UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

UIPATH, INC.,

Petitioner,

v.

RULE 14 LLC.,

Patent Owner.

Case No. IPR2024-IPR2024-01306
U.S. Patent No. 11,048,712

DECLARATION OF DESIREE L. WOLF

I, Desiree L. Wolf, make the following Declaration pursuant to 28 U.S.C. § 1746:

1. I am a paralegal with the law firm of Volpe Koenig, attorneys for Petitioner in this matter.

2. On August 15, 2024, I accessed the following URL link: https://ciir-publications.cs.umass.edu/

3. On the web page accessible at the URL link listed in Paragraph 2, on August 15, 2024, I entered the name "Zhihong Lu" into the "Authors to include" search box, and the search terms "Measures in Collection Ranking Evaluation" in the "Find publications with this content" search box. Attached as Exhibit A is a true and correct copy of a screen capture I created on August 15, 2024, showing the web page with the search terms entered.

4. On August 15, 2024, after entering the search terms, I then hit "search." A search results web page appeared. Attached as Exhibit B is a true and correct copy of a screen capture I created on August 15, 2024 of the search results web page.

5. On the search results web page listed in Exhibit B, next to the first search result for the article "Measures in Collection Ranking Evaluation," a clickable "[View bibtex]" link is shown. On August 15, 2024, I clicked that link, and a pop-up window appeared. Attached as Exhibit C is a true and correct copy of a screen

capture I created on August 15, 2024 of the search results web page with the "[View bibtex]" pop-up window appearing.

6. On August 15, 2024, I clicked the hyperlink for the article "Measures in Collection Ranking Evaluation," which brought up a copy of the article. Attached Exhibit D is a true and correct copy of a screen capture I created on August 15, 2024 showing the first screen of the article as it appeared at the URL link: https://ciir-publications.cs.umass.edu/getpdf.php?id=62

7. The URL link listed in Paragraph 6 allowed for the download of a PDF document ("Measures in Collection Ranking Evaluation"), which I downloaded on August 15, 2024. A true and correct copy of the document is attached as Exhibit E. A copy of this document is also included as Ex. 1008 in connection with the subject IPR Petition.

8. On August 15, 2024, I accessed the following URL link on the Internet Archive ("Wayback Machine") website: https://web.archive.org/web/20040405195344/http://www.cs.virginia.edu:80/

9. Attached as Exhibit F is a true and correct copy of a screen capture I created on August 15, 2024 of the Wayback Machine web page for the URL link listed in Paragraph 8.

10. On August 15, 2024, I clicked the hyperlink for "Research" and accessed the following URL link on the Wayback Machine website:

https://web.archive.org/web/20040401140136/http://www.cs.virginia.edu/research/

11. Attached as Exhibit G is a true and correct copy of a screen capture I created on August 15, 2024 of the Wayback Machine web page for the URL link listed in Paragraph 10.

12. On August 15, 2024, I clicked the hyperlink for "Cyberia - Information Retrieval Group" and accessed the following URL link on the Wayback Machine website:

https://web.archive.org/web/20040505070445/http://www.cs.virginia.edu/~cyberia/

13. Attached as Exhibit H is a true and correct copy of a screen capture I created on August 15, 2024 of the Wayback Machine web page for the URL link listed in Paragraph 12.

14. On August 15, 2024, I clicked the hyperlink for "Papers" and accessed the following URL link on the Wayback Machine website:

https://web.archive.org/web/20040625112640/http://www.cs.virginia.edu/~cyberia/papers.html

15. Attached as Exhibit I is a true and correct copy of a screen capture I created on August 15, 2024 of the Wayback Machine web page for the URL link listed in Paragraph 14.

16. On August 15, 2024, I clicked the hyperlink for "PDF" under the entry for French, J. C., A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey and Y.

Mou, "Comparing the Performance of Database Selection Algorithms," Technical Report CS-99-03, Dept. of Computer Science, University of Virginia, January 12, 1999, and accessed the following URL link on the Wayback Machine website:

https://web.archive.org/web/20040329005629/http://www.cs.virginia.edu/~cyberia/papers/SIGIR99.pdf

17. Attached as Exhibit J is a true and correct copy of a screen capture I created on August 15, 2024 of the Wayback Machine web page for the URL link listed in Paragraph 16.

18. The URL link listed in Paragraph 16 allowed for the download of a PDF document ("Comparing the Performance of Database Selection Algorithms"), which I downloaded on August 15, 2024. A true and correct copy of the document is attached as Exhibit K. A copy of this document is also included as Ex. 1009 in connection with the subject IPR Petition.

19. To create Exhibits A, B and D-K, I used the Chrome web browser and the screen capture program "Awesome Screenshot."

20. To create Exhibit C, I used the Chrome web browser and the screen capture program "Greenshot."

21. I hereby declare under penalty of perjury under the laws of the United States that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true. I understand

willful false statements and the like are punishable by fine or imprisonment, or both, under § 1001 of Title 18 of the United States Code.

Dated: August 16, 2024

By: Desiree L. Wolf
Desiree L. Wolf

EXHIBIT A

UMassAmherst

CIIR CENTER FOR INTELLIGENT INFORMATION RETRIEVAL
UNIVERSITY OF MASSACHUSETTS AMHERST

Publication Database

Fill in any of the year, author, or content fields.

INDRI SEARCH SYSTEM

Years to include (1984-Present)
Example: 2001, 2003, 2001-2003

Authors to include
Zhihong Lu
Example: Croft, James Allan

Find publications with this content
Measures in Collection Ranking Evaluation

Limit to this publication type
All

List all publications starting with the most recent , or Browse by Author

search

This site is maintained by the Department of Computer Science/Center for Intelligent Information Retrieval
© 2024 University of Massachusetts Amherst • Site Policies

EXHIBIT B

UMassAmherst

CIIR CENTER FOR INTELLIGENT INFORMATION RETRIEVAL
UNIVERSITY OF MASSACHUSETTS AMHERST

Results from CIIR Publications

Sort By: Relevance Go

IR-89: (1996) Lu, Z., "Measures in Collection Ranking Evaluation," Umass Technical Report TR96-39. [View bibtex]

> IR-89 1996 **Measures in Collection Ranking Evaluation Zhihong Lu**

IR-168: (1999) Lu, Z., "Scalable Distributed Architectures for Information Retrieval," Ph.D. dissertation. [View Abstract] [View bibtex]

> **...**1999 Scalable Distributed Architectures for Information Retrieval dissertation **Zhihong Lu** SCALABLE DISTRIBUTED ARCHITECTURES FOR INFORMATION RETRIEVAL A**...**Graduate School of the University of Massachusetts Amherst **in** partial fulfillment of the requirements for the**...**to my advisor Kathryn McKinley for her contribution **in** the development of my research and her guidance **in** my professional growth Without her broad vision**...**

Back to Search Form

This site is maintained by the Department of Computer Science/Center for Intelligent Information Retrieval
© 2024 University of Massachusetts Amherst • Site Policies

EXHIBIT C

Center for Intelligent Informatic

ciir-publications.cs.umass.edu

UMassAmherst

CIIR CENTER FOR INTELLIGENT INFORMATION RETRIEVAL
UNIVERSITY OF MASSACHUSETTS AMHERST

CIIR HOME
ABOUT THE CIIR
PUBLICATIONS
MEMBERSHIP
PERSONNEL
WHAT'S NEW
CONTACT US
CS DEPARTMENT

Results from CIIR Publications

Sort By: Relevance Go

IR-89: (1996) **Lu, Z.**, "**Measures in Collection Ranking Evaluation**," Umass Technical Report TR96-39. [View bibtex]

> IR-89 1996 **Measures in Collection Ranking Evaluation Zhihong Lu**

IR-168: (1999) **Lu, Z.**, "**Scalable Distributed Architectures for Information Retrieval**," Ph.D. dissertation. [View Abstract] [View bibtex]

> **...**1999 Scalable Distributed Architectures for Information Retrieval dissertation **Zhihong Lu** SCALABLE DISTRIBUTED ARCHITECTURES FOR INFORMATION RETRIEVAL A**...**Graduate School of the University of Massachusetts Amherst **in** partial fulfillment of the requirements for the**...**to my advisor Kathryn McKinley for her contribution **in** the development of my research and her guidance **in** my professional growth Without her broad vision**...**

Back to Search Form

This site is maintained by the Department of Computer Science/Center for Intelligent Information Retrieval
© 2024 University of Massachusetts Amherst • Site Policies

Bibliography - Google Chrome

ciir-publications.cs.umass.edu/pub/web/getbib.php?id=62

```
@techreport{IR-89,
author = {Zhihong Lu},
title = {Measures in Collection Ranking Evaluation},
year = {1996},
note = {Umass Technical Report TR96-39},
type = {IR},
number = {IR-89},
}
```

https://ciir-publications.cs.umass.edu/pub/web/search.php#

11:11 AM
8/15/2024

EXHIBIT D

Captured at: 2024/08/15 11:14 AM URL: https://ciir-publications.cs.umass.edu/getpdf.php?id=62

getpdf.php 1 / 9 110%

Measures in Collection Ranking Evaluation

Zhihong Lu James P. Callan W. Bruce Croft

Computer Science Department, University of Massachusetts
Amherst, MA 01003-4610
{zlu,callan,croft}@cs.umass.edu

Abstract

As a technique to hunt information on the Internet, ***collection location*** has received more attention. Several approaches have been proposed to solve this problem. All these approaches adopt the same procedure: ranking the collections and returning the top-ranks. But these approaches define different measures to evaluate ***collection ranking*** and the measures have significant weaknesses. In this paper, we survey the measures used in current research and propose a new pair of measures that are based on the concepts of precision and recall. The new measures overcome the problems found in the current measures.

1 Introduction

As hundreds or even thousands of collections become available on the Internet, it is practically impossible to query all of them when searching for information on a given topic: not only does such an exhaust search take a long time to complete, but network resource constraints such as limited bandwidth may also prevent the search from completing. Consequently the need to narrow the search to a few useful collections raises the issue of how to locate the most useful collections. Several approaches[1][3][2][6] have been proposed to solve this problem. All these approaches adopt the same procedure: ranking the collections and returning the top-ranks. Although collection ranking is similar to document ranking, these two problems are not identical, and thus the measures used in document ranking can not be used in collection ranking without modifications. Different measures have been defined to evaluate collection ranking. Some come from the measures used in conventional document retrieval and some do not. In this paper, we survey the measures for collection ranking used in the current research, discuss their strengthes and weaknesses, and propose new measures based on the concepts of precision and recall. In Section 2, we briefly survey the measures used in our previous research[1], NetSerf[2] and GLOSS[4][5]. In Section 3 we define the new measures, which are variations of precision and recall. In the final section, we summerize the discussions.

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT E

IPR2024-01306
UIPATH EXHIBIT 1024

IR-89

Measures in Collection Ranking Evaluation

Zhihong Lu James P. Callan W. Bruce Croft

Computer Science Department, University of Massachusetts
Amherst, MA 01003-4610
{zlu,callan,croft}@cs.umass.edu

Abstract

As a technique to hunt information on the Internet, ***collection location*** has received more attention. Several approaches have been proposed to solve this problem. All these approaches adopt the same procedure: ranking the collections and returning the top-ranks. But these approaches define different measures to evaluate ***collection ranking*** and the measures have significant weaknesses. In this paper, we survey the measures used in current research and propose a new pair of measures that are based on the concepts of precision and recall. The new measures overcome the problems found in the current measures.

1 Introduction

As hundreds or even thousands of collections become available on the Internet, it is practically impossible to query all of them when searching for information on a given topic: not only does such an exhaust search take a long time to complete, but network resource constraints such as limited bandwidth may also prevent the search from completing. Consequently the need to narrow the search to a few useful collections raises the issue of how to locate the most useful collections. Several approaches[1][3][2][6] have been proposed to solve this problem. All these approaches adopt the same procedure: ranking the collections and returning the top-ranks. Although collection ranking is similar to document ranking, these two problems are not identical, and thus the measures used in document ranking can not be used in collection ranking without modifications. Different measures have been defined to evaluate collection ranking. Some come from the measures used in conventional document retrieval and some do not. In this paper, we survey the measures for collection ranking used in the current research, discuss their strengthes and weaknesses, and propose new measures based on the concepts of precision and recall. In Section 2, we briefly survey the measures used in our previous research[1], NetSerf[2] and GLOSS[4][5]. In Section 3 we define the new measures, which are variations of precision and recall. In the final section, we summerize the discussions.

IPR2024-01306
UIPATH EXHIBIT 1024

2 A Survey of Measures in Collection Ranking Evaluation

2.1 Ranking with INQUERY

In our previous research on collection ranking [1], the mean-square error was used to compare the effectiveness of variations to the basic collection ranking algorithms.

The mean-squared error (MSE) of the collection ranking for a single query is calculated as:

$$\frac{1}{|C|} \cdot \Sigma_{i \epsilon C}(O_i - R_i)^2$$

where:

O_i = optimal rank for collection i, based on the number of relevant documents it contained (the collection with the largest number of relevant documents is ranked 1, the collection with second largest number of relevant documents is ranked 2, and so on),

R_i = the rank for collection i determined by the retrieval algorithm, and

C = the set of collections being ranked.

This measure is easy to understand (an optimal result is 0), but it overemphasizes the absolute ranks of collections. This measure can not distinguish the minor errors caused by mixing up the ranks of the collections containing nearly equal numbers of relevant documents and the major errors caused by mixing up the ranks of collections containing very different numbers of relevant documents.

