 23 of the 50 topics.

orcl99man - Oracle Text with About Queries uses the Oracle Text Boolean and statistical operators extended to use query expansion using knowledge bases or thesauri. It had an average precision of .413 and had the best average precision on 5 topics.

Iit99ma1 -- Illinois Institute of Technology. The basic retrieval strategy is a vector-space model. The best run had an average precision of .4104.

CL99Xtopt -- CLARITECH. CLARIT is a document retrieval system that incorporates natural language processing, thesauri, and the vector space retrieval model. It had an average precision of .3765

8manexT3D1N0 -- GE Research and Development Group used a Natural Language-based retrieval system. It had average precision of .3346.

The average precision of the systems evaluated in this paper is significantly greater than those evaluated in the TREC-8 Ad hoc Track. This does not imply that the document and passage retrieval technologies evaluated in this paper have better performance than the five best technologies of the TREC-8 evaluation. Precision is dependent on the size of the document set. This is why the organizers of the Text Retrieval Conferences have constructed large document sets for their evaluations. It is to be expected that if the experiments described in this paper were conducted for larger document sets, their average precision would be lower.

The Ad Hoc Query Track was discontinued after TREC 8 because the organizers concluded that the results in the track had leveled off. That is not to say that the ad hoc text retrieval problem was solved, but that there did not seem to be continued improvements in the technologies represented in the track.

Conclusions

There are significant differences in the performance of the three document retrieval technologies evaluated. The average precision of Oracle Text with word queries was .7620, NOVA .6165 and WebGlimpse .5436. However, the average precision of NOVA would have been significantly higher if the NOVA interface had allowed a higher number of passages to be retrieved and there had been time to refine the queries based on relevancy feedback. Furthermore, the average precision measures would have been lower if the collection searched had been larger.

To determine whether the performance of these technologies is scalable, a similar experiment should be conducted with a much larger corpus of Presidential electronic records. A more advanced document retrieval technology should be substituted for the Boolean document retrieval technology. To demonstrate the capability to handle not only text or HTML files but also PC files, the corpus should contain a variety of file types, e.g., WordPerfect, Lotus 123, Harvard Graphics. Data should also be collected on improvements in performance due to the use of relevancy feedback to refine queries.

Based on our analysis of the results of the NOVA queries, the performance of NOVA can possibly be improved by constructing an interface to NOVA that enables a user to specify:

- Whether to retrieve passages or documents.
- A larger number of passages (or documents) to be retrieved.
- Terms that must be included in a passage.
- That the passage may include one phrase OR another.
- Cutoff of passages retrieved based on penalty score.

If a user of NOVA could specify additional lexical knowledge, it should also improve the performance. For instance, a phrase such as "Strategic Arms Reduction Talks" should be equivalent to its acronym START. A phrase such as "United States" should be equivalent to its abbreviations, U.S. and US. It would also help if the parser used for creating the index and the query parser could distinguish abbreviations and acronyms. Then it might be possible to distinguish a capitalized acronym or abbreviation from the corresponding term, e.g., verb or noun "start" and the acronym "START."

The conceptual index that NOVA creates for a collection resembles a back-of-the-book index. It provides users an alternative way of finding documents relevant to their needs. It also provides a method for query refinement. Experiments should be conducted to determine whether responses to FOIA requests might be satisfied by navigating this index, either in place of queries, or as a supplement to NOVA queries.

References

Ambroziak, J. and Woods, W. A. (1998) Natural Language Technology and Precision Content Retrieval. Technical Report SMLI TR-98-69, Sun Microsystems Laboratory, December 1998.

Manber, U., Gopal, B. and Wu, S. (1998) Glimpse Documentation. Department of Computer Science, University of Arizona. http://webglimpse.org/gdocs/glimpsehelp.html

Manber, U. and Wu, S. (1993). GLIMPSE: A Tool to Search Through Entire File Systems. Department of Computer Science, University of Arizona. October 1993.

McGregor, C. (1999). Oracle8i *inter*Media Text Reference, Release 2 (8.1.6). Part No. A77063-01. Oracle Corporation. December 1999. http://download-east.oracle.com/docs/cd/A81042_01/DOC/index.htm

McGregor, C. (2002). Oracle Text Reference, Release 9.2. Part No. A96518-01. Oracle Corporation. March 2002. http://technet.oracle.com/docs/products/oracle9i/doc_library/release2/text.920/a96518/toc.htm

Underwood, W. E., Hayslett-Keck, M. and Laib, S. (2002) The Archival Processing Tool (APT): User's Guide, Version 2.04, PERPOS Technical Report ITTL/CISTD 02-2, November.

Voorhees, E. M. and Harman, D. Overview of the Eighth Text Retrieval Conference (TREC-8) (1999) The Eighth Text Retrieval Conference (TREC-8), NIST Special Publication 500-246. http://trec.nist.gov/pubs.html

Voorhees, E. M. and Harman, D. K. (eds.) (2000) The Ninth Text REtrieval Conference (TREC-9), NIST Special Publication 500-249.

Woods, W. A. (1998) Guidelines for Precision Content Retrieval. Sun Microsystems Laboratory Technical Report.

Appendix A: FOIA Requests Used in Experiments

1. Documents pertaining to United States Information Agency (USIA), Voice of America (VOA) and China. (98-0001-F)
2. Documents pertaining to Radio Marti or TV Marti and broadcasting to Cuba. (98-0001-F)
3. Documents pertaining to aid to nonpublic schools, e.g., tuition tax credits and vouchers during the George Bush Presidency. (98-0002-F)
4. Documents pertaining to the Domestic Policy Council. (98-0004-F)
5. I am requesting a copy of all documents which concern POW/MIAs in Southeast Asia, including Vietnam and Laos. I am a family member of an MIA.(98-0029-F)
6. Records, briefings and reports held by the Bush Library which pertain to the bombing of Pan Am Flight 103 over Lockerbie, Scotland on December 21, 1988 . (98-0034-F)
7. Materials pertaining to the nomination of former Senator John Tower of Texas as Secretary of Defense. I am interested particularly in Tower's nomination fight for defense secretary. (98-0041-F)
8. Materials pertaining to Human Immunosuppressant Virus or HIV, and Acquired Immune Deficiency Syndrome or AIDS. [HIV is actually Human Immunodeficiency Virus] (98-0091-F)
9. Files dealing with the following subject matter: refugee / asylum procedures for Central America (e.g., El Salvador, Nicaragua, and Guatemala). (98-0097-F)
10. Records relating to US military intervention against Iraq (i.e., Desert Shield, Desert Storm) (98-0099-F)
11. Records relating to US military intervention in Somalia (i.e., Operation Restore Hope). (98-0101-F)
12. Records pertaining to US military intervention (or non-intervention) in Bosnia. (98-0102-F)
13. All documents and materials that relate to the NATO Heads of State Summit at Rome on Nov 7-8, 1991. This is also referred to as a meeting of the North Atlantic Council at Rome on Nov 7-8, 1991. This includes, but is not limited to speeches and formal statements regarding the Rome Summit in general, and specifically, the NATO Strategic Concept presented at that summit. (98-0142-F)
14. Documents or material dealing with March 14, 1991: Meeting between President Bush and French President Francois Mitterand. (98-0142-F)
15. Records and references relating to George Bush and Iran-contra independent counsel Lawrence Walsh. (98-0189-F)
16. All records related to President Bush trip to Hungary during July 11-14, 1989 and his diplomatic efforts in Hungary. (98-0194-F)
17. Papers pertaining to the nomination of US Supreme Court Justice Clarence Thomas. (98-205-F)
18. A request for materials relating to the making, revising, adopting of the FY 1990 Federal Budget or what White House aids called the "budget deal of the century. (98-0251-F)
19. A request for materials relating to or reflecting President Bush's concern for ethical conduct in the Federal executive, specifically, the creation, operations, and staffing of the Office of Government Ethics as a separate agency in October 1989, 1989. (98-0251-F)
20. I am interested in the various foreign and domestic policy experts that were assembled to assist President Bush assume power. I would like to request access to any materials which link President Bush to the following think tank: Council on Foreign Relations. (98-0255-F).
21. Documents related to the creation of a solution to the Savings and Loan crisis between the date of the Bush Administration coming into office until August 9, 1989 (the latter date being the passage of the Financial Institutions Reform, Recovery and Enforcement Act. (98-0356-F)
22. Documents concerned with US foreign relations with Pakistan. (98-0387-F)
23. Documents concerned with nuclear proliferation with respect to India or Pakistan. (98-0387-F)
24. Information on U.S. efforts to convene an Arab-Israeli peace conference in 1991. The conference ultimately occurred in Madrid, Spain, and opened on October 10, 1991. The conference was orchestrated by the United States and included Palestinian representatives as well as officials from Egypt, Israel, Jordan, Saudi Arabia, Syria and a number of other states.

    I am requesting information on the efforts to convene the conference and convince all relevant parties to attend. I believe this primarily took place between February and October, 1991 (Though U.S. discussions might have begun as early as late 1990).

    One final note: according to Secretary Baker the decision to hold the conference in Madrid only came in the last few weeks before the conference was held. Thus it is unlikely that most file headings or titles will refer to Madrid. (98-0497-F)
25. In general the topic is NAFTA (North American Free Trade Agreement)
    --Any business involvement in the passage of NAFTA - labor involvement.
    --During Bush's Presidency the evolution and history of how and why NAFTA came about.
    --Other peoples, advisors involved.