Example 2.1 *Consider a query q and 5 collections c_1, c_2, c_3, c_4 and c_5. Table 1 shows the optimal rank O and two estimated ranks $Rank_1$ and $Rank_2$ generated by different ranking algorithms.*

O	*Relevant Documents*	$Rank_1$	$Rank_2$
c_1	*20*	c_1	c_2
c_2	*19*	c_3	c_1
c_3	*8*	c_2	c_3
c_4	*1*	c_4	c_4
c_5	*0*	c_5	c_5

Table 1: The optimal and two estimated ranks for Example 2.1

$Rank_1$ and $Rank_2$ have the same mean square error, 0.667. However $Rank_2$ is better than $Rank_3$ in some sense, since $Rank_2$ only mixes up the ranks of c_1 (20 relevant documents) and c_2 (19 relevant documents), which might not be noticeable to users, while $Rank_1$ mixes up c_2 (19 relevant documents) and c_3 (8 relevant documents), which is a noticeable error. Considering the case that the system suggests to search only first two collections, $Rank_2$ can let the user get 39 relevant documents (the same as the optimal rank), and $Rank_3$ can only let the user get

IPR2024-01306
UIPATH EXHIBIT 1024

28 relevant documents (here we assume that the search engine at each individual collection can retrieve all relevant documents).

2.2 NetSerf

NetSerf[2] defines its measure upon a set of queries: Given the set of relevant archives (collections in our context), count the number of queries for which any of the relevant archives were found in the first n top-ranked hits returned by the systems.

For each query in the query set, this measure only uses a binary value to record the contribution of the query: if the query hits any relevant archive in the top n hits, return 1, otherwise return 0. It does not count how many relevant archives have been hit after the first relevant one is hit. All these characteristics make this measure unable to distinguish between the archive with a big number of relevant documents and that with a small number of relevant documents, and thus it ignores many significant errors.

Example 2.2 *Consider a query q and 5 archives a_1, a_2, a_3, a_4 and a_5. Table 2 shows the optimal rank O and three estimated ranks $Rank_2$,$Rank_3$ and $Rank_4$ generated by different ranking algorithms. Table 3 shows the values of the measure in the first n top-ranked hits for this query.*

O	*Relevant Documents*	$Rank_2$	$Rank_3$	$Rank_4$
a_1	*20*	a_2	a_4	a_4
a_2	*19*	a_1	a_3	a_5
a_3	*8*	a_3	a_1	a_1
a_4	*1*	a_4	a_2	a_2
a_5	*0*	a_5	a_5	a_3

Table 2: The optimal and three estimated ranks for Example 2.2

first n hits	$Rank_2$	$Rank_3$	$Rank_4$
1	*1*	*1*	*1*
2	*1*	*1*	*1*
3	*1*	*1*	*1*
4	*1*	*1*	*1*
5	*1*	*1*	*1*

Table 3: The values of the measure for the ranks in Table 2

The measure treats $Rank_2$, $Rank_3$ and $Rank_4$ as equally effective. However $Rank_3$ and $Rank_4$ are much worse rank than $Rank_2$, since $Rank_3$ and $Rank_4$ can not rank the archives with large relevant documents as the top ones. Considering the case that the system suggests to search two archives, $Rank_2$ can retrieve 39 relevant documents, $Rank_3$ can retrieve 9 relevant document, and $Rank_4$ can retrieve 1 relevant document.

IPR2024-01306
UIPATH EXHIBIT 1024

2.3 GLOSS

GLOSS defines its measures based on the well-known precision and recall in conventional document retrieval. GLOSS has two versions: the Boolean version[4] and the vector-space version[5]. The definitions of the measures in these two versions have slight differences. GLOSS uses *databases* to refer to *collections* in our context.

2.3.1 The Boolean Version

In the Boolean version of GLOSS, GLOSS defines $Right(q, DB)$, where q is a query and DB is a set of databases. There are two definitions of $Right(q, DB)$.

- The first definition is that $Right(q, DB)$ is the set of all databases in DB containing at least one document that matches query q.
- The second definition is that $Right(q, DB)$ is the set of those databases containing the most matching documents, more formally,

$$\begin{aligned} Right(q, DB) &= \{db \in DB | RSize(q, db) > 0 \wedge \\ &\quad |\frac{RSize(q, db) - hreal}{hreal}| \leq \delta\} \end{aligned}$$

  where $RSize(q, db)$ is the number of documents that match q in database db, and $hreal = max_{db \in DB} RSize(q, db)$.

The first definition is a special case of the second definition (by setting δ big enough).

GLOSS defines the measures R_{Right} and P_{Right} as:

$$R_{Right}(DB) = \frac{|Chosen \cap Right|}{|Right|}$$

$$P_{Right}(DB) = \frac{|Chosen \cap Right|}{|Chosen|}.$$

$R_{Right}(DB)$ is the fraction of the $Right$ databases that are selected and $P_{Right}(DB)$ is the fraction of selected databases that are $Right$ ones.

These measures treat the databases in $Right(q, DB)$ as equally important. If the number of relevant documents in each database varies significantly, these measures can not distinguish between the databases with large numbers of relevant documents and those with small numbers.

Example 2.3.1 *Consider a query q and 5 databases db_1, db_2, db_3, db_4 and db_5. Table 4 shows the optimal rank O and two estimated ranks $Rank_2$ and $Rank_3$ generated by different GLOSS estimators. Table 5 shows the values of the measures in the top n databases for this query.*

IPR2024-01306
UIPATH EXHIBIT 1024

O	*Relevant Documents*	$Rank_2$	$Rank_3$
db_1	*20*	db_2	db_4
db_2	*19*	db_1	db_3
db_3	*8*	db_3	db_1
db_4	*1*	db_4	db_2
db_5	*0*	db_5	db_5

Table 4: The optimal and two estimated ranks for Example 2.3.1

	R_{right}		P_{right}	
Top n dbs	$Rank_2$	$Rank_3$	$Rank_2$	$Rank_3$
1	*25%*	*25%*	*100%*	*100%*
2	*50%*	*50%*	*100%*	*100%*
3	*75%*	*75%*	*100%*	*100%*
4	*100%*	*100%*	*100%*	*100%*
5	*100%*	*100%*	*80%*	*80%*

Table 5: The values of the measures for the ranks in Table 4.

Table 5 shows that $Rank_2$ and $Rank_3$ are equally effective, but actually $Rank_3$ is a much worse one (considering the example in Example 2.2).

2.3.2 The Vector-Space Version

In the vector-space version of GLOSS (called gGLOSS in GLOSS terminology), the optimal rank is determined by sorting the databases according to their goodness $Goodness(l, q, db)$ for a given query q. $Goodness(l, q, db)$ is defined as:

$$Goodness(l, q, db) = \Sigma_{d \in Rank(l,q,db)} sim(q, d),$$

where $sim(q, d)$ is the similarity between query q and document d, and $Rank(l, q, db) = \{d \in db | sim(q, d) > l\}$.

It is noteworthy that gGLOSS does not use the relevance information in the definition of $Goodness(l, q, db)$. This definition does not offer benefit than using relevant information directly, since $Goodness(l, q, db)$ is highly correlated with relevant information ([5] gives some discussions for this).

GLOSS defines measures R_n and P_n for a given query as:

$$R_n = \frac{g_n}{i_n}$$

where:
i_n is the accumulated goodness of the top n databases in the ideal rank,
g_n is the accumulated goodness of the top n databases in the rank that gGloss generates.

IPR2024-01306
UIPATH EXHIBIT 1024

$$P_n = \frac{|\{db \in Top_n(G) | Goodness(l, q, db) > 0\}|}{|Top_n(G)|}$$

where G is the rank gGLOSS generates.

The measure P_n suffers the same problem as P_{right} in the Boolean version. The measure R_n behaves much better than R_{right}.

Example 2.3.2 *Consider a query q and 5 databases db_1, db_2, db_3, db_4 and db_5. Table 6 shows the optimal rank O and two estimated ranks $Rank_2$ and $Rank_3$ generated by different gGLOSS estimators. Table 5 shows the values of the measures in the top n databases for this query.*

O	$Goodness$	$Rank_2$	$Rank_3$
db_1	0.9	db_2	db_4
db_2	0.6	db_1	db_3
db_3	0.3	db_3	db_1
db_4	0.2	db_4	db_2
db_5	0.0	db_5	db_5

Table 6: The optimal and two estimated ranks for Example 2.3.2

	R_n		P_n	
Top n dbs	$Rank_2$	$Rank_3$	$Rank_2$	$Rank_3$
1	66.7%	22.2%	100%	100%
2	100%	33.3%	100%	100%
3	100%	77.8%	100%	100%
4	100%	100%	100%	100%
5	100%	100%	80%	80%

Table 7: The values of the measures for the ranks in Table 6.

The measure R_n shows that $Rank_2$ is better than $Rank_3$ (compared with R_{right} in the Boolean version, this measure is better), but the measure P_n does not (the same as P_{right}).

3 New Measures

By analyzing the current measures, it is easy to see that the mean square error is in the sense of "too strict", because it does not distinguish between minor errors and major errors. The measures developed for NetSerf and GLOSS are "too loose", because they ignore many errors. New measures need to be defined to overcome these problems.

Like GLOSS, our new measures are based on the concepts of precision and recall. But we modify the conventional precision and recall by considering the numbers of the relevant documents in each collection.

IPR2024-01306
UIPATH EXHIBIT 1024

As in our previous research[1], the optimal rank is determined by sorting collections according to the numbers of relevant documents they contain. The new measures are defined as:

For a given query and top n collections,

$$R_n = \frac{\Sigma_{i=1}^{n} r_i}{Rel}$$

$$P_n = \frac{\Sigma_{i=1}^{n} r_i}{n}$$

where:

r_i is the number of relevant documents in collection c_i,
c_i is the ith collection in the rank generated by a ranking algorithm;
Rel is the total number of relevant documents in all collections; and
n is the number of collections ranked.

R_n indicates the fraction of the relevant documents that can be retrieved when searching the top n collections. P_n indicates the average number of the relevant documents that each collection contains when searching the top n collections.

The advantages of these measures are that collections with many relevant documents can be distinguished from those with few relevant documents, and that the minor errors of ranks can also be distinguished from the major errors. The disadvantage is that the concept of P_n deviates from traditional concept of precision.

Example 3 *Consider a query q, 5 collections c_1, c_2, c_3, c_4 and c_5, and all rank examples used in previous examples (Table 8). Table 9 shows the values of R_n in the top n collections for a given query. Table 10 shows the values of P_n in the top n collections.*

O	*Relevant Documents*	$Rank_1$	$Rank_2$	$Rank_3$	$Rank_4$
c_1	*20*	c_1	c_2	c_4	c_4
c_2	*19*	c_3	c_1	c_3	c_5
c_3	*8*	c_2	c_3	c_1	c_1
c_4	*1*	c_4	c_4	c_2	c_2
c_5	*0*	c_5	c_5	c_5	c_3

Table 8: The optimal and the estimated ranks for Example 3

	R_n				
Top n colls	*Optimal*	$Rank_1$	$Rank_2$	$Rank_3$	$Rank_4$
1	*41.7%*	*41.7%*	*39.6%*	*2.1%*	*2.1%*
2	*81.3%*	*58.3%*	*81.3%*	*18.7%*	*2.1%*
3	*97.9%*	*97.9%*	*97.9%*	*60.4%*	*43.8%*
4	*100%*	*100%*	*100%*	*100%*	*83.3%*
5	*100%*	*100%*	*100%*	*100%*	*100%*

Table 9: The values of R_n for the ranks in Table 8.

IPR2024-01306
UIPATH EXHIBIT 1024

	P_n				
Top n colls	*Optimal*	$Rank_1$	$Rank_2$	$Rank_3$	$Rank_4$
1	*20*	*20*	*19*	*1*	*1*
2	*19.5*	*14*	*19.5*	*4.5*	*0.5*
3	*15.7*	*15.7*	*15.7*	*9.7*	*7*
4	*12*	*12*	*12*	*12*	*10*
5	*9.6*	*9.6*	*9.6*	*9.6*	*9.6*

Table 10: The values of P_n for the ranks in Table 8.

Considering $Rank_1$, the data on the second and the third rows in Tables 9 and 10 indicate that $Rank_1$ makes some big mistakes between rank 2 and rank 3. Considering $Rank_2$, the mistakes between rank 1 and rank 2 are minor. Considering $Rank_2$, $Rank_3$ and $Rank_4$, it is easy to see that $Rank_2$ is much better than $Rank_3$ and $Rank_4$. Tables 9 and 10 show that the new measures make it possible to distinguish between major errors and minor errors.

4 Conclusion

This paper surveys the collection ranking measures used in current research. The measures used in NetSerf and GLOSS do not consider information about the relevant documents in each collection, which causes them to ignore significant errors in the rankings. The mean square error overemphasizes the absolute ranks of each collection, which make it count some minor errors as major errors. By analyzing the mistakes that the current measures made, a new pair of measures are proposed that are based on the concepts of precision and recall, and that include the information about relevant documents in each collection. The examples show that the new measures overcome the problems found in the other measures.

Acknowledgements

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

References

[1] James P. Callan, Zhihong Lu and W. Bruce Croft. Searching Distributed Collections with Inference Networks. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 21-28, Seattle, July 1995. Association for Computing Machinery.

IPR2024-01306
UIPATH EXHIBIT 1024

[2] NetSerf: Using Semantic Knowledge to Find Internet Information Archives. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 4-11, Seattle, July 1995. Association for Computing Machinery.

[3] Glossary of Servers *http://gloss.stanford.edu*

[4] Luis Gravano and Hector Garcia-Molina. Precision and Recall of GLOSS Estimator for Database Discovery. *Stanford Univerdity Technical Note Number STAN-CS-TN-94-10.*

[5] Luis Gravano and Hector Garcia-Molina. Generalizing GLOSS to Vector-Space Databases and Broker Hierarchies. *Proceedings of the 21st VLDB Conference*, Zurich, Switchland 1995.