In general, I am looking for both the internal process of how NAFTA passed and also external support/non- support of it. Mexico's point man was Jaime Serra de Puche. I can't remember the US's point man - it may be Robert Mossbach's [Mosbacher]. (99-0062-F)

26. Request for records re: NASA, specifically. U.S.-Russia joint efforts including but not limited to Shuttle-MIR. (99-0093-F)
27. I am working with former Defense Secretary Caspar Weinberger on his autobiography, and we are interested in any Bush presidential records relating to Secretary Weinberger - in general, but particularly regarding the pardon, which President Bush granted to Weinberger in December of 1992. (99-0099-F)
28. I would like to file a Freedom of Information Act request to gain access to all documents that focused on the internal situation in South Africa between 1948 and 1994. I am interested in studies that examine the internal political dynamics of South Africa and attempt to predict the actions of the white government and the black opposition as they struggled over apartheid. (99-0103-F)
29. Materials relating to the Bush Administration's role in supporting NIH research (including material on the Human Genome Project, and AIDS Research). (99-105-F)
30. Material relating to the Bush Administration's consideration of legislation to expand access to healthcare. (99-0105-F)
31. Records related to 1992 UNCED ("Earth Summit") in Rio. [UNCED is the acronym for UN Conference on Environment and Development] (99-0128-F)
32. Records related to the White House council on competitiveness (V. P. Quayle). (99-0129-F)
33. Documents regarding the decision to invade Panama on 12/20/89. I will not require any documents dated after 12/20/89. I would like any documents relevant to the decisional process. (99-0186-F)
34. All documents that relate in whole or in part to President Bush and General Secretary Gorbachev's meeting in Malta on December 2-3, 1989. (99-0273-F)
35. Documents related to U.S. - Soviet relations, particularly, arms control negotiations - START I and START II (99-0302-F)
36. All materials pertaining to attempted coup against Gorbachev, 8/91 (99-0303-F)
37. All materials pertaining to breakdown of the Soviet Union, 12/91 (99-303-F)
38. Documents and materials that pertain to early policy decisions on the administration's human rights approach in foreign policy. [Early is interpreted as 1989] (99-0461-F)
39. Documents and materials held by the Bush Presidential Library that pertain to Human Rights as a part of foreign developmental aid, particularly for the years post-1989. (99-0461-F)
40. Documents and materials that pertain to Human rights in the bilateral relations between the U.S. and China. (99-0461-F)
41. Documents and materials that pertain to Human rights as an issue at the Drug Summit San Antonio in February 1992. (99-0461-F)
42. Documents and materials that pertain to trade issues in the bilateral relations between the U.S. and China. (99-0461-F)
43. Documents and materials pertaining to Foreign Financial and Military Assistance issues in the bilateral relations between the U.S. and Turkey. (99-0461-F)
44. Documents and materials pertaining to Arms Exports issues in the bilateral relations between the U.S. and Iraq. (99-0461-F)
45. Copies of all documents that relate to the Central Valley Project Improvement Act of 1992 (Public Law Number 102-575). This law was introduced in 1991 as H.R. 429. (99-0551-F)
46. Records relating to SII (Structural Impediments Initiative) (99-0584-F)
47. Records relating to SUPER 301 / Omnibus Trade and Competitiveness Act of 1988 (99-0584-F)
48. Records relating to Japan (Trade and Economic Policy) (99-0584-F)
49. Records relating to Bilaterals (Ministerial Meetings)
    3/2-3/90 Bush - Kaifu Ministerial Meetings (Palm Springs, CA)
    4/4/91 Bush - Kaifu Misisterial Meetings (Newport Beach, CA)
    7/11/91 Bush - Kaifu Ministerial Meeting (Kennebunkport, Maine)
    9/1/89 Bush - Kaifu Ministerial Meetings (Washington)
    (99-0584-F)
50. Request for records pertaining to Hurricane Andrew (99-0727-F2)
51. Material on "Millie", Mrs. Bush's deceased canine. (200-0590-F)

Appendix B: Nova Precision Content Retrieval Queries

NovaQuery-01 = Voice of America VOA USIA China
NovaQuery-02 = Radio TV Marti broadcast Cuba
NovaQuery-03 = Tuition credit voucher public education
NovaQuery-04 = Domestic Policy Council
NovaQuery-05 = POW MIA Pow's MIA's Southeast Asia Laos Vietnam
NovaQuery-06 = Bomb Pan Am Flight 103 Lockerbie December 21 1988
NovaQuery-07 = Nomination Senator John Tower Secretary of Defense
NovaQuery-08 = HIV Human immunodeficiency virus AIDS
NovaQuery-09 = Refugee Asylum central america
NovaQuery-10 = Operation Desert Shield Storm Persian Gulf War IRAQ Kuwait
NovaQuery-11 = Operation Restore Hope Somalia
NovaQuery-12 = bosnia
NovaQuery-13 = North Atlantic Council NATO Rome Summit
NovaQuery-14 = Francois Mitterand Bush March 1991
NovaQuery-15 = Independent Counsel Walsh Iran-Contra Affair
NovaQuery-16 = July 1989 Bush Hungary
NovaQuery-17 = nomination Clarence Thomas Supreme Court Justice
NovaQuery-18 = FY 1990 Federal Budget
NovaQuery-19 = Office of Government Ethics
NovaQuery-20 = Council on Foreign Relations
NovaQuery-21 = Savings loan crisis
NovaQuery-22 = Pakistan
NovaQuery-23 = India Pakistan nuclear proliferation
NovaQuery-24 = Arab Israel Israeli peace conference 1991
NovaQuery-25 = NAFTA North American Free Trade Agreement
NovaQuery-26 = NASA Russia joint space
NovaQuery-27 = Caspar Weinberger
NovaQuery-28 = apartheid South Africa
NovaQuery-29 = NIH National Institute of Health Research Funding
NovaQuery-30 = health care Legislation
NovaQuery-31 = UNCED Earth Summit Rio 1992
NovaQuery-32 = Council on Competitiveness
NovaQuery-33 = panama situation 1989
NovaQuery-34 = Gorbachev Malta 1989
NovaQuery-35 = Start Soviet arms control negotiations
NovaQuery-36 = coup Gorbachev
NovaQuery-37 = revolution Soviet Union December 1991
NovaQuery-38 = human rights
NovaQuery-39 = foreign development aid human rights
NovaQuery-40 = human rights China
NovaQuery-41 = human rights drug summit san antonio
NovaQuery-42 = trade china
NovaQuery-43 = Financial Military Assistance Turkey
NovaQuery-44 = Arms Exports IRAQ
NovaQuery-45 = Central Valley 1992
NovaQuery-46 = Structural Impediments Initiative SII
NovaQuery-47 = Super 301
NovaQuery-48 = Japan Trade and Economic Policy
NovaQuery-49 = Kaifu Bush meeting
NovaQuery-50 = Hurricane Andrew
NovaQuery-51 = Millie

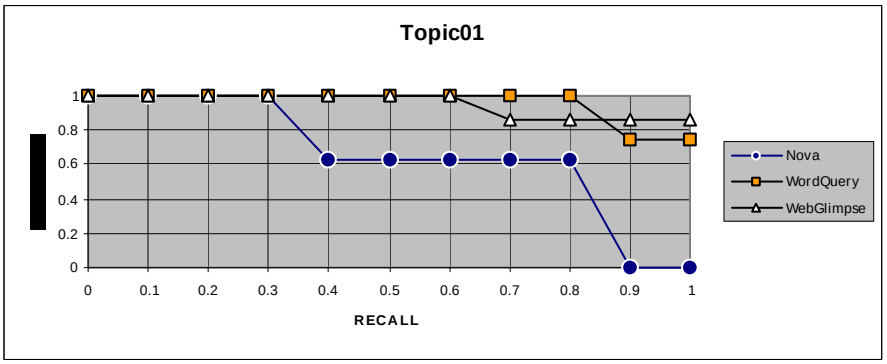
Appendix C: Oracle Text (Word Queries)

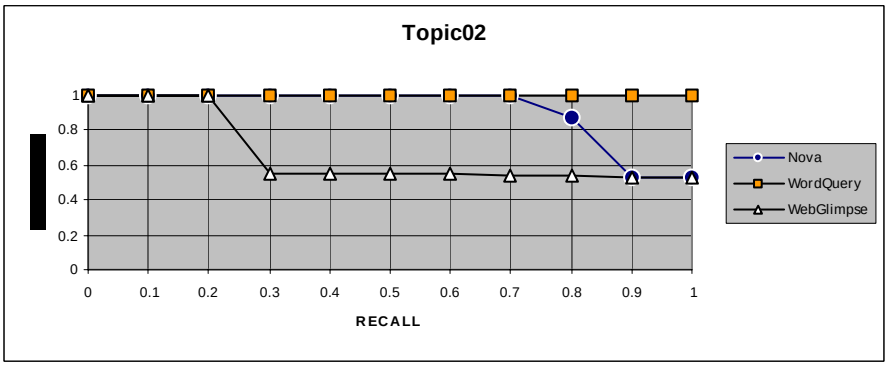
```
WordQuery-01 = '$china AND (({USIA}|united states information agency|u.s. information agency) OR ({VOA}|voice of
               america))'
WordQuery-02 = '(radio marti) OR (tv marti) OR ($cuba AND $broadcast)'
WordQuery-03 = '(aid=tuition=credit=credits=voucher=vouchers NEAR private=nonpublic=religious NEAR
               education=educational=school=schools)'
WordQuery-04 = '(domestic policy council)'
WordQuery-05 = '(southeast asia|vietnam|laos) AND (({POW}|prisoner of war) OR ({MIA}|missing in action))'
WordQuery-06 = '(pan am|flight 103|lockerbie) AND (bomb|bombing|terrorist|terrorism)'
WordQuery-07 = '(nominate|nominee|nomination) AND (john tower|senator tower|(defense AND tower))'
WordQuery-08 = '({HIV}|human immunodeficiency virus|{AIDS}|acquired immune deficiency syndrome)'
WordQuery-09 = '($refugee OR asylum) AND (central $america|$nicaragua|$guatemala|el $salvador)'
WordQuery-10 = '(military intervention,against Iraq,military force,military action,desert shield,desert storm,Gulf War)
               AND (Iraq)'
WordQuery-11 = '(somalia NEAR military) OR (operation restore hope)'
WordQuery-12 = '($bosnia)'
WordQuery-13 = '(rome summit|heads of state summit|north atlantic council) OR ({NATO} AND strategic concept)'
WordQuery-14 = '(march,14=14th,91=1991,meeting,Mitterand,Bush) AND (Mitterrand AND {March})'
WordQuery-15 = '(Iran,Contra,affair) AND (Walsh|Independent Counsel)'
WordQuery-16 = 'Hungary AND ($trip|$visit|$diplomacy|diplomatic|July NEAR 89=1989)'
WordQuery-17 = '(nominate=nominee=nomination,justice,judge,Supreme Court*5,Clarence Thomas*10,(Thomas
               NEAR nomination)*10) AND (Thomas) AND (supreme court)'
WordQuery-18 = '(FY=fiscal=budget) NEAR (90=1990)'
WordQuery-19 = '(Office of Government Ethics)'
WordQuery-20 = '(Council on Foreign Relations)'
WordQuery-21 = '($saving {and} $loan)' AND DOCNAME < 89080900
WordQuery-22 = '($pakistan)'
WordQuery-23 = '(nuclear=proliferation=nonproliferation) NEAR (India=Pakistan)'
WordQuery-24 = '(peace,conference,Madrid,(October Near 91=1991)) AND ($Arab|$Israel|Israeli|Middle East)'
WordQuery-25 = '({NAFTA}|North American Free Trade Agreement)'
WordQuery-26 = '(States) NEAR (Soviet=Russia) NEAR (NASA=space)'
WordQuery-27 = '(caspar|weinberger) ACCUM (caspar weinberger)'
WordQuery-28 = '(apartheid,South Africa,internal,situation,political,struggle) AND ((apartheid) OR (South Africa))'
WordQuery-29 = '({NIH}|National Institutes of Health) AND ($research|$funding)'
WordQuery-30 = '(health care) NEAR ($legislation=$expand=$access)'
WordQuery-31 = '({UNCED}|earth summit) OR (Rio NEAR 92=1992)'
WordQuery-32 = '(competitive=competitiveness NEAR council=Quayle)'
WordQuery-33 = '(military=decision=situation=89=1989) AND Panama' AND DOCNAME < 89122100
WordQuery-34 = '(Gorbachev) AND (Malta|NEAR((November=December,89=1989),5))'
WordQuery-35 = '(({U.S.}|United States) AND ($Soviet|$Russia)  AND (arms control)) OR (START I=II)'
WordQuery-36 = '($coup AND $Gorbachev)'
WordQuery-37 = '(($revolution) NEAR ($Soviet=$Russia)) OR ((December NEAR 91=1991) AND ($Soviet=$Russia))'
               AND DOCNAME > 91119999
WordQuery-38 = '(human $right)' AND DOCNAME < 90000000
WordQuery-39 = '(human $right NEAR foreign=development=aid)'
WordQuery-40 = '(human right=rights) NEAR (china=chinese)'
WordQuery-41 = '(human $right AND drug summit AND San Antonio AND (February NEAR 92=1992))'
WordQuery-42 = '(trade NEAR china)'
WordQuery-43 = '(turkey NEAR finance=financial NEAR assist=assistance) OR (turkey NEAR military NEAR
               assist=assistance)'
WordQuery-44 = '(export=exports=arms) NEAR Iraq'
WordQuery-45 = '(Central Valley Project)'
WordQuery-46 = '({SII}|Structural Impediments Initiative)'
WordQuery-47 = '(Super 301)'
WordQuery-48 = '($Japan=Japanese NEAR $trade=$economic)'
WordQuery-49 = '(Kaifu,Bush,$minister,ministerial,$meeting,$discuss,$discussion,(March NEAR 90=1990),(April
               NEAR 91=1991),(July NEAR 91=1991),(September NEAR 89)) AND (Kaifu)'
WordQuery-50 = '($hurricane,$andrew,($hurricane $andrew)*10) AND ($hurricane AND $andrew)'
WordQuery-51 = '(dog,pet,canine,Millie) AND Millie'
```

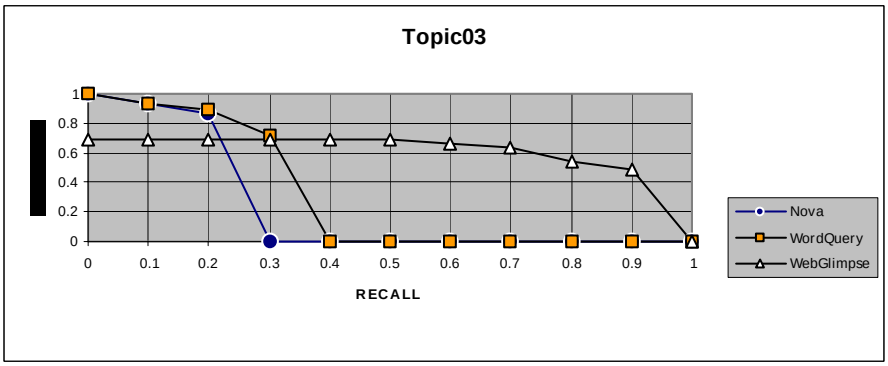
Appendix D: WebGlimpse Queries

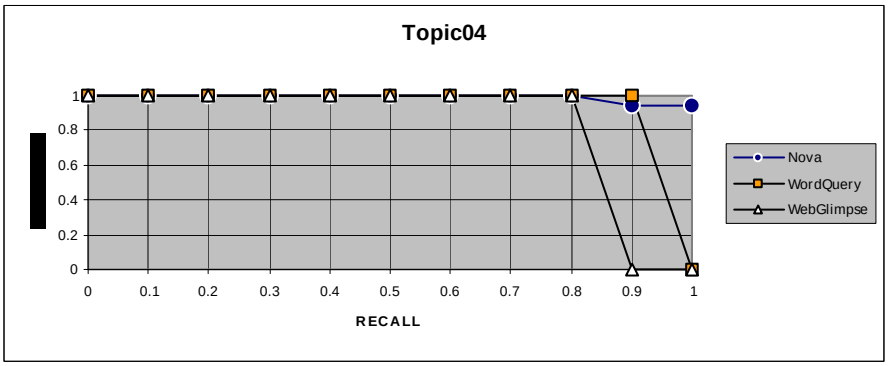
WebGlimpseQuery-01 = China;USIA,VOA,{United States Information Agency},{U%S% Information Agency},{Voice of America}
WebGlimpseQuery-02 = Radio Marti,TV Marti,{Cuba;broadcast}
WebGlimpseQuery-03 = education;{tuition,credit,credits,voucher,vouchers}
WebGlimpseQuery-04 = Domestic Policy Council
WebGlimpseQuery-05 = {POW,MIA};{Southeast Asia,Vietnam,Laos}
WebGlimpseQuery-06 = {Pan Am,Flight 103,Lockerbie};{bomb,bombing,terrorist,terrorism}
WebGlimpseQuery-07 = {nominate,nominee,nomination};{Senator Tower,John Tower,{defense;Tower}}
WebGlimpseQuery-08 = HIV,{Human Immunodeficiency Virus},AIDS,{Acquired Immune Deficiency Syndrome}
WebGlimpseQuery-09 = {refugee,asylum};{Central America},{El Salvador},{Nicaragua},{Guatemala}
WebGlimpseQuery-10 = Iraq;{{military intervention},{against Iraq},{military force},{military action},{Desert Shield},{Desert Storm},{Gulf War}}
WebGlimpseQuery-11 = {Somalia;military},{Operation Restore Hope}
WebGlimpseQuery-12 = Bosnia
WebGlimpseQuery-13 = {Rome;Summit},{North Atlantic Council},{Heads of State Summit},{NATO;strategic concept}
WebGlimpseQuery-14 = Mitterand;Bush;{March,meeting}
WebGlimpseQuery-15 = {Iran;contra};{Independent Counsel},{Walsh}
WebGlimpseQuery-16 = Hungary;{M07;Y89},{trip,visit,diplomacy,diplomatic}
WebGlimpseQuery-17 = {nominate,nominee,nomination};{{Clarence Thomas},{Thomas;Supreme Court}}
WebGlimpseQuery-18 = Budget;{Fiscal Year 1990},{FY 1990}
WebGlimpseQuery-19 = {Office of Government Ethics}
WebGlimpseQuery-20 = {Council on Foreign Relations}
WebGlimpseQuery-21 = {{Savings and Loan},{S&L}};{Y89};{M01,M02,M03,M04,M05,M06,M07,M08}
WebGlimpseQuery-22 = Pakistan
WebGlimpseQuery-23 = {India,Pakistan};{nuclear,proliferation}
WebGlimpseQuery-24 = Arab;{Israel,Israeli};{peace,conference,Madrid,{M10;Y91}}
WebGlimpseQuery-25 = {NAFTA,{North American Free Trade Agreement}}
WebGlimpseQuery-26 = {U%S%,United States};{Soviet,Russia};{NASA,space}
WebGlimpseQuery-27 = {Caspar,Weinberger}
WebGlimpseQuery-28 = {South Africa};{apartheid,discrimination}
WebGlimpseQuery-29 = {NIH,{National Institutes of Health}};{research,funding}
WebGlimpseQuery-30 = {health care};{legislation,law}
WebGlimpseQuery-31 = UNCED,{Earth Summit},{Rio;{Y92}}
WebGlimpseQuery-32 = {{Council on Competitiveness},{Quayle;{competition,competitive,competitiveness}}}
WebGlimpseQuery-33 = Panama;{military,decision,situation};{Y89}
WebGlimpseQuery-34 = Bush;Gorbachev;Malta
WebGlimpseQuery-35 = {United States};{Soviet,Russia};{arms control,START I,START II}
WebGlimpseQuery-36 = Gorbachev;coup
WebGlimpseQuery-37 = {U%S%S%R,Soviet,Russia};{breakdown,collapse,revolt,revolution,{{Y91};{M12}}}
WebGlimpseQuery-38 = {human rights};{Y89}
WebGlimpseQuery-39 = {human rights};{developmental aid,foreign development,foreign aid,aid development}
WebGlimpseQuery-40 = China;{human rights}
WebGlimpseQuery-41 = {human rights};{{drug summit},{San Antonio;summit}}
WebGlimpseQuery-42 = China;trade
WebGlimpseQuery-43 = Turkey;{financ,militar}
WebGlimpseQuery-44 = Iraq;{arms,weapon};export
WebGlimpseQuery-45 = {central valley project}
WebGlimpseQuery-46 = {SII},{Structural Impediments Initiative}
WebGlimpseQuery-47 = {Super 301}
WebGlimpseQuery-48 = Japan;{trade,economic}
WebGlimpseQuery-49 = Kaifu;{Bush,meeting}
WebGlimpseQuery-50 = hurricane;Andrew
WebGlimpseQuery-51 = Millie