[6] WAIS 2.0: Technical Description *http://www.wais.com/newhomepages/techtalk.html*

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT F

Captured at: 2024/08/15 11:25 AM URL: https://web.archive.org/web/20040405195344/http:/www.cs.virginia.edu:80/

INTERNET ARCHIVE WaybackMachine http://www.cs.virginia.edu/ Go JAN APR JUN 05 2003 2004 2005 About this capture
1,576 captures
31 Dec 1996 - 15 Aug 2024

UNIVERSITY of VIRGINIA

Computer Science

Research | Teaching | People | Community

Of interest to: Prospective Students, Members

Search
Directory
Contact Us

Search
Advanced Search - Site Directory

Research
Projects, Publications, Technical Reports [CS TRs], Computing Facilities, Awards

People
Undergraduate Students, CLAS Program Students, Graduate Students, Graduating PhDs, Faculty, New Faculty, Staff, Alumni, Faculty and Staff Directory, Web Team, Awards
Chair: Jack Stankovic

Teaching
Degrees: CS Major (Curriculum), Computer Engineering Major, CS Minor, Graduate
CLAS CS Program — Applications Due 6 April 2004

Courses: Spring 04 [COD], Summer 04 [COD], Fall 04 [COD], Previous Years, Course Registration Requirements, Course Waiting Lists, Prerequisite Policy
Teaching Resource Center, Anonymous Feedback

Community
Why us?, Colloquia, News, ACM Student Chapter, Graduate Student Group, Graduate Orientation, Professional Societies, Corporate Partners, Happy Hour, Lounge, Tea Time

MRRL: Analysis tools for Memory Reference Reuse Latency allow sampled architecture simulation to defeat cold-start bias by rigourously identifying the minimum warmup duration. More

CS854: Microprocessor Survey

More Highlights

SEAS Open House [more pictures]

Department Chair Search

General Faculty Search

Apply for **Undergraduate Teaching Assistant Positions**

Upcoming Events
- Tomorrow (6 April), 2:00pm - Dr. Paul Davis, RAND / RAND Graduate School. *Rethinking Families of Models and Games for New Circumstances*

More Colloquia - UVa Events Calendar

News
- 28 Mar - High-speed network may boost region's economy (Marty Humphrey) [Associated Press]
- 05 Mar - What's Good About Computer Viruses (David Evans) [TechNewsWorld]
- 03 Nov - Students to Face Off in 'Battle of the Brains' at U.Va. (David Brogan) [U.Va. Top News Daily]

Previous News - UVa News

Recent Awards
- Jim Van Dyke won the 2004 Virginia Engineering Foundation Outstanding Student Award
- Arsalan Tavakoli-Shiraji won a Goldwater Scholarship
- David Chu was awarded an NSF Graduate Fellowship.
- Three UVa students have won CRA Outstanding Undergraduate Honorable Mentions: **David Chu**, **Kristen Neal** and **James Van Dyke**.
- **Siva Velusamy and Karthik Sankaranarayanan** won the ISCA best student paper award (for a paper co-authored with Kevin Skadron, Mircea Stan, Wei Huang and David Tarjan). [Paper] (10 June 2003)
- **Tian He, John Stankovic, Chenyang Lu and Tarek Abdelzaher** were nominated for the ICDCS best paper award (29 April 2003)

More Awards

Announcements
- ACM Tutoring

Reference
Address, Area Maps, Office Locator, Libraries, Links to Universities, Miscellaneous Links

Help
Computer Support, HelpNet, Web Team, Web Tutorials

Fun
Computer Museum, Happy Hour, Quotes, Oracle of Bacon

Documents
Brochure, Message from Our Chair, Posters, Newsletter, Department Overview

Pictures
Classic Moments, Graduation, Photo Collections

Join Us
Admissions, Applications, Job Listings, Why UVA CS?

Surroundings
School of Engineering, University of Virginia, Charlottesville, Virginia

Vision
New Building, Strategic Plan, Promotion & Tenure

For Members
Login, Newsgroups, Reservations, Course Feedback, Course Waiting Lists, Logos, Mail

Department of Computer Science
School of Engineering, University of Virginia
151 Engineer's Way, P.O. Box 400740
Charlottesville, Virginia 22904-4740
(434) 982-2200 Fax: (434) 982-2214

Web Comments: *webteam@cs.virginia.edu*
Admissions Inquiries: *inquiry@cs.virginia.edu*
Site directory, Other addresses
Server statistics
© Created by the CS Web Team

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT G

Captured at: 2024/08/15 11:28 AM URL: https://web.archive.org/web/20040401140136/http:/www.cs.virginia.edu/research/

INTERNET ARCHIVE Wayback Machine
http://www.cs.virginia.edu/research/ Go
321 captures
5 Jun 1997 - 4 Dec 2023
FEB APR JUN
01
2003 2004 2005
About this capture

Research

Overview - Project Descriptions - List of Projects by Principal Investigator
Sponsors - External Funding - Resources - Technical Reports and Dissertations - Honors and Awards

Project Home Pages

Architecture

LAVA - *Laboratory for Computer Architecture at Virginia*
VLSI CAD - *Physical Design and Layout*

Artificial Intelligence

Computer Vision
Programming the Swarm

Computational Biology

Computational Biology

Dependability

Survivability Achitechtures
Programming the Swarm
Zeus - *Practical Formal Techniques for Software Developement*
Dependability Research Group

Distributed Computing

Grid Computing Group
Isotach - *Concurrency control without locks or barriers*
Legion - *World-Wide Virtual Computer*

Electronic Commerce

Internet Technology Innovation Center
SurveySuite

Embedded Systems

Component Based OSs for Embedded Systems
VEST - *Virginia Embedded Systems Toolkit*

Graphics

Chromium - *Scalable Interactive Graphics with Commodity Technology*
Dynamically Simulated Characters in Graphical Environments
Perceptually-Driven Rendering Using Gaze-Based Simplification
Scanning Monticello
VLSIR - *Very Large Scale Interactive Rendering*
VDSLib - *Free Software for View-Dependent Simplification*

Information Management

Networking

Isotach - *Concurrency control without locks or barriers*
Multimedia Networks Group (MNG)
Web QoS Group

Programming Languages

Naccio - *Policy-Directed Code Safety*
Programming the Swarm
Zephyr - *National Compiler Infrastructure*

Real-Time Systems

BeeHive - *Global Multimedia Database Support for Dependable, Real-Time Application*
Feedback Control Real-Time Scheduling
HoLSt - *Hierarchical Loadable Schedulers*
VEST - *Virginia Embedded Systems Toolkit*
Web QoS Group

Security

Intrusion Detection
Naccio - *Policy-Directed Code Safety*
Physicrypt - *Physical Cryptography and Security*
Splint - *Secure Programming Lint*
SurvIS - *Center for Survivable Information Systems*
Survivability Architectures

Software Engineering

Galileo - *An Advanced Fault Tree Analysis Tool*
HoLSt - *Hierarchical Loadable Schedulers*
Programming the Swarm
Splint - *Annotation-Assisted Lightweight Static Checking*
Strategic Software Engineering
Survivability Architectures
Zeus - *Practical Formal Techniques for Software Developement*

Theory

Computational Geometry

IPR2024-01306
UIPATH EXHIBIT 1024

Department of Computer Science
School of Engineering, University of Virginia
151 Engineer's Way, P.O. Box 400740
Charlottesville, Virginia 22904-4740
(434) 982-2200 Fax: (434) 982-2214

Web Comments: *webteam@cs.virginia.edu*
Admissions Inquiries: *inquiry@cs.virginia.edu*
Site directory, Other addresses
Server statistics
© *Created by the CS Web Team*

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT H

Captured at: 2024/08/15 11:29 AM URL: https://web.archive.org/web/20040505070445/http:/www.cs.virginia.edu/~cyberia/

INTERNET ARCHIVE WaybackMachine
http://www.cs.virginia.edu/~cyberia/ Go
190 captures
9 Oct 1999 - 18 Apr 2024
MAR 2003 | MAY 05 2004 | JUN 2005
About this capture

Cyberia

The University of Virginia Information Retrieval Group

Cyberia Projects

CBIR

Extending the Vocabulary for Cross-Disciplinary Searching of EOS Data

Personal Information Environments

Testbed

Multiple Viewpoints

Publications

Papers
Posters
Presentations

The Group

People
Contact

A major part of our work is research in distributed Information Retrieval. We are developing new techniques aimed at constructing a user-customized Personalized Information Environment (PIE) that can be tailored to the needs of specific users.

We have recently constructed a testbed for evaluating distributed information retrieval approaches and are currently working to compare different approaches using this testbed.

When dealing with distributed collections, different collections may employ different conventions for naming entities in the collection. This can adversely affect retrieval performance by obscuring similarities among those entities. This problem can be lessened by the use of authority files which list accepted variants of names; however, the construction of these authority files is labor-intensive. We have been studying the potential for partially automating this process, as in the EVOC Project.

We are also investigating the use of multiple viewpoints of a document collection during searching or browsing to facilitate the location of useful documents. Our research in this area is focused on specific techniques for interacting with document collections. We are investigating the representations, framework and processes required to support these interactions.

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT I

Captured at: 2024/08/15 11:31 AM URL: https://web.archive.org/web/20040625112640/http:/www.cs.virginia.edu/~cyberia/papers.html

INTERNET ARCHIVE WaybackMachine http://www.cs.virginia.edu/~cyberia/papers.html Go DEC JUN AUG 25 2003 2004 2005 About this capture
60 captures
9 Oct 1999 - 3 Dec 2016

Papers

The University of Virginia Information Retrieval Group

Presentations | Posters
Journal Articles | Articles & Talks | Conference Papers
Unrefereed Papers | Graduate Works | Undergraduate Theses

Journal Articles

top

French, J. C., A. C. Chapin, W. N. Martin
"Multiple Viewpoints as an Approach to Digital Library Interfaces," Journal of the American Society for Information Science and Technology (to appear).

Powell, A. L. and J. C. French
"Comparing the Performance of Collection Selection Algorithms,"" ACM Transactions on Information Systems, vol. 21, no. 4, October 2003, pp. 412-456
PDF

French, J. C., A. L. Powell, F. Gey, N. Perelman
"Exploiting Manual Indexing to Improve Collection Selection and Retrieval Effectiveness" Information Retrieval, vol. 5, no. 4, October 2002, pp. 323-351.
PDF

French, J. C. and A. L. Powell
"Metrics for Evaluating Database Selection Techniques," World Wide Web, Vol. 3, no. 3, November 2000, pp.153-163
postscript

French, J. C., A. L. Powell, E. Schulman
"Using Clustering Strategies for Creating Authority Files" Journal of the American Society for Information Science
PS version

French, J. C., C. L. Viles
" Personalized Information Environments: An Architecture for Customizable Access to Distributed Digital Libraries " D-Lib Magazine, vol. 5 no. 6 (June 1999).

Viles, C. L. and J. C. French
"Content Locality in Distributed Digital Libraries," Information Processing and Management, vol 35, no. 3 (May 1999), pp 317-336.

Schulman, E., J. C. French, A. L. Powell, G. Eichhorn, M. J. Kurtz and S. S. Murray
"Trends in Astronomical Publication between 1975 and 1996" Publications of the Astronomical Society of the Pacific, vol. 109 (Nov. 1997), pp. 1278-1284.

French, J. C., D. E. Brown, and N. H. Kim
"A Classification Approach to Boolean Query Reformulation" Journal of the American Society for Information Science, vol. 48, no. 8 (August 1997), pp. 694-706.

Kim, N. H., J. C. French, and D. E. Brown
"Development of New Retrieval Performance Measures for Query Reformulation Algorithms" Transactions of the Korean Information Processing Society, vol. 4, no. 4 (April 1997), pp. 963-972.

Kim, N. H., D. E. Brown, and J. C. French
"Time Complexity Analysis of Boolean Query Formulation Algorithms" Transactions of the Korean Information Processing Society, vol. 4, no. 3 (March 1997), pp. 709-719.

French, J. C., J. C. Knight, and A. L. Powell
"Applying Hypertext Structures to Software Documentation" Information Processing and Management, vol. 33, no. 2 (March 1997), pp. 219-231.
PS version
PDF version

French, J. C. and C. L. Viles

IPR2024-01306
UIPATH EXHIBIT 1024

French, J. C. and C. L. Viles
"Ensuring Retrieval Effectiveness in Distributed Digital Libraries" ournal of Visual Communication and Image Representation, vol. 7, no. 1 (March 1996), pp. 61-73.

French, J. C., E. Fox, K. Maly, A. Selman
"Wide Area Technical Report Service: Technical Reports Online" Communications of the ACM, vol. 38, no. 4 (April 1995), pp. 45.
via ACM DL

Viles, C. L. and J. C. French
"Availability and Latency of World Wide Web Information Servers" Computing Systems, vol. 8, no. 1 (1995), pp. 61-91.
PS version

Heyman, B. L., J. C. French, and G. L. Bull
"The Library Access, Search and Retrieval Pilot Project for the Global Change Data and Information System" Library Hi Tech, issue 49-50, vol. 13 no. 1-2 (1995), pp. 19-21.

French, J. C.
"DIRE: An Approach to Improving Scientific Communication" Information and Decision Technologies, vol. 19 (1994), pp. 527-541.