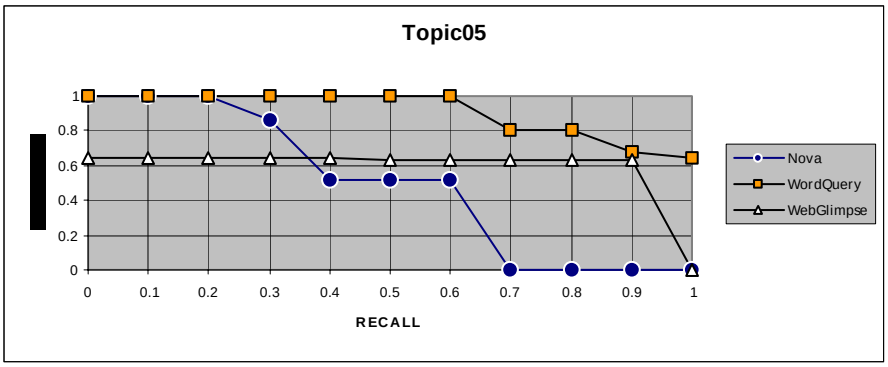
Appendix E: Precision Recall Graphs

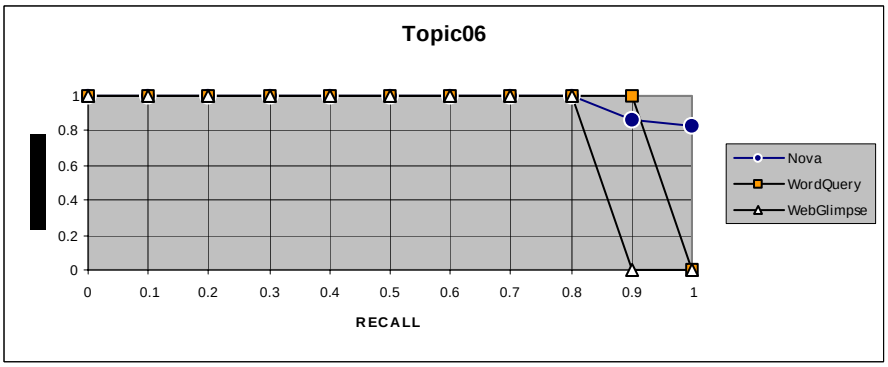


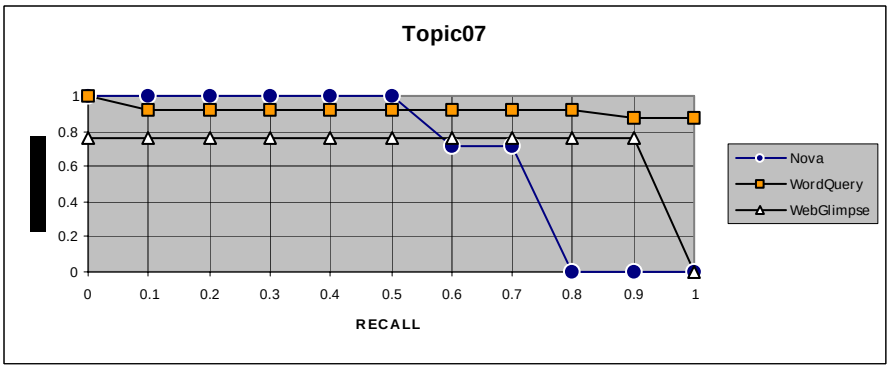


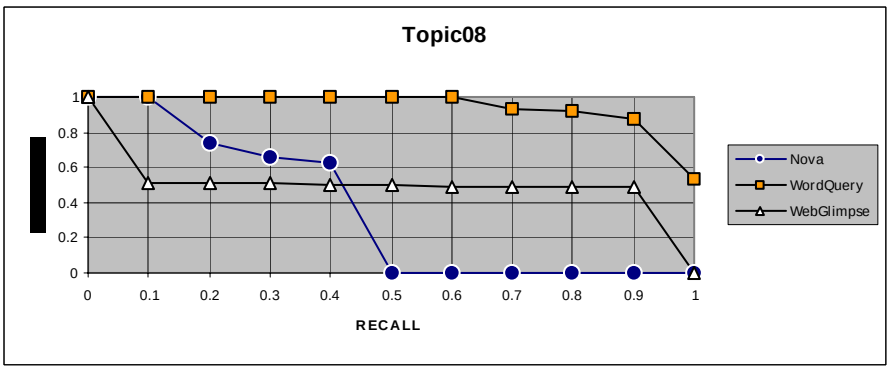


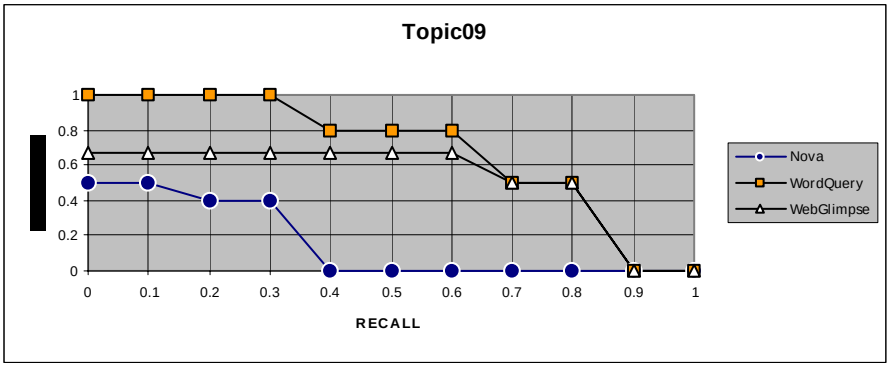


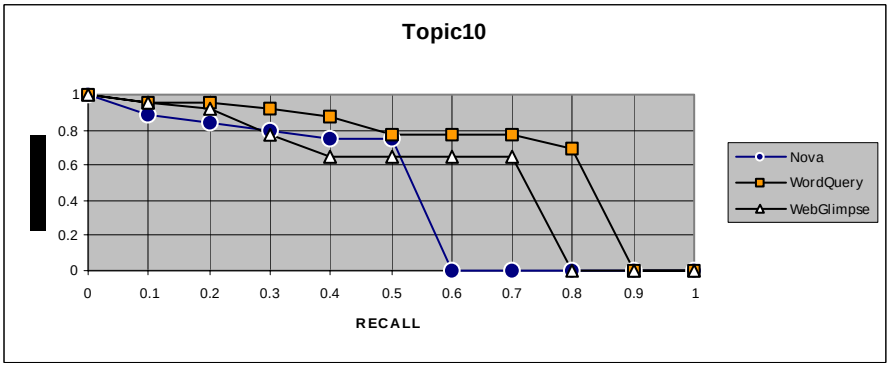

Topic05
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

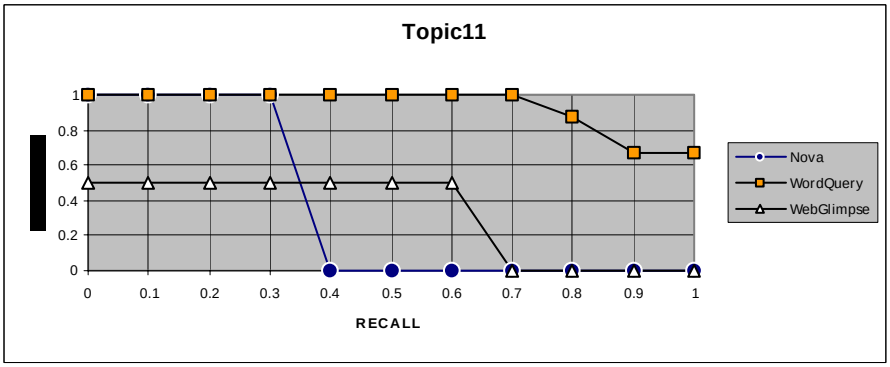

Topic06
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

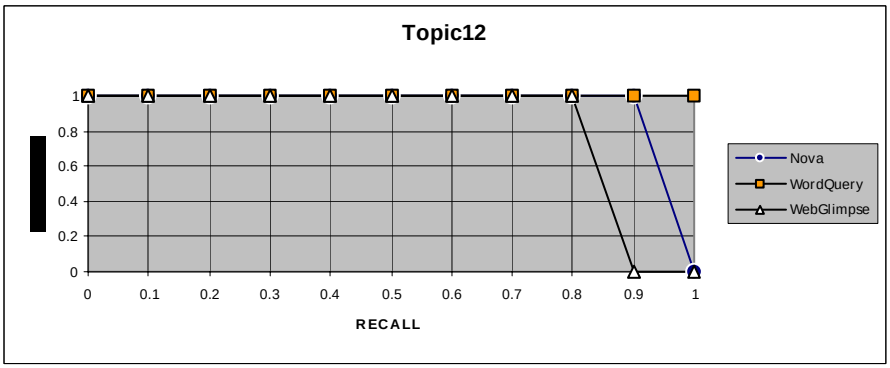

Topic07
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

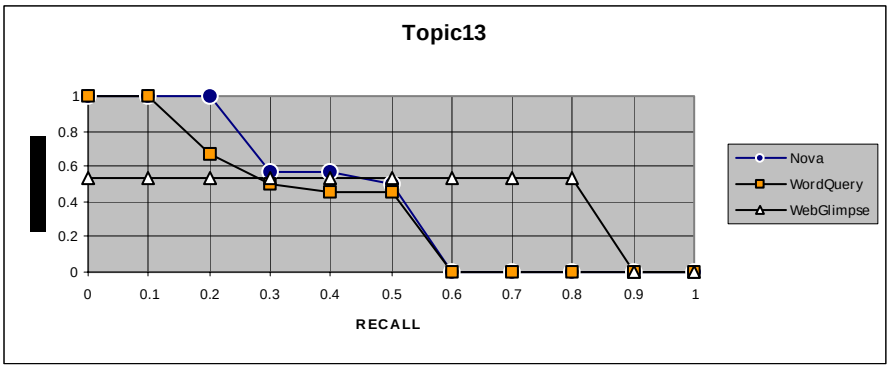

Topic08
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

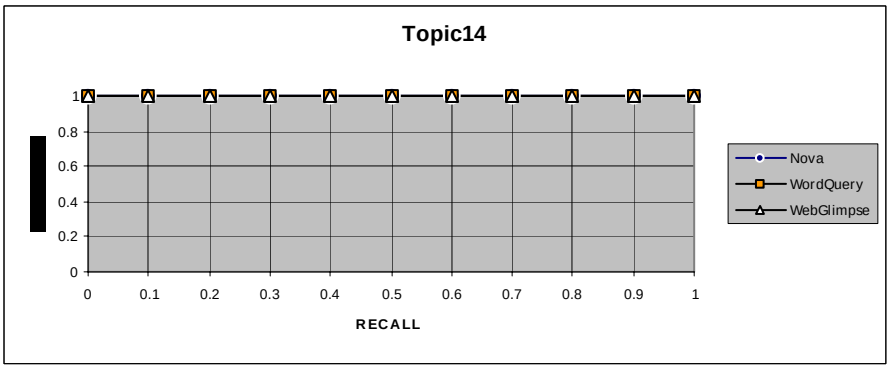

Topic09
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

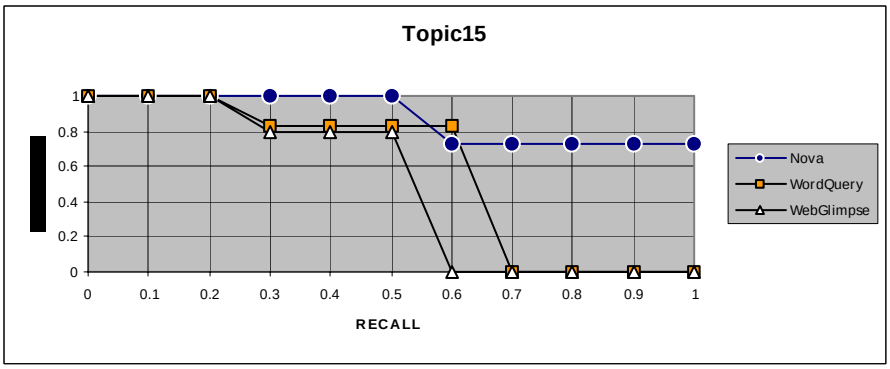

Topic10
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

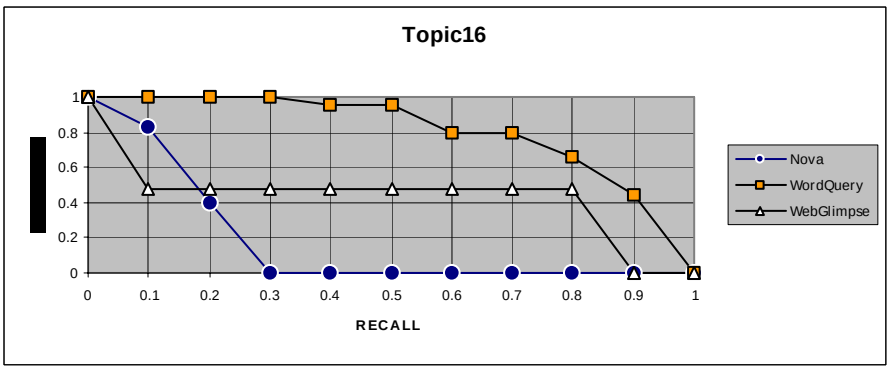

Topic11
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

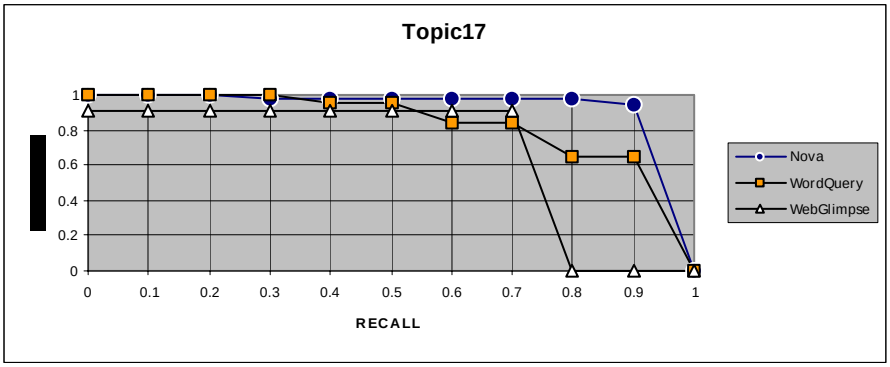

Topic12
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

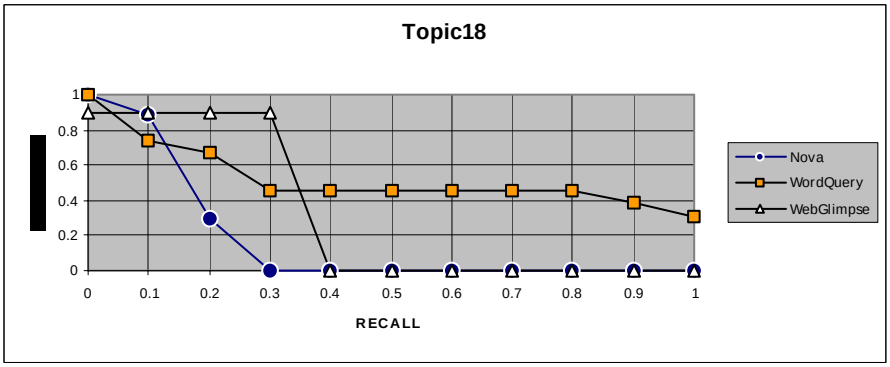

Topic13
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

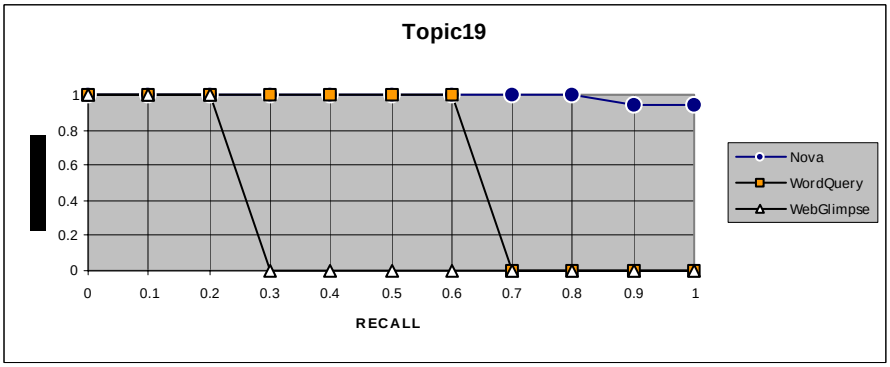

Topic14
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

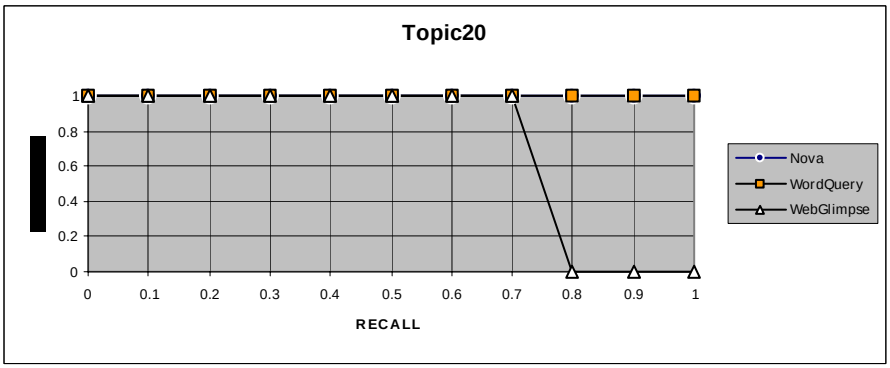

Topic15
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

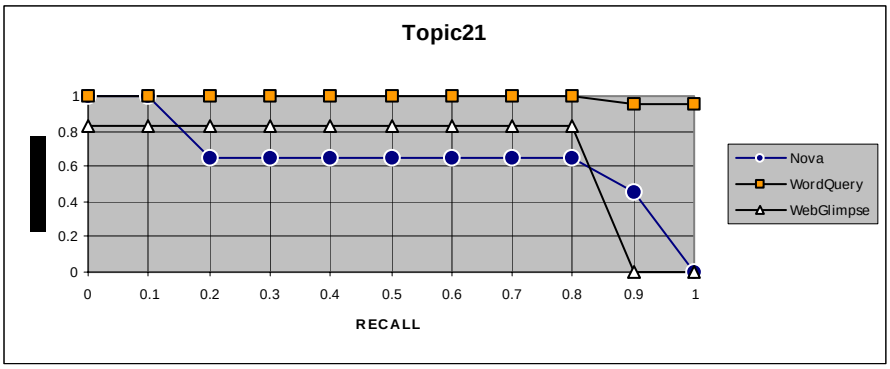

Topic16
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

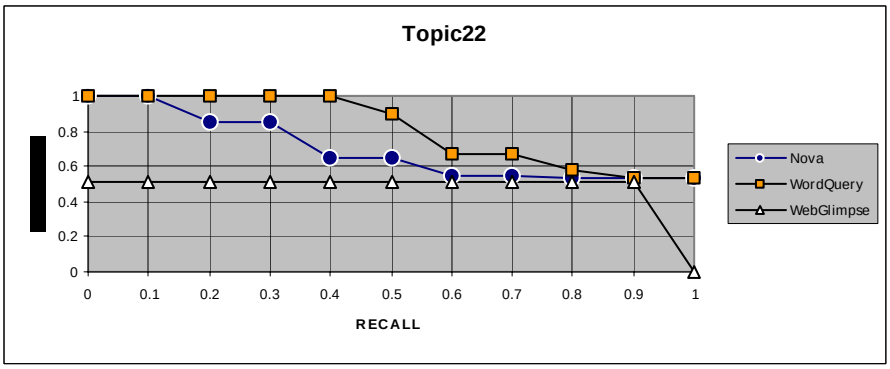

Topic17
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

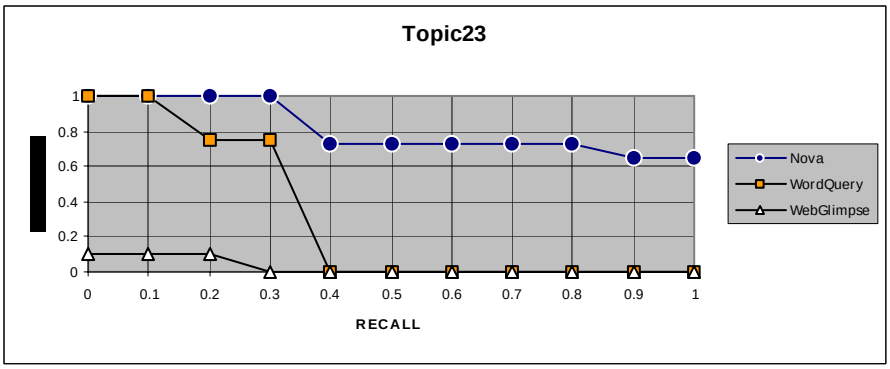

Topic18
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

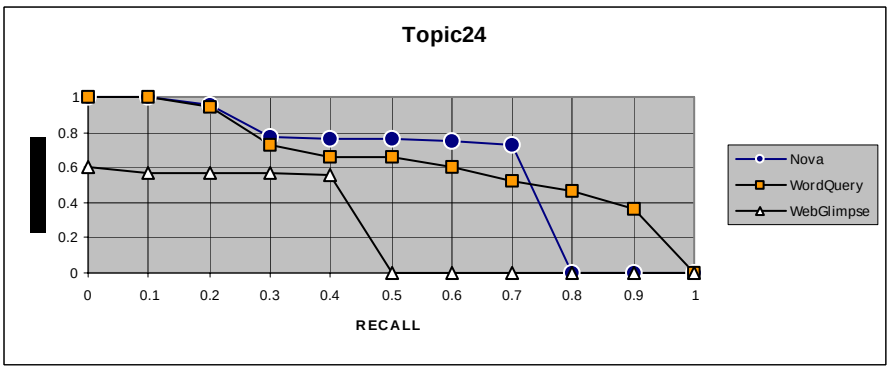

Topic19
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

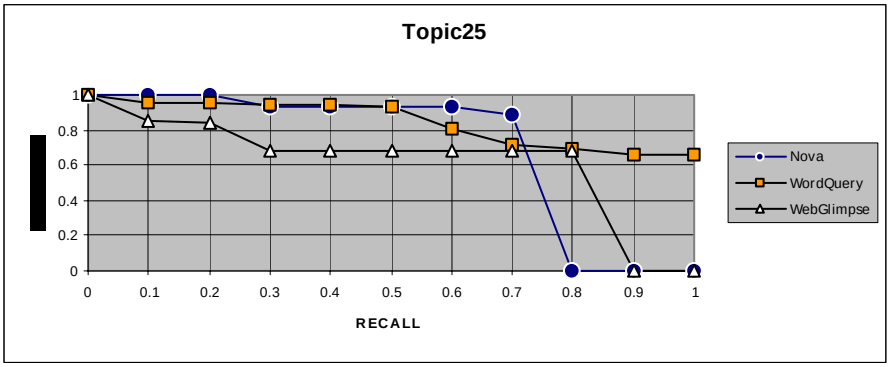

Topic20
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

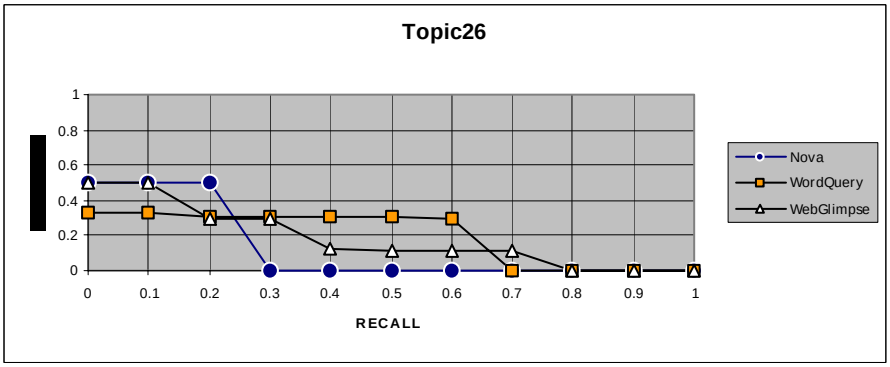

Topic21
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

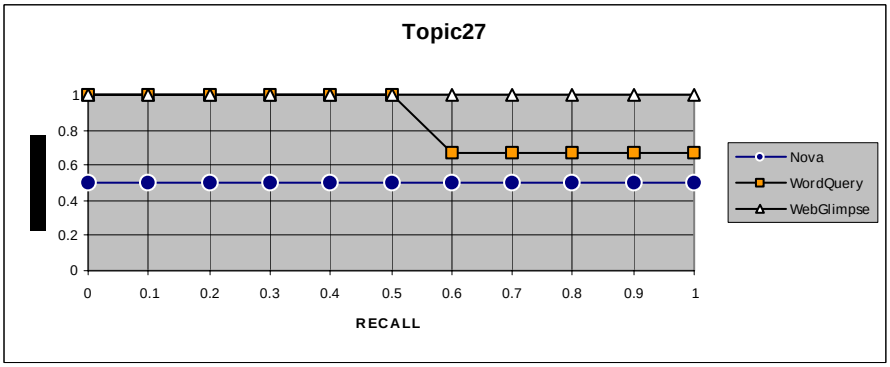