Conference Papers

top

Jin, X. and J. C. French
"Improving Image Retrieval Effectiveness via Multiple Queries," First ACM International Workshop on Multimedia Databases (ACM-MMDB 2003), New Orleans, Louisiana, November 7, 2003, pp. 86-93.
PDF

French, J. C., J. V. S. Watson, X. Jin and W. N. Martin
"An Exogenous Approach for Adding Multiple Image Representations to Content-Based Image Retrieval Systems", Seventh International Symposium on Signal Processing and its Applications (ISSPA 2003), Paris, July 1-4, 2003.
PDF

French, J. C, A. C. Chapin and W. N. Martin
"Multiple Viewpoints: A Strategy for Searching Multimedia Content," Workshop on Multimedia Content in Digital Libraries Chania, Crete, Greece, June 2-3, 2003.

French, J. C., A. C. Chapin and W. N. Martin,
"An Application of Multiple Viewpoints to Content-Based Image Retrieval," Joint Conference on Digital Libraries (JCDL 2003), Houston, Texas, May 27-31, 2003.
PDF

French, J. C., J. V. S. Watson, X. Jin and W. N. Martin,
"Integrating Multiple Multi-Channel CBIR Systems," International Workshop on Multimedia Information Systems (MIS 2003), Ischia, Italy, May 26-28, 2003, pp. 85-95.
PDF

French, JC
"Modeling Web Data," Joint Conference on Digital Libraries (JCDL'02), Portland, OR, July, 2002, pp. 320-321.
PDF

Monroe, G. A., J. C. French and A. L. Powell,
"Obtaining Language Models of Web Collections Using Query-Based Sampling Techniques," Hawaii International Conference on System Sciences (HICSS-35): Data Mining and Information Retrieval, Jan. 2002 (to appear).
PDF version

Hauver, D. B. and J. C. French,
"Flycasting: Using Collaborative Filtering to Generate a Playlist for Online Radio," International Conference on Web Delivery of Music, Florence, Italy, Nov. 2001 (to appear).
PS

French, J. C., A. L. Powell, F. Gey, N. Perelman,
"Exploiting A Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness," Tenth International Conference on Information and Knowledge Management (CIKM 2001), Nov. 2001 (to appear).
PS version

French, J. C. and D. B. Hauver,
"Flycasting: On the Fly Broadcasting," Joint DELOS-NSF International Workshop on Personalization and Recommender Systems in Digital Libraries, Dublin, Ireland, June 18-20, 2001, pp. 89-93.
PDF version

Powell, A.L., J. C. French, J. Callan, M. Connell, and C. Viles
"The Impact of Database Selection on Distributed Searching", 23rd ACM Confrence on Information Retrieval, SIGIR 2000.
PS version

Powell, A.L. and J. C. French
"Growth and Server Availability of the NCSTRL Digital Library", Fifth ACM Conference on Digital Libraries, 2000

IPR2024-01306
UIPATH EXHIBIT 1024

"Growth and Server Availability of the NCSTRL Digital Library", Fifth ACM Conference on Digital Libraries, DL 2000.
PS version
PDF version

Dushay, N., J. C. French and C. Lagoze
"Predicting Indexer Performance in a Distributed Digital Library", European Conference on Digital Libraries, Paris, France, September 1999.
PS version

French, J. C. and A. L. Powell,
"Metrics for Evaluating Database Selection Techniques," Proc. Tenth International Workshop on Database and Expert Systems Applications (DEXA'99), Florence, Italy, Sept. 1999, pp. 726-730.
PS version
PDF version

Dushay, N., J. C. French and C. Lagoze
"Using Query Mediators for Distributed Searching in Federated Digital Libraries", Proc. Fourth ACM Conference on Digital Libraries, Berkeley, CA, August 1999, pp. 171-178.
PDF version

Sharretts, C. and J. C. French
"Electronic Theses and Dissertations at the University of Virginia Library," Proc. Fourth ACM Conference on Digital Libraries, Berkley, CA, August 1999, pp. 246-247.
PS version
PDF version

French, J. C., A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey, Y. Mou
"Compairing the Performance of Database Selection Alogrithms" Proc. 22st International Conference on Research and Development in Information Retrieval (SIGIR'99), Berkley, CA, August 1999, pp. 238-245.
PS version
PDF version

Ganti, V., R. Ramakrishnan, J. Gehrke, A. Powell, J. French
"Clustering Large Datasets in Arbitrary Metric Spaces" Proc. 15th International Conference on Data Engineering (ICDE'99), Sydney, Australia, March 1999
PS version
PDF version

French, J. C., A. L. Powell, C. L. Viles, T. Emmitt and K. Prey
"Evaluating Database Selection Techniques: A Testbed and Experiment" Proc. 21st International Conference on Research and Development in Information Retrieval (SIGIR'98), Melbourne, Australia, August 1998.
PS version
PDF version

French, J. C., A. L. Powell, and W. R. Creighton III
"Efficient Searching in Distributed Digital Libraries" Proc. 3rd ACM International Conference on Digital Libraries (DL'98), Pittsburgh, PA, June 23-26, 1998.
PS version
PDF version

French, J. C., A. L. Powell and E. Schulman
"Applications of Approximate Word Matching in Information Retrieval," Proc. 6th International Conference on Information and Knowledge Management (CIKM'97), Las Vegas, Nevada, Nov. 1997, pp. 9-15.
PS version
PDF version
HTML Slide Show of Presentations

French, J. C., A. L. Powell, E. Schulman and J. L. Pfaltz
"Automating the Construction of Authority Files in Digital Libraries: A Case Study," Proc. First European Conference on Research and Advanced Technology forDigital Libraries (ECDL'97), Lecture Notes in Computer Science 1324, C. Peters and C. Thanos (eds.), Springer-Verlag, Pisa, 1-3 September 1997, pp. 55-71.
PS version
PDF version
HTML Slide Show of Presentation

Powell, A. L., J. C. French and J. C. Knight
"A Systematic Approach to Creating and Maintaining Software Documentation"
Proc. 1996 ACM Symposium on Applied Computing (SAC 96), Philadelphia, PA, Feb. 1996, pp. 201-208.
PS version
PDF version

Barros, J. E., J. C. French, W. N. Martin, P. Kelly, and M. Cannon
"Using the Triangle Inequality to Reduce the Number of Comparisons Required for Similarity-Based Retrieval" Electronic Imaging '96, Storage and Retrieval for Image and Video Databases IV, San Jose, CA, (Feb. 1996), 392-403.

Viles, C. L. and J. C. French
"On the Update of Term Weights in Dynamic Information Retrieval Systems" Proc. 4th International Conference on Information and Knowledge Management (CIKM'95), Baltimore, Maryland, Nov. 1995, pp. 167-174.
via ACM DL

Barros, J. E., J. C. French, and W. N. Martin
"Trading Efficiency for Effectiveness in Similarity-Based Indexing for Image Databases" in Digital Image Storage and Archiving Systems, C.-C. Jay Kuo, Editor, Proc. SPIE 2606 (1995), pp. 276-287.

Viles, C. L. and J. C. French
"Dissemination of Collection Wide Information in a Distributed Information Retrieval System" Proc. 18th International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, July 1995, pp. 12-20.
PS version

Barros, J. E., J. C. French, W. N. Martin, and P. Kelly
"System for Indexing Multispectral Satellite Images for Efficient Content-Based Retrieval" in Storage and Retrieval for Image and Video Databases III, W. Niblack, R. C. Jain, editors, Proc. SPIE 2420, Feb. 1995, pp. 228-237.
PS version

Maly, K., J. C. French, A. Selman, E. A. Fox
"Wide Area Technical Report Service" Proc. 2nd International World Wide Web Conference, WWW'94 Oct. 1994, pp. 523-533.

Barros, J. E., J. C. French, W. N. Martin, P. Kelly, and J. M. White
"Indexing Image Databases for Content Based Retrieval" in Image and Information Systems: Applications and Opportunities (23rd AIPR Workshop), Proc. SPIE 2368, Oct. 1994, pp. 25-36.
PS version

Kim, N. H., J. C. French and D. E. Brown
"Boolean Query Reformulation with the Query Tree Classifier" in Advances in Classification Research, vol. 4, Learned Information, 1994, pp. 123-138.

French, J. C.
"An Environment for Passively Sharing Information" Proc. International Space Year Conference on Earth and Space Science Information Systems, Feb. 1992, pp 632-637.

Invited Articles and Talks

top

Powell, A. L. and J. C. French
"Using Multiple Views of a Document Collection in Information Exploration" CHI'98 Workshop on Innovation and Evaluation in Information Exploration Interfaces, Los Angeles, CA, April 19, 1998.

Powell, A. L. and J. C. French
"Using Topic Information to Improve Access to Scientific Document Collections" SIGIR Workshop on Summarization and Visualization for IR: Reducing the Information Overload, Philadelphia, PA, July 31, 1997.

French, J. C.
"Database Support for Correlation and Fusion Algorithms" Workshop on Data Mapping and Matching, NASA CESDIS, Goddard Space Flight Center, Greenbelt, MD, Nov. 7, 1996.

French, J. C.
"Networked Computer Science Technical Reports Library" SURA Workshop on Electronic Theses and Dissertations, Charlotte, NC, August 1-2, 1996.

French, J. C.
"Networked Computer Science Technical Reports Library" Dagstuhl Workshop on Networked Information Systems: Discovery, Retrieval, Dissemination, Saarbrucken, Germany, Feb. 26 - Mar. 1, 1996.

French, J. C.
"NCSTRL: Preliminary Evaluation" NCSTRL Technical Review, CNRI, Reston, VA, March 20, 1996.

French, J. C.
"Library Access, Search, and Retrieval (LASR) Pilot" GCDIS Subgroup of the CENR Task Force on Observations and Data Management, Data Management Working Group of the U.S. Global Change Research Program, Washington, DC, March 14, 1996.

French, J. C.
"Library Access, Search, and Retrieval (LASR) Pilot" Data Management Working Group of the U.S. Global Change Research Program, Washington, DC, March 14, 1996.

French, J. C. and C. Lagoze
"Networked Computer Science Technical Reports Library" Dagstuhl Workshop on Networked Information Systems: Discovery, Retrieval, Dissemination, Saarbrucken, Germany, Feb. 26 - Mar. 1, 1996.

Lagoze, C. and J. C. French
"Networked Computer Science Technical Reports Library" ACM Computing Week, Philadelphia, PA, Feb. 16, 1996.

French, J. C.
"Networked Computer Science Technical Reports Library" CRA Board of Directors, Washington, DC, Dec. 1995.

IPR2024-01306
UIPATH EXHIBIT 1024

French, J. C.
"Library Access, Search, and Retrieval (LASR) Pilot" Global Climate Data and Information System (GCDIS) Data Center Workshop, Oak Ridge, TN, Sept. 12-14, 1995.

Unrefereed Publications

top

Prey, Kevin; J. C. French; Allison L. Powell; Charles L. Viles
"Inverse Document Frequency and Web Search Engines" Technical Report CS-2001-07, February 5, 2001.
PS version

O'Neil, Edward K.; J. C. French
"A Description of the LAMBWeb-Derived Language Model Builder," Technical Report CS-2000-31, May 16, 2000.
PS version

Monroe, Gary A.; David R. Mikesell; J. C. French
"Determining Stopping Criteria in the Generation of Web-Derived Language Models," Technical Report CS-2000-30, May 10, 2000.
PS version

Srinivasa, R., T. Phan, N. Mohanraj, A. L. Powell and J. C. French,
"Database Selection Using Document and Collection Term Frequencies," Technical Report CS-2000-32, Dept. of Computer Science, University of Virginia, May 2000.

Callan, J., A. L. Powell, J. C. French and M. Connell,
"The Effects of Query-Based Sampling on Automatic Database Selection Algorithms," Technical Report CMU-LTI-00-162, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2000.

Dushay, N., J. C. French and C. Lagoze,
"Predicting Indexer Performance in a Distributed Digital Library," Technical Report TR99-1743, Dept. of Computer Science," Cornell University, May 1999.
HTML

French, J. C., A. L. Powell and J. Callan,
"Effective and Efficient Automatic Database Selection," Technical Report CS-99-08, Dept. of Computer Science, University of Virginia, February 22, 1999.
Postscript

Dushay, N., J. C. French and C. Lagoze,
"Using Query Mediators for Distributed Searching in Federated Digital Libraries," Computing Research Repository (xxx.cs.DL/9902020), February 1, 1999.
HTML

French, J. C., A. L. Powell, J. Callan, C. L. Viles, T. Emmitt, K. J. Prey and Y. Mou,
"Comparing the Performance of Database Selection Algorithms," Technical Report CS-99-03, Dept. of Computer Science, University of Virginia, January 12, 1999.
postscript
PDF

Dushay, N., J. C. French and C. Lagoze,
"A Characterization Study of NCSTRL Distributed Searching," Technical Report TR99-1725, Dept. of Computer Science, Cornell University, January 1999.
HTML

Powell, A. L. and J. C. French,
"The Potential to Improve Retrieval Effectiveness with Multiple Viewpoints," Technical Report CS-98-15, Dept. of Computer Science, University of Virginia, May 14, 1998.
postscript.Z

Natrajan, A., A. L. Powell, and J. C. French
"Using N-grams to Process Hindi Queries with Transliteration Variations" Technical Report CS-97-17, Dept. of Computer Science, Univ. of Virginia, July 1997.
PS version

French, J. C., A. L. Powell, and E. Schulman
"Automating the Construction of Authority Files in Digital Libraries: A Case Study" Technical Report CS-97-02, Dept. of Computer Science, Univ. of Virginia, Jan. 1997.
PS version

French, J. C., A. L. Powell, and E. Schulman
"Applications of Approximate Word Matching in Information Retrieval" Technical Report CS-97-01, Dept. of Computer Science, Univ. of Virginia, Jan. 1997.
PS version

Schulman, E., J. C. French, A. L. Powell, S. S. Murray, G. Eichhorn, and M. J. Kurtz
"The Sociology of Astronomical Publication Using ADS and ADAMS" in Astronomical Data Analysis Software and Systems VI, edited by G. C. Hunt and H. E. Payne, A.S.P. Conference Series, Vol. 1[illegible] 364.