Topic22
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

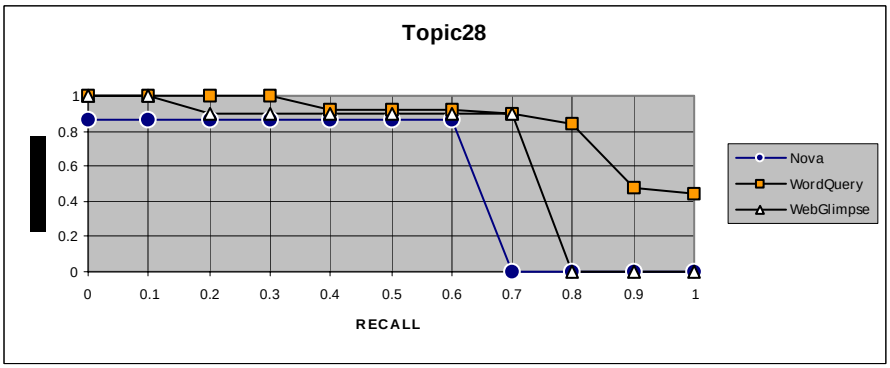

Topic23
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

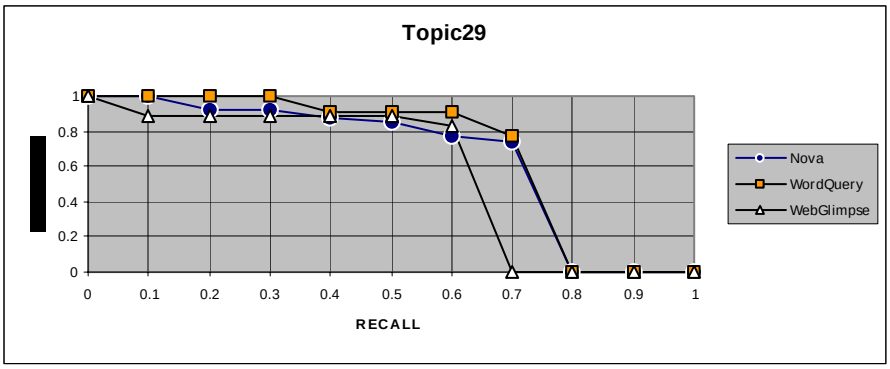

Topic24
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

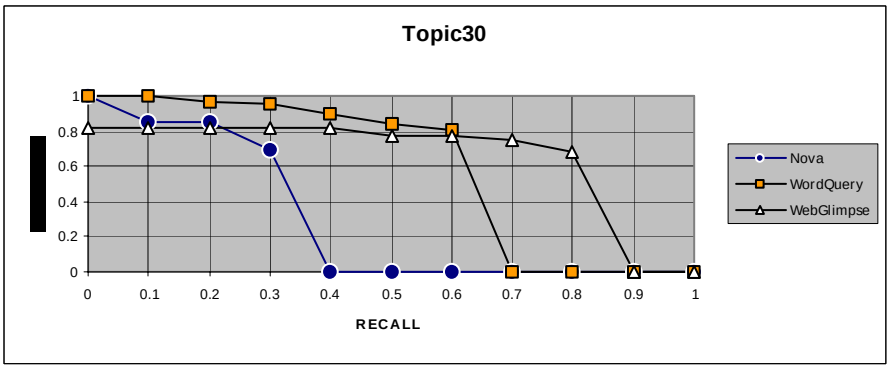

Topic25
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

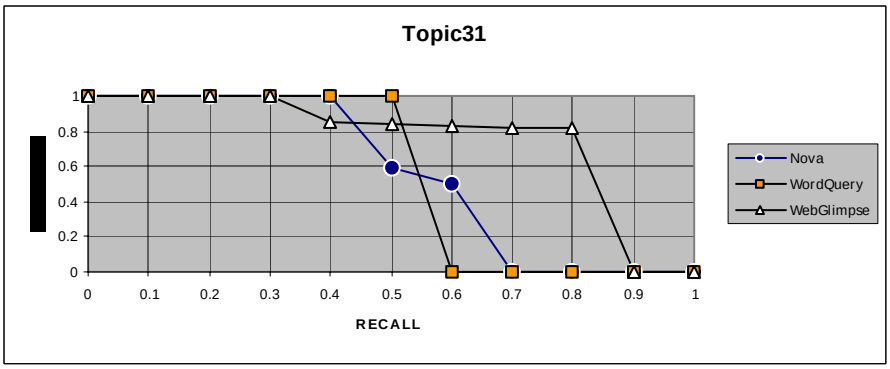

Topic26
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

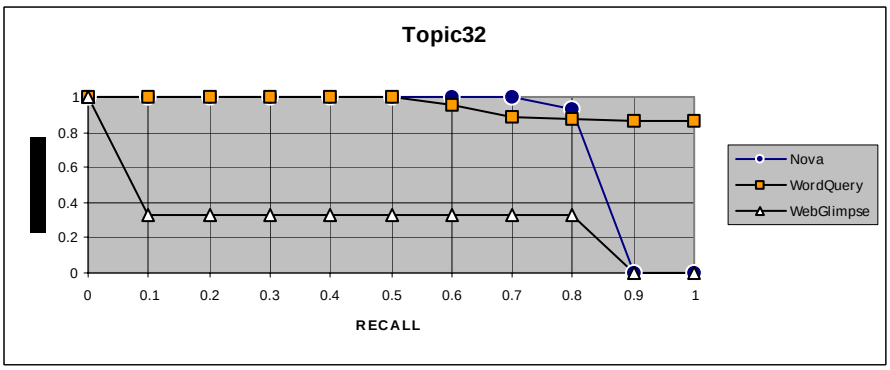

Topic27
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

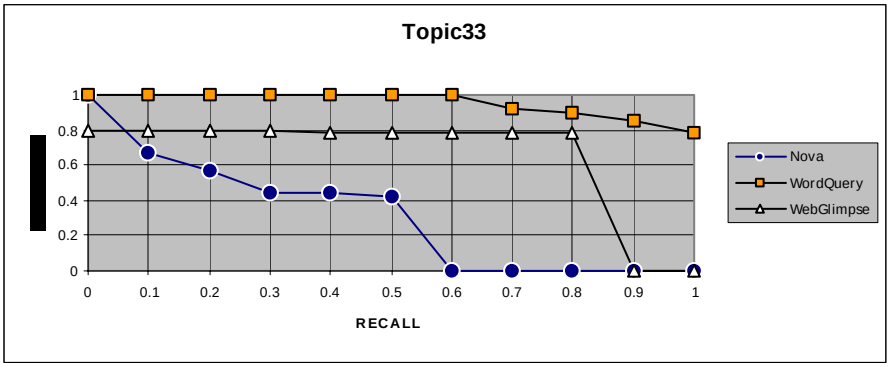

Topic28
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

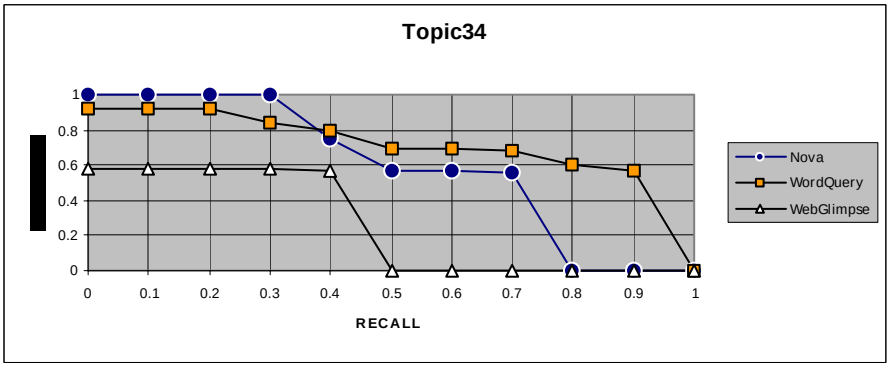

Topic29
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

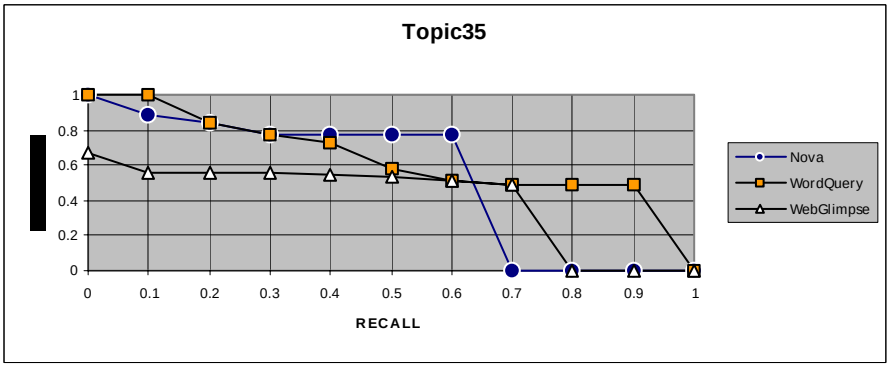

Topic30
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

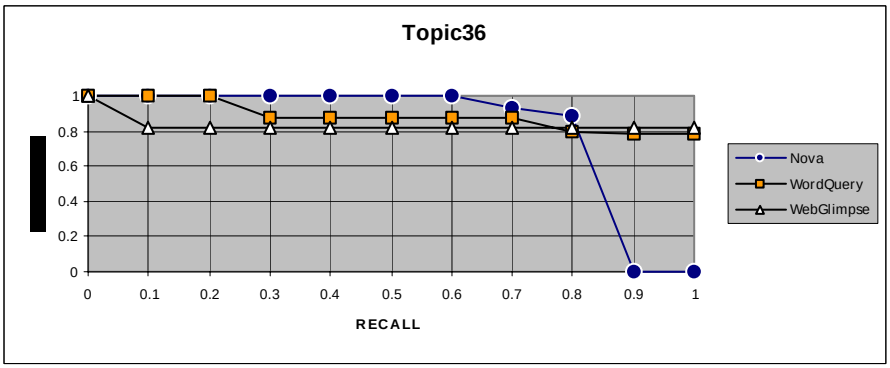

Topic31
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

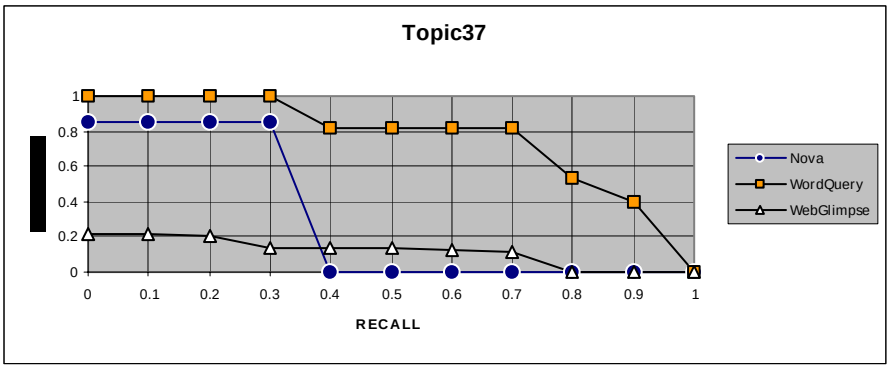