IPR2024-01306
UIPATH EXHIBIT 1024

Schulman, E., A. L. Powell, J. C. French, G. Eichhorn, M. J. Kurtz, and S. S. Murray
"Using the ADS Database to Study Trends in Astronomical Publication" Bulletin of the American Astronomical Society, vol. 28, 4 (1996), pp. 1281.

French, J. C. and C. L. Viles
"Exploiting Coauthorship to Infer Topicality in a Digital Library of Computer Science Technical Reports" Technical Report CS-96-20, Dept. of Computer Science, Univ. of Virginia, Dec. 1996.
PS version

Viles, C. L. and J. C. French
"TREC4 Experiments using DRIFT" 4th Text Retrieval Conference, Gaithersburg, MD, NIST Special Publication 500-236 (Oct. 1996), pp. 549-557.

French, J. C., G. L. Bull, M. Tarrant, and A. L. Powell
"Library Access, Search and Retrieval (LASR) Pilot (Final Report)" Technical Report CS-96-07, Dept. of Computer Science, Univ. of Virginia, Mar. 1996.
PS version

French, J. C., J. C. Knight, and A. L. Powell
"Hypertext Structures and Software Documentation" Technical Report CS-96-04, Dept. of Computer Science, Univ. of Virginia, Feb. 1996.
PS version

Barros, J., J. C. French, W. Martin, and P. Kelly
"System for Indexing Multi-spectral Satellite Images for Efficient Content-based Retrieval" Technical Report CS-95-06, Dept. of Computer Science, Univ. of Virginia, Jan. 1995.
PS version

Viles, C. L. and J. C. French
"Dissemination of Collection Wide Information in a Distributed Information Retrieval System" Technical Report CS-95-02, Dept. of Computer Science, Univ. of Virginia, Jan. 1995.
PS version

Barros, J. E., J. C. French, W. N. Martin, P. Kelly, J. M. White
"Indexing Multispectral Images for Content-Based Retrieval" Technical Report CS-94-40, Dept. of Computer Science, Univ. of Virginia, Oct. 1994.
PS version

French, J. C., E. Fox, K. Maly, A. Selman
"Wide Area Technical Report Service - technical reports online" SIGACT News, vol. 25, no. 3 (Sept. 1994), pp. 124-127.

Viles, C. L. and J. C. French
"Availability and Latency of World Wide Web Information Servers" Technical Report CS-94-36, Dept. of Computer Science, Univ. of Virginia, Sept. 1994.
PS version

Jones, J. D. and J. C. French
"An Archive Service with Persistent Naming for Objects" Technical Report CS-94- 35, Dept. of Computer Science, Univ. of Virginia, Aug. 1994.
PS version

Maly, K. J., J. C. French, A. Selman, and E. A. Fox
"Wide Area Technical Report Service" Technical Report TR-94-13, Dept. of Computer Science, Old Dominion Univ., June, 1994.
HTML at Cornell

Barros, J. E. and J. C. French
"Scalable Database Support for Correlation and Fusion Algorithms" Technical Report IPC-93-05, Institute for Parallel Computation, University of Virginia, October 1993.
PS version

French, J. C.
"Estimating Retrieval Regions for Correlation and Fusion Algorithms" Technical Report IPC-93-06, Institute for Parallel Computation, University of Virginia, October 1993.

Maly, K., J. French, E. Fox, A. Selman
"Prototype System Puts Technical Reports On Line" Computing Research News, vol. 5, no. 2 (Mar. 1993), pp. 14.

Maly, K. J., E. A. Fox, J. C. French and A. Selman
"Wide Area Technical Report Service" Technical Report TR-92-44, Dept. of Computer Science, Old Dominion Univ., Dec. 1992.
HTML at Cornell

French, J. C.
"Heuristic Clustering of Signature Files: A New Approach (Preliminary Results)" Technical Report CS-92-42, Dept. of Computer Science, Univ. of Virginia, Sept. 1992.
PS version

French, J. C.,
"Electronic Distribution of Technical Reports and Working Papers: A Simple Cooperative Approach" Technical Report CS-92-27, Dept. of Computer Science, Univ. of Virginia, Sept. 1992.
PS version

IPR2024-01306
UIPATH EXHIBIT 1024

PS version

French, J. C. and C. L. Viles
"A Software Toolkit for Developing Distributed Applications" Technical Report CS-92-26, Dept. of Computer Science, Univ. of Virginia, Sept. 1992.
PS version

Theses, Dissertations and Project Reports-Graduate Works

top

Tang, C.
"TRAPS: A Translator for Postscript Documents" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1991.

Das, M.
"A Study of Some Algorithms for Partitioning Signature Files" Master's Thesis, Dept. of Computer Science, Univ. of Virginia, 1991.

Terry, J.
"Collection and Distribution of Technical Reports" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1992.

Durant, L.
"A Bibliographic Citations Query System in ADAMS" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1992.

Barros, J.
"Scalable Database Support for Correlation and Fusion Algorithms" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1993.

Henry, S.
"Evaluation of the Vector Space Model of Information Retrieval as a Filtering Mechanism" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1993.

Kim, N.
"A Classification Approach to the Automalation of Boolean Queries in Information Retrieval" Ph.D. Dissertation, Dept. of Systems Engineering, Univ. of Virginia, 1993 (coadvisor with D. E. Brown).

Zhu, X.
"IUFROS: An Individual-based User Flexible and Redefinable Open Simulator of Forest Growth and Succession" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1994.

Julian, F.
"Speeding Up Searches with Clustered Signatures" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1994.

Jones, J.
"An Archive Service with Persistent Naming for Objects" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1994.

Wohlford, G.
"Evaluation of N-Grams as Index Elements in Information Retrieval" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1995.

Stumbaugh, M.
"A World-Wide Web Interface for Technical Report Archives" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1995.

Powell, A.
"A Systematic Approach to Creating and Maintaining Software Documentation" Master's Project, Dept. of Computer Science, Univ. of Virginia, 1995.

Viles, C. L.
"Maintaining Retrieval Effectiveness in Distributed, Dynamic Information Retrieval Systems" Ph.D. Dissertation, Dept. of Computer Science, Univ. of Virginia, 1996.
zipped PS
PDF

Emmitt, T.,
"Cue-Validity Variance Database Selection Algorithm Enhancement," Master's Project, Dept. of Computer Science, Univ. of Virginia, 1999.

Wu, B.,
"Building a Digital Library for the Publications of the Computer Science Department," Master's Project, Dept. of Computer Science, Univ. of Virginia, 2000.

Phan, T.,
"A Design and Implementation of Access Control for Distributed Digital Objects and Repositories," Master's Project, Dept. of Computer Science, Univ. of Virginia, 2000.

Monroe, G. A.,
"An Investigation of Language Models of Web Databases Generated Using Query-Based Sampling Techniques," Master's Project, Dept. of Computer Science, Univ. of Virginia, 2000.

IPR2024-01306
UIPATH EXHIBIT 1024

Powell, A. L.,
"Database Selection in Distributed Information Retrieval: A Study of Multi-Collection Information Retrieval," Ph.D. Dissertation, Dept. of Computer Science, Univ. of Virginia, 2001.
(2001 UMI Doctoral Dissertation Award, American Society for Information Science and Technology)
Zipped PS
PDF

O'Neil, E. K.,
"Selective Dissemination of Information in the Dynamic Web Environment," Master's Thesis, Dept. of Computer Science, Univ. of Virginia, 2001.
PS version
PDF version

Undergraduate Theses

top

Werbel, D.
"Design of a Data Visualization System" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1992.

Gomez, C.
"University Course Information System: A Class Scheduler and Course Information Database" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1993.

Gabel, J.
"Designing a Portable User Interface for Posting Technical Reports to WATERS" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1994.

Ryan, M.
"Geographic Information Systems on the World Wide Web" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, January 1997.

Moore, M.
"Electronic Submission of Undergraduate Theses: A Case Study for the Implementation of a University-Wide Program for the Electronic Submission and Online Archiving of Theses and Dissertations" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1997.

Zha, Y.
"Electronic Publishing for a Professional Society: Design and Implementation of Online ACM SIGIR Conference Proceedings" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1997.

Nelson, E.
"Results Merging in the Networked Computer Science Technical Reports Library (NCSTRL)" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, December 1997.

Young, B.
"Selective Searching over the Networked Computer Science Technical Report Library (NCSTRL)" Senior Thesis, Dept. of Computer Science, Univ. of Virginia,
December 1997.

Connelly, Z.
"Electronic Submission of Undergraduate Theses: A Continuation and Enhancement of the Online Archiving of Undergraduate Theses" Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1998.

McDonald, J.,
"Electronic Submission of Undergraduate Theses: Aesthetic and Security Enhancements to the University of Virginia Electronic Submission System," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2000.

Moore, J.,
"Design and Implementation of a Java-based Course Scheduler Application for Undergraduate and Graduate Students Attending the University of Virginia," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2000.

Ogle, V.,
"Study of Search Engine Strategies through Black Box Testing," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2000.

Wong, D.,
"A Convenient Storage Catalog for Graphic Novels: Creating an Online Searchable Database," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2000.

Robbins, B.,
"An Interactive Search Assistant for the EOS Database," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, August 2000.

Funkhouser, T. J.,
``The Technology, Culture, and Communication Toolkit: A Web-based Application for TCC Advisors and Students," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 1999.
(Undergraduate Research and Design Symposium Finalist.)

Ho, S.,
"The Design and Implementation of the Online Ordering and Delivery Tracking System," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, December 2000.

IPR2024-01306
UIPATH EXHIBIT 1024

Hauver, D.,
``Flycasting: Using Collaborative Filtering to Generate a Playlist for an Online Station," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2001.
(Undergraduate Research and Design Symposium Finalist)

Chatlani, P.,
``Relevance Feedback for Web Searches," Senior Thesis, Dept. of Computer Science, Univ. of Virginia, May 2001.
Demo (requires plug-in, may not work under Netscape)

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT J

Captured at: 2024/08/15 11:52 AM URL: https://web.archive.org/web/20040329005629/http:/www.cs.virginia.edu/~cyberia/papers/SIGIR99.pdf

http://www.cs.virginia.edu/~cyberia/papers/SIGIR99.pdf Go FEB MAR MAR

12 captures 29 Mar 2004 - 18 Sep 2011 2003 29 2004 2006 About this capture

camera.dvi 1 / 8 123%

Comparing the Performance of Database Selection Algorithms

James C. French[1], Allison L. Powell[1], Jamie Callan[2], Charles L. Viles[3]
Travis Emmitt[1], Kevin J. Prey[1], Yun Mou[2]

[1]Dept. of Computer Science*
University of Virginia
Charlottesville, VA

[2]Computer Science Dept.†
Univ. of Massachusetts
Amherst, MA

[3]School of Info. and Library Science‡
Univ. of North Carolina–Chapel Hill
Chapel Hill, NC

Abstract We compare the performance of two database selection algorithms reported in the literature. Their performance is compared using a common testbed designed specifically for database selection techniques. The testbed is a decomposition of the TREC/TIPSTER data into 236 subcollections. The databases from our testbed were ranked using both the *gGlOSS* and *CORI* techniques and compared to a baseline derived from TREC relevance judgements. We examined the degree to which *CORI* and *gGlOSS* approximate this baseline. Our results confirm our earlier observation that the *gGlOSS Ideal(l)* ranks do not estimate relevance-based ranks well. We also find that *CORI* is a uniformly better estimator of relevance-based ranks than *gGlOSS* for the test environment used in this study. Part of the advantage of the *CORI* algorithm can be explained by a strong correlation between *gGlOSS* and a size-based baseline (SBR). We also find that *CORI* produces consistently accurate rankings on testbeds ranging from 100–921 sites. However for a given level of recall, search effort appears to scale linearly with the number of databases.

1 Introduction

of collection information to increase retrieval effectiveness [10, 21, 20].

In this paper we focus on *database selection* as a fundamental problem in the distributed environment, providing the first direct comparison of two database selection techniques described in the literature. The problem can be stated intuitively as the process of selecting a (hopefully) small set of databases to which to send a query. Database selection is the first step in a process that continues with search at the distributed sites and fusing or merging of result lists from the sites. The primary goal in this database selection step is to select as small a set of collections as possible to send a query to without sacrificing retrieval effectiveness.

Several techniques for database selection have been proposed and independently evaluated; however the test environments have varied in both the underlying data and the evaluation methods. There is very little literature on the direct comparison of competing techniques. This study directly compares two techniques, *gGlOSS* [13, 14] and *CORI* [7], in a common test environment using the same data and evaluation techniques. We had two goals for this study. The first goal was to demonstrate a sound methodology for the

IPR2024-01306
UIPATH EXHIBIT 1024

EXHIBIT K

Comparing the Performance of Database Selection Algorithms

James C. French[1], Allison L. Powell[1], Jamie Callan[2], Charles L. Viles[3]
Travis Emmitt[1], Kevin J. Prey[1], Yun Mou[2]

[1]Dept. of Computer Science*
University of Virginia
Charlottesville, VA

[2]Computer Science Dept.†
Univ. of Massachusetts
Amherst, MA

[3]School of Info. and Library Science‡
Univ. of North Carolina–Chapel Hill
Chapel Hill, NC

Abstract We compare the performance of two database selection algorithms reported in the literature. Their performance is compared using a common testbed designed specifically for database selection techniques. The testbed is a decomposition of the TREC/TIPSTER data into 236 subcollections. The databases from our testbed were ranked using both the *gGlOSS* and *CORI* techniques and compared to a baseline derived from TREC relevance judgements. We examined the degree to which *CORI* and *gGlOSS* approximate this baseline. Our results confirm our earlier observation that the *gGlOSS Ideal*(l) ranks do not estimate relevance-based ranks well. We also find that *CORI* is a uniformly better estimator of relevance-based ranks than *gGlOSS* for the test environment used in this study. Part of the advantage of the *CORI* algorithm can be explained by a strong correlation between *gGlOSS* and a size-based baseline (SBR). We also find that *CORI* produces consistently accurate rankings on testbeds ranging from 100–921 sites. However for a given level of recall, search effort appears to scale linearly with the number of databases.

1 Introduction

The proliferation of online resources, the growing need to conduct searches across many of these resources, and the de-facto requirement of pruning the resource set of interest to manageable size has increased attention on retrieval in the distributed environment. Distributed information retrieval encompasses many important problems, including database or collection selection [7, 11, 15, 13, 18, 24, 25, 14], collection fusion or results merging [2, 7, 8, 23, 22], and dissemination of collection information to increase retrieval effectiveness [10, 21, 20].

In this paper we focus on *database selection* as a fundamental problem in the distributed environment, providing the first direct comparison of two database selection techniques described in the literature. The problem can be stated intuitively as the process of selecting a (hopefully) small set of databases to which to send a query. Database selection is the first step in a process that continues with search at the distributed sites and fusing or merging of result lists from the sites. The primary goal in this database selection step is to select as small a set of collections as possible to send a query to without sacrificing retrieval effectiveness.

Several techniques for database selection have been proposed and independently evaluated; however the test environments have varied in both the underlying data and the evaluation methods. There is very little literature on the direct comparison of competing techniques. This study directly compares two techniques, *gGlOSS*[13, 14] and *CORI*[7], in a common test environment using the same data and evaluation techniques. We had two goals for this study. The first goal was to demonstrate a sound methodology for the systematic study of these algorithms and their relative performance. The second, and more important, goal was to gain insight into both the collective and individual behavior of these algorithms.

2 Database Selection Experiment

Distributed searching is composed of three fundamental activities: (1) choosing the specific databases to search, (2) searching the chosen databases, and (3) merging the results into a cohesive response. In these experiments we focus specifically on the first activity. Callan *et al.*[7] call this the *collection selection* problem while Gravano *et al.*[15] refer to it as the *text database resource discovery* problem. In French *et al.*[9] we refer to this as *database selection* and will retain that terminology here for consistency.

In earlier work[9] we described a testbed and methodology for examining the behavior of algorithms such as these. We also proposed metrics and analyses that were designed to illuminate the behavior of the algorithm under study. In that work[9], we determined that *gGlOSS*[13] does not estimate relevance-based rankings very well. In the present paper we are concerned with determining how *gGlOSS* com-

*This work supported in part by DARPA contract N66001-97-C-8542 and NASA GSRP NGT5-50062.

†This work supported in part by NSF, the Library of Congress, and the Department of Commerce under agreement EEC-9209623, and by the U.S. Patent and Trademark Office and DARPA/ITO under contract F19628-95-C-0235.

‡This work supported in part by DARPA contract N66001-97-C-8542.

IPR2024-01306
UIPATH EXHIBIT 1024

pares with another database selection algorithm, *CORI*[7]. In this section, we describe the *gGlOSS* and *CORI* algorithms and review the testbed and metrics for comparison.

2.1 The Testbed

A number of researchers have been working on issues in distributed information retrieval systems but the test environments are idiosyncratic in both data and evaluation measures, making it impossible to compare results. In French *et al.*[9] we proposed a test environment for the systematic study of distributed information retrieval algorithms. Our testbed is based on the TIPSTER data used in the TREC[16] conferences. We decompose the large collections into smaller subcollections that serve as hypothetical "sites" in our distributed information retrieval test environment. The data is decomposed by source, year, and month resulting in 236 sites. We used TREC topics 51-150 as the test queries in our earlier study[9]. The characteristics of this testbed, the queries used, and other details can be found in French *et al.*[9].

2.2 The Algorithms Considered

2.2.1 gGlOSS

Gravano *et al.*[15] proposed *GlOSS*, the *Glossary-of-Servers Server*, as an approach to the database selection problem for the Boolean IR model. Later *GlOSS* was generalized to *gGlOSS*[13] to handle the vector space information retrieval model. This generalization can be used for any IR model that computes a score to determine how well a document satisfies a query, provided that certain collection statistics can be made available to *gGlOSS*.

gGlOSS assumes that the databases can be characterized according to their *goodness* with respect to any particular query. *gGlOSS*'s job is then to estimate the goodness of each candidate database with respect to a particular query and then suggest a ranking of the databases according to the estimated goodness.

Goodness for each database, *db*, is defined as follows.

$$Goodness(l, q, db) = \sum_{d \in \{db | sim(q,d) > l\}} sim(q, d) \qquad (1)$$

where $sim(q, d)$ is a function that calculates the similarity between a query q and a document d. Once $Goodness(l, q, db)$ has been calculated for each database db with respect to q at threshold l, the ideal rank for the query at threshold l, $Ideal(l)$ can be formed by sorting the databases in descending order of their goodness.

Note that *gGlOSS* does not compute $Ideal(l)$ rather it is advanced as the goal to which *gGlOSS* estimated ranks $Max(l)$ and $Sum(l)$, defined in [13], will be compared. In French *et al.*[9] we showed that *gGlOSS* $Max(l)$ and $Sum(l)$ estimators do a good job of estimating $Ideal(l)$. We also showed that $Ideal(l)$ is not well-correlated to relevance.

Complete details for calculating the $Max(l)$ and $Sum(l)$ estimators are given in [13] and are not reproduced here. But, for later reference we note that

$$Max(0) = Sum(0) = Ideal(0), \qquad (2)$$

that is, at threshold $l = 0$ both estimators give identically the $Ideal(0)$ ranking of databases for all queries. In addition, $l = 0$ allows a consistent comparison of $Ideal(l)$ rankings when comparing different underlying retrieval systems that produce differently scaled similarity values (i.e. $sim(q, d)$ in Equation 1). Hence, in the evaluation to follow we will use $Ideal(0)$ as the *gGlOSS* estimate since *gGlOSS* can compute this exactly.

Note that *gGlOSS* needs two vectors of information from each database db_i in order to make its estimates.

1. the document frequency df_{ij} for each term t_j in db_i; and
2. the sum of the weight of each term t_j over all documents in db_i.

If the underlying database cannot be made to divulge this information directly, it is in principle possible to recover the information by issuing a single-term query for each vocabulary term. Our choice of $Ideal(0)$ obviates this; we can compute $Ideal(0)$ directly from the databases by simply issuing the test queries.

2.2.2 *CORI*

Given a set of databases to search, the *CORI* approach creates a *database selection index* in which each database is represented by its terms and their document frequencies *df*. Databases are ranked for a query q by a variant of the Inquery document ranking algorithm. The belief $p(r_k|db_i)$ in database db_i due to observing query term r_k is determined by:

$$\begin{aligned} T &= \frac{df}{df + 50 + 150 \cdot cw/\overline{cw}} \\ I &= \frac{\log\left(\frac{|DB|+0.5}{cf}\right)}{\log\left(|DB| + 1.0\right)} \\ p(r_k|db_i) &= 0.4 + 0.6 \cdot T \cdot I \end{aligned} \qquad (3)$$

where:

df	is the number of documents in db_i containing r_k,
cf	is the number of databases containing r_k,
$\|DB\|$	is the number of databases being ranked,
cw	is the number of words in db_i, and
$\overline{cw}$	is the mean cw of the databases being ranked.

The belief in a database depends upon the query structure, but is usually just the average of the $p(r_k|db_i)$ values for each query term [7].

The *CORI* approach to ranking collections can be summarized as $df.icf$, where icf is inverse collection frequency.

2.3 Controlling the Indexing Vocabulary

In [9] we prepared the test collection by using SMART version 11.0[3] using the same parameters as Gravano *et al.*[13]. Note that for those experiments each of the 236 sites used the same parameters and search engine (SMART) to process queries. Then we prepared a union vocabulary incorporating all the terms appearing at any of the separate collections. This gave us a canonical global vocabulary with which to store the document frequencies and weight sums required by *gGlOSS* to make its estimates.

IPR2024-01306
UIPATH EXHIBIT 1024

To control the *CORI* evaluation properly it was necessary to guarantee that the same indexing vocabulary was used by *gGlOSS* and *CORI*. Sources of variability arise from: parsing the original TREC documents; tokenizing a document into words; stemming algorithm; and stoplist. To control for all these factors we synthesized pseudo-documents for each collection. We used the stemmed and stopped vocabulary from the SMART indexing done in [9], emitted each term tf times, and padded the document back to it original length with a fake word. Inquery was then run with stemming off and a single word stoplist. It was also necessary to handle special characters differently. The end result was that Inquery indexed exactly the same vocabulary that SMART did in the earlier experiments.

We verified that the steps we took to control the indexing vocabulary did not hinder the *CORI* performance. We ranked the collections using *CORI* with Inquery defaults then evaluated the ranking. The average performance results were nearly identical to the *CORI* results using the controlled vocabulary.

2.4 Evaluation—Baselines for Comparison

We refer to a number of baselines in the evaluation below, specifically: the *gGlOSS* baseline, $Ideal(0)$; the relevance-based ranking (RBR); and the size-based ranking (SBR). They are defined as follows.

$Ideal(0)$: This ranking is produced by processing each query at each of the 236 subcollections and then using the goodness (see Equation 1) to rank the subcollections.

RBR: These rankings were produced for each query by using the relevance judgements supplied with the TREC data. Databases are simply ordered by the number of relevant documents they contain.

SBR: Databases are ordered by the total number of documents they contain. Note that this ranking is constant for all queries.

2.5 Evaluation—Metrics for Comparison

As in our earlier work we use mean squared error, two recall metrics, $\mathcal{R}_n$ [13] and $\widehat{\mathcal{R}}_n$ [9], and a precision measure, $\mathcal{P}_n$ [13]. In addition we will also use the Spearman coefficient of rank correlation[12].

2.5.1 Recall and Precision Analogs

In [9] we discussed performance metrics that are analogous to the well known IR metrics of *recall* and *precision*. We briefly review the metrics here.

We provide a *baseline ranking*, B, that represents agoal for each query. An algorithm produces some *estimated ranking* for the query, E, and our goal is to decide how well E approximates B. We assume that each database db_i in the collection has some merit, $merit(q, db_i)$, to the query q. The baseline is expressed in terms of this merit; the estimate is formed by implicitly or explicitly estimating merit.

Let db_{b_i} and db_{e_i} denote the database in the i-th ranked position of rankings B and E respectively. Let

$$B_i = merit(q, db_{b_i}) \text{ and } E_i = merit(q, db_{e_i}) \tag{4}$$

denote the merit associated with the i-th ranked database in the baseline and estimated rankings respectively. In the results that follow we have the following convention. For the $Ideal(l)$ calculations $merit(q, db) = Goodness(l, q, db)$; for RBR we define $merit(q, db)$ to be the number of relevant documents in db; and for SBR we define $merit(q, db)$ to be the total number of documents in db.

Gravano *et al.*[13] defined $\mathcal{R}_n$ as follows.

$$\mathcal{R}_n = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n} B_i}. \tag{5}$$

This is a measure of how much of the available merit in the top n ranked databases of the baseline has been accumulated via the top n databases in the estimated ranking.

An alternative definition[9] is given by

$$\widehat{\mathcal{R}}_n = \frac{\sum_{i=1}^{n} E_i}{\sum_{i=1}^{n^*} B_i} \tag{6}$$

where

$$n^* = \max k \text{ such that } B_k \neq 0. \tag{7}$$

Intuitively, n^* is the breakpoint between the useful and useless databases. The denominator is just the total merit contributed by all the databases that are useful to the query. Thus, $\widehat{\mathcal{R}}_n$ is a measure of how much of the total merit has been accumulated via the top n databases in the estimated ranking. Lu *et al.*[17] have also suggested using this measure.

Gravano *et al.*[13] have also proposed a precision-related measure, $\mathcal{P}_n$. It is defined as follows.

$$\mathcal{P}_n = \frac{|\{db \in Top_n(E) | merit(q, db) > 0\}|}{|Top_n(E)|} \tag{8}$$

This gives the fraction of the top n databases in the estimated ranking that have non-zero merit.

2.5.2 Spearman Coefficient of Rank Correlation

The Spearman coefficient of rank correlation, ρ, is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \tag{9}$$

where D_i is the difference in the i-th paired ranks. We have $-1 \leq \rho \leq 1$ where $\rho = 1$ when two rankings are in perfect agreement and $\rho = -1$ when they are in perfect disagreement.

3 Results

In the results that follow, we report the comparison of *gGlOSS* and *CORI*; we use the baseline $Ideal(0)$ to represent *gGlOSS*. For both algorithms, the evaluation was conducted using the full 236 subcollections of the testbed described earlier. We used the TREC topics 51-150 as the test query set. We examine potential reasons for differences in performance and discuss a heuristic for a lower bound on performance.

On the graphs in this section, we label each curve E.B where E is the database selection algorithm (estimator) employed and B is the baseline to which E is being compared. A curve labeled B.B, e.g, RBR.RBR, is intended to show the best possible behavior, that is, when the baseline is used to estimate itself.