Topic32
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

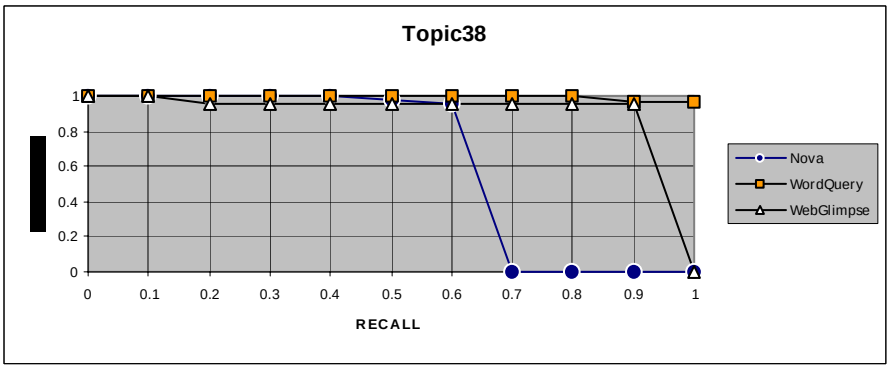

Topic33
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

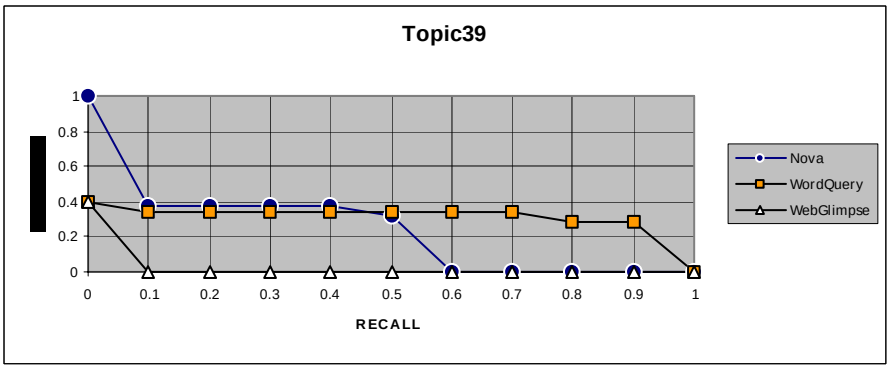

Topic34
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

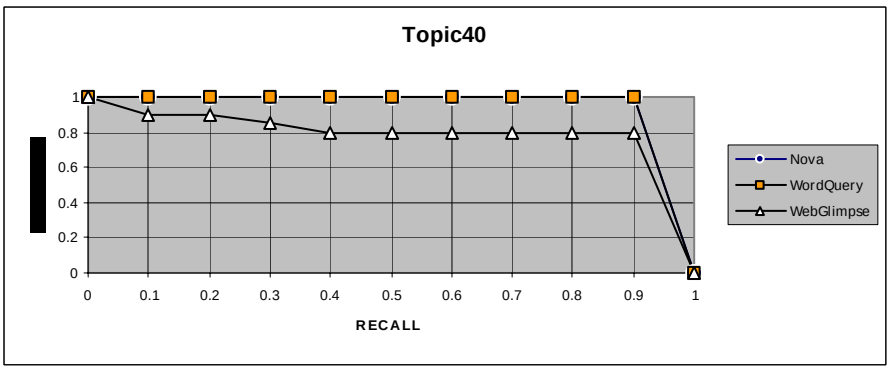

Topic35
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

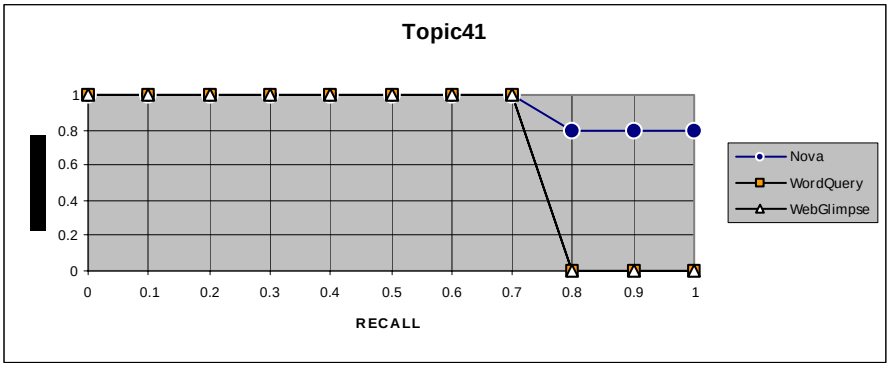

Topic36
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

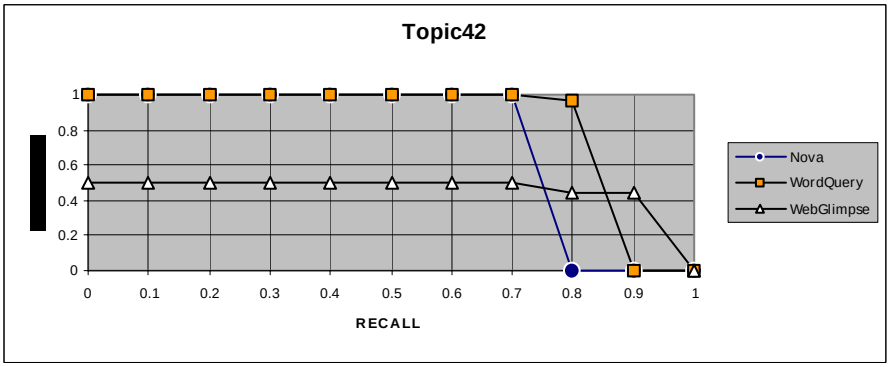

Topic37
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

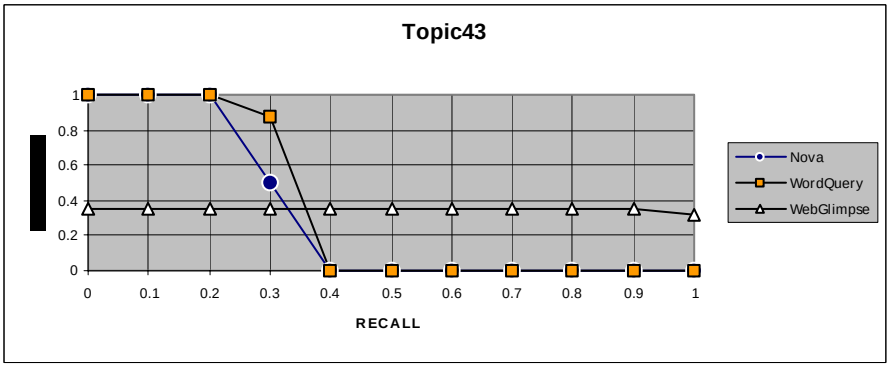

Topic38
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

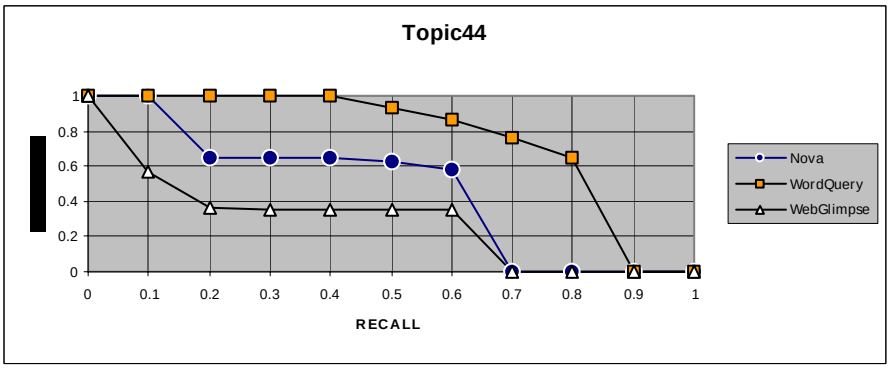

Topic39
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

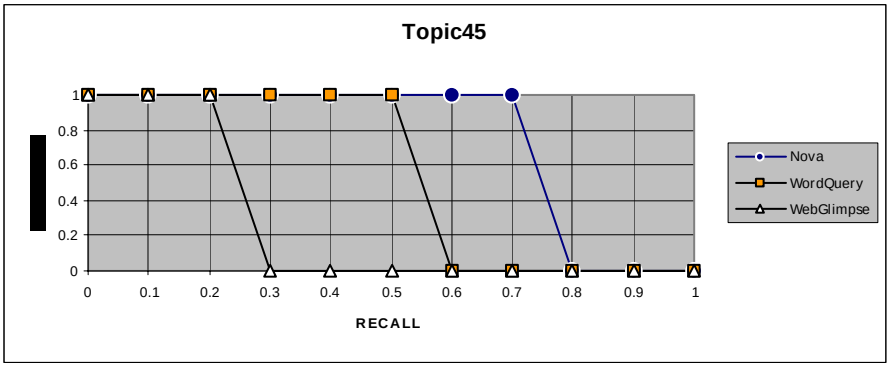

Topic40
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

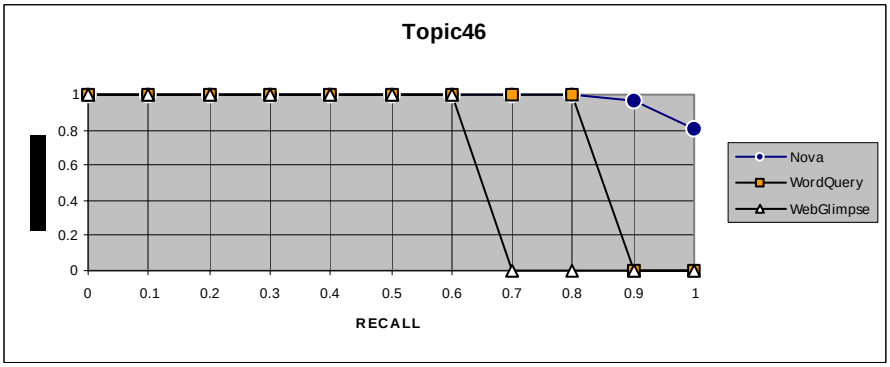

Topic41
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