IPR2024-01306
UIPATH EXHIBIT 1024

3.1 *CORI* Evaluation

The first step was to evaluate the performance of the *CORI* database selection algorithm using the evaluation measures described in French *et al.*[9]. The evaluation measures include mean-squared error, a precision analog, $\mathcal{P}_n$ and two recall analogs, $\mathcal{R}_n$ and $\widehat{\mathcal{R}}_n$. *CORI* was evaluated using all four measures. For comparison, we also show the performance of *gGlOSS* $Ideal(0)$ [9] for all measures.

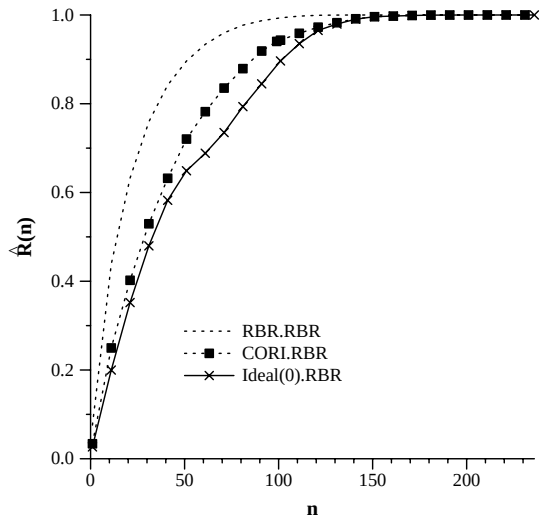

Figure 1: *CORI* results compared to *gGlOSS* and maximum achievable performance.

For these experiments, we used the RBR baseline as the standard against which to compare algorithm performance; the best possible performance under a given measure is to approximate RBR exactly.

We first considered the mean-squared error of the differences between the $Ideal(0)$ and *CORI* rankings and the RBR baseline. MSE can provide a general impression of the similarity of rankings on a query by query basis. The MSE values for both approaches were in the same range, with a minimum value of 940 and a maximum of 4942. However, the MSE values for *CORI* are lower on average with *CORI* having MSE values less than $Ideal(0)$ for 84% of the queries. This suggests that *CORI* may approximate RBR more closely than *gGlOSS*.

We then evaluated *CORI* using the $\mathcal{P}_n$, $\mathcal{R}_n$ and $\widehat{\mathcal{R}}_n$ measures. These measures can help reveal the nature and degree of the discrepancies shown by MSE. The results for $\widehat{\mathcal{R}}_n$ are shown in Figure 1; results for $\mathcal{P}_n$ and $\mathcal{R}_n$ are included in summary Figures 4 and 5.

In Figures 1, 2, 4, 5 and 6, the plots for $Ideal(0)$ and *CORI* should be compared to the plot RBR.RBR. For all measures *CORI* approximates the RBR baseline more closely than *gGlOSS* $Ideal(0)$.

3.2 $\widehat{\mathcal{R}}_n$ Difference

To get a different perspective on the impact of the performance difference, we examined the data plotted in Figure 1. Figure 1 displays the average performance over 100 queries of the $\widehat{\mathcal{R}}_n$ measure. This measure shows the rate at which relevant documents can be accrued as additional collections are considered. Another way to view this data is to determine how many collections from a ranking must be considered to reach a given value for $\widehat{\mathcal{R}}_n$. Table 1 shows the values for *CORI* and *gGlOSS* $Ideal(0)$ compared with the highest achievable value.

Examining Figure 1, it is readily apparent that fewer collections need be considered to reach each level of $\widehat{\mathcal{R}}_n$ when the *CORI* ranking was used than when the $Ideal(0)$ ranking was used. This is reinforced by Table 1. However, it is more apparent from Table 1 that the *CORI* ranking has a greater advantage for $\widehat{\mathcal{R}}_n$ values of 0.7 and above. If the goal is exhaustive searching, using the *CORI* ranking will be more efficient.

There are three regions of interest in Figure 1. These regions can also be seen in Figures 4, 5 and 6. For all measures, for n roughly less than 50, the performance of all of the algorithms is fairly close. For n greater than roughly 120, all algorithms are faced with a situation in which most of the relevant documents have been located and little improvement is possible. However for n in the range of 50-120, *CORI* has a more evident advantage, as seen in Table 1.

$\widehat{\mathcal{R}}_n$	Number of collections required to achieve $\widehat{\mathcal{R}}_n$		
	CORI	$Ideal(0)$	**Best Achievable**
0.1	4	5	2
0.2	9	12	4
0.3	15	18	7
0.4	21	25	10
0.5	29	33	14
0.6	38	43	19
0.7	49	64	26
0.8	64	83	36
0.9	86	102	52
1.0	209	225	145

Table 1: $\widehat{\mathcal{R}}_n$ for *CORI* and $Ideal(0)$

4 Analysis and Discussion

4.1 *CORI* and *gGlOSS* Comparison

The immediate question was why *CORI* approximated the RBR baseline more closely than *gGlOSS* $Ideal(0)$. Our initial observation was that the performance differences could be due to indexing differences. Our initial *gGlOSS* experiments were intended to reproduce the experimental setup of Gravano and Garcia-Molina [13]. As a result, we used the exact underlying retrieval engine and parameter settings reported—SMART using *ntc* weights for document terms and *nnn* weights for query terms.

It has been shown that using different weights and normalization approaches in SMART improves retrieval performance for the large TREC collections [4, 5, 19], so we considered that the improvement shown by *CORI* could be due simply to better-tuned indexing information. There-

IPR2024-01306
UIPATH EXHIBIT 1024

fore, we used the underlying Inquery collection indexes—used to generate the *CORI* inference net—to generate an additional $Ideal(0)$ baseline based upon the Inquery indexing information. We will refer to the new $Ideal(0)$ baseline as *Ideal(0)-inquery*. If *Ideal(0)-inquery* approximated RBR more accurately than $Ideal(0)$, it would imply that the indexing weights used by Inquery provided a better input to the *gGlOSS* algorithm. In this case, it would be more representative to compare the *CORI* performance to that of *Ideal(0)-inquery* than to that of $Ideal(0)$. The comparisons of $Ideal(0)$ and *Ideal(0)-inquery* to baseline RBR for evaluation measure $\widehat{\mathcal{R}}_n$ is shown in Figure 2. The comparisons for $\mathcal{P}_n$ and $\mathcal{R}_n$ are included in summary Figures 4 and 5.

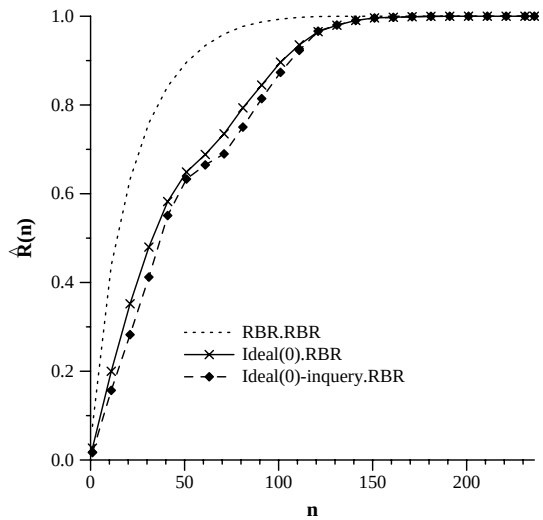

Figure 2: The two $Ideal(0)$s

Note that, in fact, the performance of *gGlOSS* declined when the Inquery weighting information was used. Therefore, we conclude that the improvement in performance of Inquery is not due to differences in term weighting. At present, we are not certain what aspect of the *CORI* database selection algorithm is responsible for its performance difference. We have conjectured that to *gGlOSS* a database with many documents of low similarity may appear more useful than a database with a few documents of high similarity [9]. It is possible that *CORI* is utilizing a better length normalization strategy that allows it to avoid this difficulty.

4.2 Formulation of SBR

Our 236 collection testbed is produced by partitioning the TREC data by publishing source, year and month. As a result, the number of documents in a collection varies widely for our testbed. During a detailed examination of results from our *gGlOSS* experiments [9], we noted some recurring themes in the rankings produced by $Ideal(0)$.

It appeared that $Ideal(0)$ routinely ranked highly collections with very large numbers of documents. An examination of the RBR baseline revealed that collections with a large number of documents also tended to have a good representation of relevant documents. Voorhees noticed a similar phenomenon in her work[22], specifically that the distribution of relevant documents was uneven across document sources. She noted that a strategy of retrieving from AP and WSJ was often very effective with TREC data. In our decomposition, subcollections derived from AP and WSJ tend to have a large number of documents.

We created a new baseline, size-based ranking (SBR), in which databases are ordered by the total number of documents that they contain. We used the SBR baseline to determine if a correlation between collection size and $Ideal(0)$ or RBR did in fact exist. We used Spearman's ρ to measure the correlation between $Ideal(0)$ and SBR and between RBR and SBR. The results are shown in Figure 3.

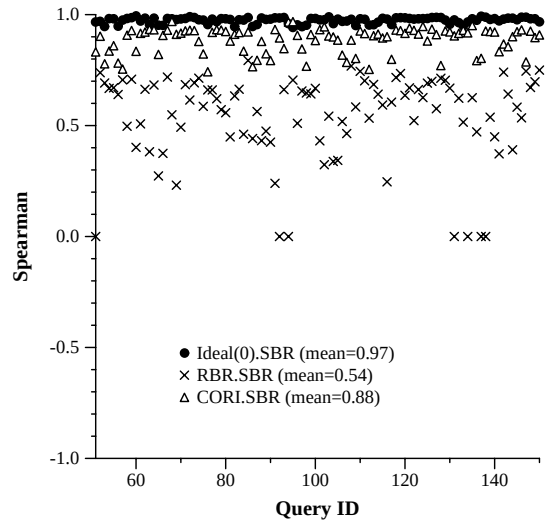

Figure 3: Spearman's ρ values for $Ideal(0)$ and RBR compared to SBR.

We noted that there is a very strong positive correlation between $Ideal(0)$ and SBR for all queries. The correlation between *CORI* and SBR is not as pronounced, but still strong. There is also a moderate positive correlation between RBR and SBR for most queries. SBR is a simple heuristic that is constant for all queries and therefore not at all computationally intensive. We will show later that SBR can closely approximate the performance of $Ideal(0)$ for this testbed. We consider SBR a useful lower bound on database selection performance for this collection.

4.3 Relationship between SBR and Ideal(0)

gGlOSS computes collection scores by summing over documents in a collection. As a result, *gGlOSS* favors collections with a large number of documents. For this testbed, the performance of *gGlOSS* $Ideal(0)$ is strongly influenced by this feature of the algorithm. $Ideal(0)$ is strongly positively correlated with SBR. Evidence of this correlation can be seen in Figures 4, 5 and 6. The approximation of RBR by $Ideal(0)$ closely tracks the approximation of RBR by SBR. However, $Ideal(0)$ does consistently outperform SBR.

Figures 4, 5 and 6 also show how all of the approaches discussed here approximate RBR. For all measures, *CORI*

IPR2024-01306
UIPATH EXHIBIT 1024

outperforms *gGlOSS*. Also note that SBR represents a useful lower bound for performance in this testbed.

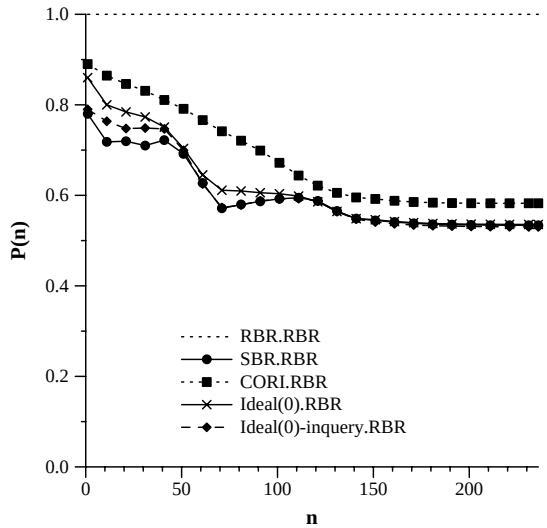

Figure 4: *CORI* results compared to *gGlOSS* and maximum achievable performance using $\mathcal{P}_n$.

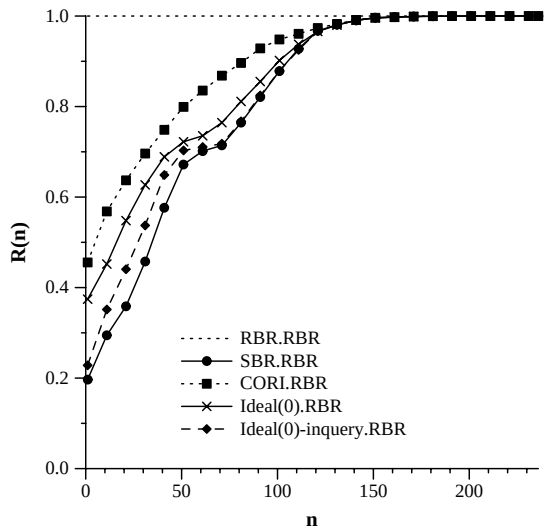

Figure 5: *CORI* results compared to *gGlOSS* and maximum achievable performance using $\mathcal{R}_n$.

The *CORI* and $Ideal(0)$ plots appear similar in Figure 6, making it difficult to judge the performance difference between *CORI* and $Ideal(0)$. Examining the plotted data, we determined that *CORI* shows at least 25% improvement over $Ideal(0)$ for $n \leq 10$ and approximately 10% improvement for $n \leq 90$. We also note that $Ideal(0)$ shows at least 20% improvment over SBR for $n < 50$; however, for $n > 50$, $Ideal(0)$ tracks SBR much more closely.