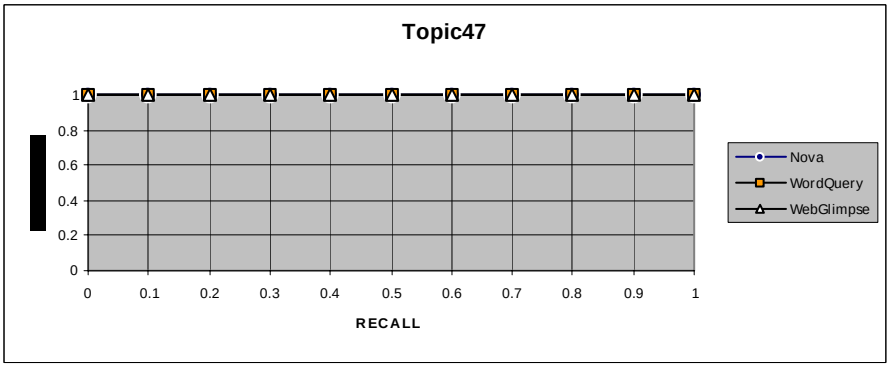

Topic42
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

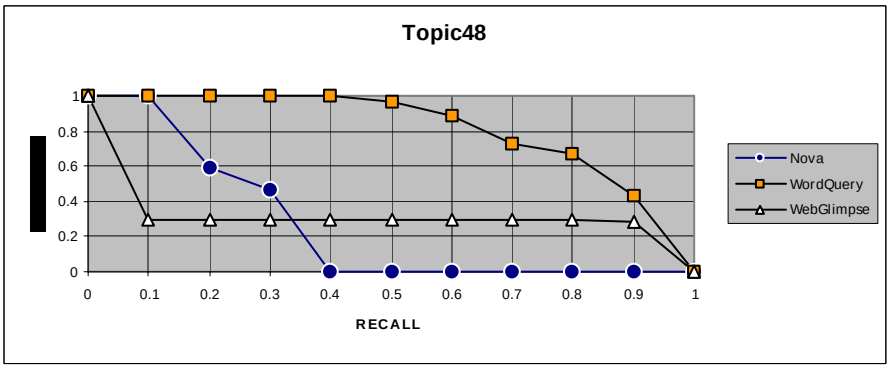

Topic43
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

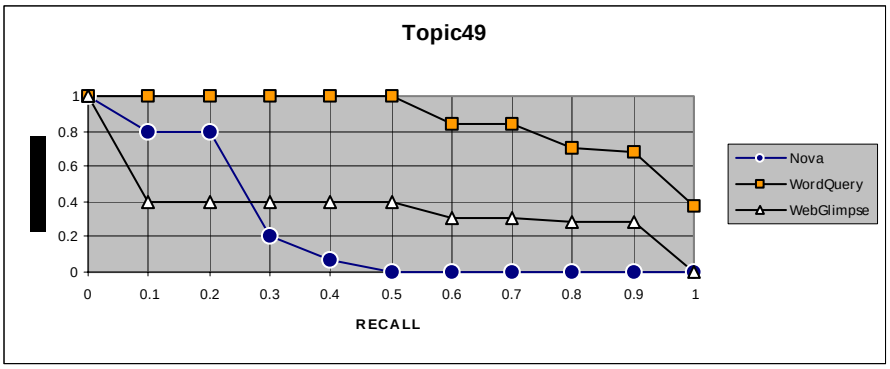

Topic44
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

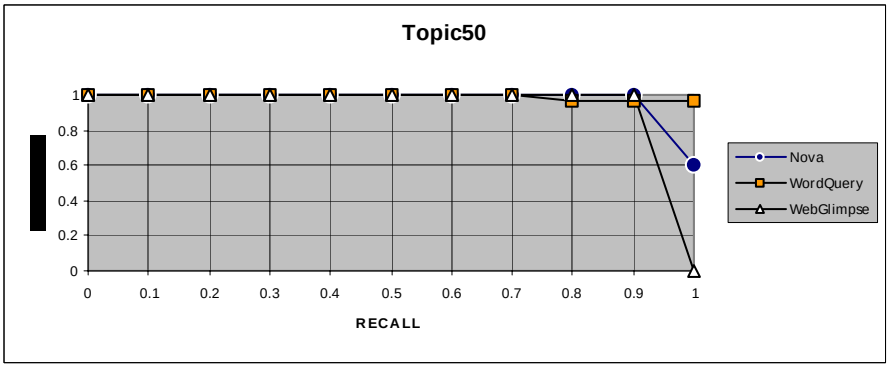

Topic45
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

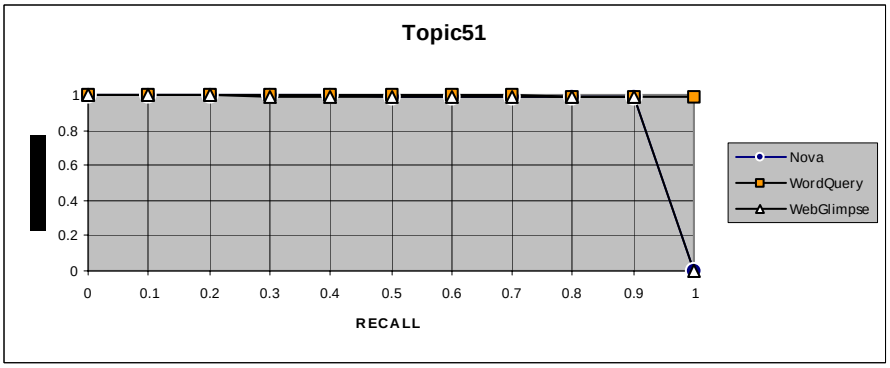

Topic46
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse

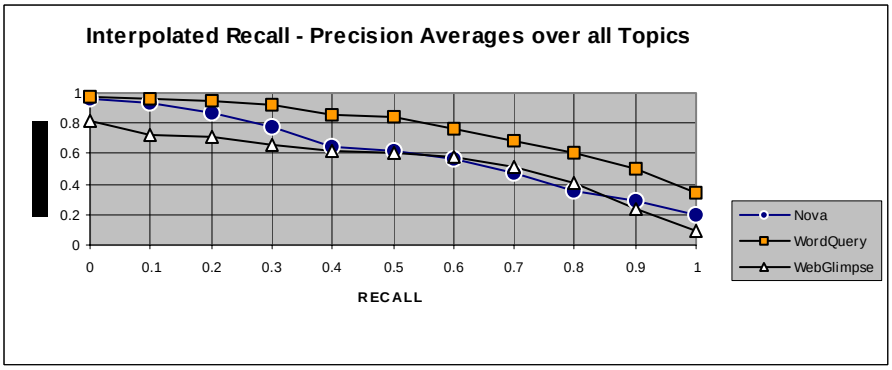

Topic47
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse


Topic48
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse


Topic49
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse


Topic50
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse


Topic51
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse


Interpolated Recall - Precision Averages over all Topics
1
0.8
0.6
0.4
0.2
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
RECALL
Nova
WordQuery
WebGlimpse