Future experiments will need to examine the implications of the correlation between $Ideal(0)$ and SBR. For this

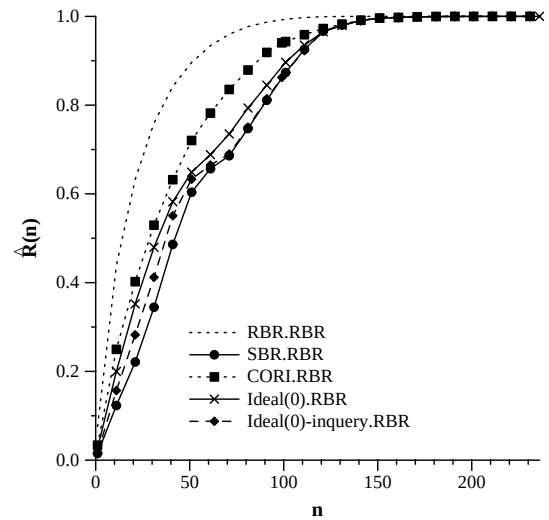

Figure 6: *CORI* results compared to *gGlOSS* and maximum achievable performance using $\widehat{\mathcal{R}}_n$.

testbed and the experiments reported here, it is not clear if the preference for large collections degraded the performance of *gGlOSS*; for this testbed there was also a positive correlation between SBR and RBR. However, given a testbed with a negative correlation between SBR and RBR, a preference for large collections could degrade the performance of *gGlOSS* or any other database selection algorithm with this feature.

In future experiments, we will create additional testbeds in which the number of documents in a collection will be held constant. This will allow us to examine algorithm performance when collection-size preference is controlled for.

5 Scaling Up Database Selection Algorithms

Database selection algorithms become increasingly necessary as the number of databases grows from $O(10)$, to $O(100)$, to $O(1,000)$ databases. Early research on database selection was based on testbeds of $O(10)$ databases [7, 15], and suggested that selection algorithms were very accurate. However, as the algorithms were improved and applied to testbeds of $O(100)$ databases, the experimental results became more ambiguous [13, 17, 24]. A high priority for our recent research has been to determine whether the *CORI* algorithm remains effective as the number of databases increases.

5.1 Testbeds

Experiments were conducted with the 236 collection testbed used throughout this paper, as well as two additional testbeds of 100 collections and 921 collections. Summary characteristics of these three testbeds are provided in Table 2.

The 100 collection testbed was created by dividing TREC CDs 1, 2, and 3 into 100 smaller databases of roughly equal size (about 30 megabytes each). Each database contained documents from a single source, ordered as they occur on the TREC CDs; hence documents in a database were also usu-

IPR2024-01306
UIPATH EXHIBIT 1024

Num. DB	Source	Queries	Size (GB)	Total Docs
100	TREC 1+2+3	51-100	3.3	1,078,166
236	TREC 1+2+3 subset	51-150	2.7	691,058
921	TREC-6 VLC	301-350	20.1	7,492,048

Table 2: Characteristics of testbeds used in scaling experiments.

ally from similar timeframes. CD 1 contributed 37 databases, CD 2 contributed 27 databases, and CD 3 contributed 36 databases.

The 921 collection testbed was created by dividing the TREC-6 VLC corpus into 921 smaller databases of roughly equal size, subject to the same source and document adjacency constraints used for creating the 100 collection testbed.

Queries were based on TREC topics 51-150 and 301-350. We used query sets INQ001, INQ026, and INQ411, all created by the UMass CIIR as part of its participation in TREC and Tipster evaluations [1, 6]. Queries in these sets are long, complex, and have undergone automatic query expansion.

The relevance assessments were the standard TREC relevance assessments supplied by the U.S. National Institute for Standards and Technology. Relevance judgements for the VLC data are relatively sparse, so additional relevance judgements for 20 queries were gathered by trained undergraduates. The additional relevance judgements had little impact on results, and so were not used in the results reported here.

5.2 Experiments

The effectiveness of the *CORI* database selection algorithm was measured with $\mathcal{R}$ and $\widehat{\mathcal{R}}$, described above. In these tests, $\mathcal{R}$ and $\widehat{\mathcal{R}}$ were applied to the *percentage* of databases searched, so that experiments on testbeds of different sizes could be compared. We call these variations $\mathcal{R}(p)$ and $\widehat{\mathcal{R}}(p)$, where $p = n/N$, and N is the number of databases in the testbed.

5.3 Results

Figures 7 and 8 summarize the experimental results. Measurements with $\widehat{\mathcal{R}}(p)$ (Figure 7) show that on all testbeds the collections ranked highly by the *CORI* selection algorithm contain a large percentage of the relevant documents. This is an encouraging sign.

It may appear that the algorithm is *more* effective as more databases are available for searching, but this is an artifact of the distribution of relevant documents across the databases. Measurements with $\mathcal{R}$ (Figure 8), which normalizes $\widehat{\mathcal{R}}$ by the best result possible from searching that percentage of databases, show that effectiveness on the 100 and 921 collection testbeds is very similar. The only difference is when a very small percentage of collections is searched; in this case, effectiveness with the large testbed is substantially lower.

The 236 collection testbed appears to be a slightly easier testbed than the 100 collection and 921 collection testbeds,

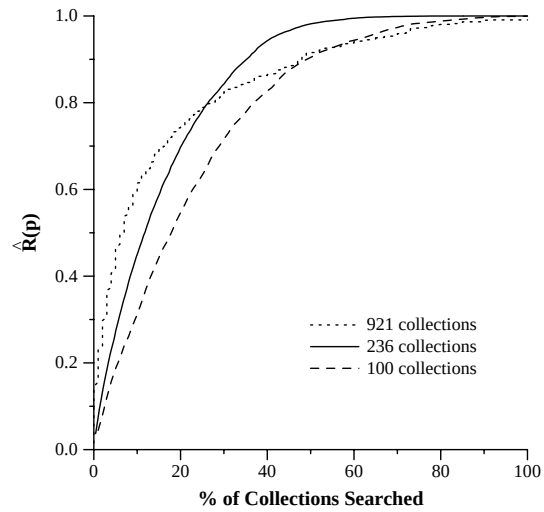

Figure 7: Database selection accuracy on testbeds of different sizes.

but the results are otherwise similar.

The effectiveness of the *CORI* algorithm appears to be related consistently to the percentage of databases searched, irrespective of the testbed size. This is not inappropriate for recall-oriented metrics such as $\widehat{\mathcal{R}}$ and $\mathcal{R}$, but it implies that a tenfold increase in available databases requires a tenfold increase in search effort to achieve a given level of recall. We would prefer a less linear relationship between testbed size and search effort, which suggests that additional research on database selection algorithms is required.

6 Conclusions

We have examined the performance of two database selection algorithms, *gGlOSS* and *CORI*. We have demonstrated the high correlation between $Ideal(l)$ ranks and the SBR baseline. This supports our conjecture[9] that to *gGlOSS* a database with very many documents of marginal similarity will appear, in aggregate, to be more useful than a database having a few documents with large similarity. We have shown that *CORI* is a better estimator of RBR than *gGlOSS* and is at least 10% more effective through $\widehat{\mathcal{R}}_n$ levels of 90%. In part this may be due to a better length normalization strategy so that *CORI* is not as confounded by databases with many documents of small similarity. We have also shown that the difference in performance between *CORI* and *gGlOSS* is not due to Inquery weighting versus SMART weighting.

In our experiments, the *CORI* approach required fewer databases to be searched than *gGlOSS* on average to achieve a specific recall level. This implies that *CORI* will be more cost-effective in large, distributed environments having lower latency for searches.

Our experiments investigating how well the *CORI* algorithm scales to large numbers of databases produced a more mixed message. The *CORI* selection algorithm produced consistently accurate rankings on testbeds ranging from 100

IPR2024-01306
UIPATH EXHIBIT 1024

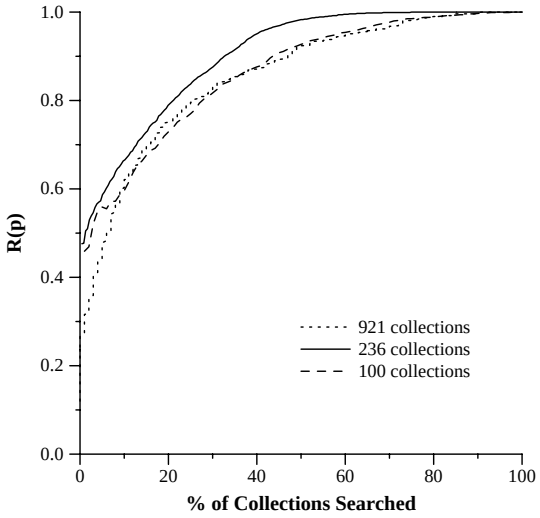

Figure 8: Scaled database selection accuracy on testbeds of different sizes.

to 921 collections. This is clearly a positive result. However, consistency was observed in the *percentage* of databases searched, not in the *number* of databases searched. For a given level of recall, search effort appears to scale linearly with the number of available databases. A less linear relationship is desirable, and perhaps required, in environments containing very large numbers of databases. Our results suggest that further research on database selection algorithms is required.

We have demonstrated a systematic methodology for examining database selection algorithms. This is the first study of this kind. We can now investigate other database selection algorithms in the same framework and be able to compare performance in a common experimental environment.

It will be instructive to repartition our testbed into a similar number of sites, but with each site having about the same number of documents. There are a number of ways to do this, but each has drawbacks. Such a partition would completely defeat the SBR baseline because every site would have equal merit and thus all sites would be equally useful under the SBR baseline. We think that this might be a more demanding search environment and offer greater insight into the behavior of these kinds of algorithms.

Acknowledgements. We thank Margaret Connell for her assistance with experiments reported here.

References

[1] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. INQUERY does battle with TREC-6. In *Proc. TREC-6*.

[2] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Informa tion Retrieval. *Information Processing & Management*, 31(4):431–448, 1995.

[3] C. Buckley. SMART version 11.0, 1992. ftp://ftp.cs.cornell.edu/pub/smart.

[4] C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART : TREC 2. In *Proc. TREC-2*, pages 45–56, 1994.

[5] C. Buckley, G. Salton, and J. Allan. Automatic retrieval with locality information using SMART. In *Proc. TREC-1*, pages 59–72, 1993.

[6] J. P. Callan, W. B. Croft, and J. Broglio. TREC and TIPSTER experiments with INQUERY. *Information Processing & Management*, 31(3):327–343, 1995.

[7] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *Proc. SIGIR'95*, pages 21–29, 1995.

[8] E. A. Fox, M. P. Koushik, J. Shaw, R. M. odlin, and D. Rao. Combining Evidence from Multiple Searches. In *Proc. TREC-1*, pages 319–328, 1992.

[9] J. C. French, A. L. Powell, C. L. Viles, T. Emmitt, and K. J. Prey. Evaluating Database Selection Techniques: A Testbed and Experiment. In *Proc. SIGIR'98*, pages 121–129, 1998.

[10] J. C. French and C. L. Viles. Ensuring Retrieval Effectiveness in Distributed Digital Libraries. *Journal of Visual Communication and Image Representation*, 7(1):61–73, 1996.

[11] N. Fuhr. A Decision-Theoretic Approach to Database Selection in Networked IR. *ACM Trans. on Information Systems*. to appear.

[12] J. D. Gibbons. *Nonparametric Methods for Quantative Analysis*. Holt, Rinehart and Winston, 1976.

[13] L. Gravano and H. Garcia-Molina. Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. In *Proc. VLDB'95*, 1995.

[14] L. Gravano, H. Garcia-Molina, and A. Tomasic. GlOSS: Text-source discovery over the internet. *ACM Trans. on Database Systems*, To appear.

[15] L. Gravano, H. Garcia-Molina, and A. Tomasic. The Effectiveness of GlOSS for the Text Database Discovery Problem. In *Proc. SIGMOD'94*, pages 126–137, May 1994.

[16] D. Harman. Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proc. TREC-4*, 1996.

[17] Z. Lu, J. P. Callan, and W. B. Croft. Measures in collection ranking evaluation. Technical Report TR-96-39, Computer Science Department, University of Massachusetts, 1996.

[18] A. Moffat and J. Zobel. Information Retrieval Systems for Large Document Collections. In *Proc. TREC-3*, pages 85–94, 1995.

[19] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. SIGIR'96*, pages 21–29, 1996.

[20] C. L. Viles and J. C. French. TREC4 Experiments Using DRIFT. In *Proc. TREC-4*, 1996.

[21] C. L. Viles and J. C. French. Dissemination of Collection Wide Information in a Distributed Inform ation Retrieval System. In *Proc. SIGIR'95*, pages 12–20, July 1995.

[22] E. Voorhees, N. Gupta, and B. Johnson-Laird. Learning Collection Fusion Strategies. In *Proc. SIGIR'95*, pages 172–179, 1995.

[23] E. Voorhees, N. Gupta, and B. Johnson-Laird. The Collection Fusion Problem. In *Proc. TREC-3*, pages 95–104, 1995.

[24] J. Xu and J. Callan. Effective Retrieval with Distributed Collections. In *Proc. SIGIR'98*, pages 112–120, 1998.

[25] B. Yuwono and D. Lee. Server Ranking for Distributed Text Retrieval Systems on Internet. In *Proc. of the Int. Conf. on Database Systems for Adv. Applications*, pages 41–49, 1997.

IPR2024-01306
UIPATH EXHIBIT 1024