Selection of Overlapped Databases

Joaquim Macedo, António Costa, and Vasco Freitas
{macedo,costa,vf}@uminho.pt

Departamento de Informática,
Universidade do Minho, Campus de Gualtar,
4710-057 Braga, Portugal

Abstract. Database selection is one of the key functions of a distributed information retrieval system. In presence of uncoordinated replication, some databases may have a large subset of its returned documents overlapped on responses of best ranked databases. This will lead to degradation of the selection algorithm performance.
The work reported in this paper is concerned with selection of overlapped databases and how the use of overlapping information may prevent the expected performance degradation. A well know database selection algorithm is evaluated in Laboratory using an overlapped database context. Experimental results show that the performance degradation of the selection can be reduced if overlapping among different databases is used as a parameter of the selection function.
Keywords: Network Information Retrieval, Database Selection, Replication

1 Introduction

Internet results from the interconnection of autonomous organizations networks, each one with its own user policy and without any general technical coordination. Thus, one may expect to detect non intentional replication in distributed applications such as distributed information retrieval systems. In more technical coordinated environments, replication may be used as the key for achieve good performance, high availability and fault tolerance in a distributed system. In both scenarios, the replicated system must be accessed by its users in a consistent way.

Replication in distributed information systems is due to the existence of replicated documents in the indexed information space or due to the overlap of the indexed information space (see Figure 1).

Replication is a source of additional complexity: the text database must avoid the re-indexing of the same document; the database selection algorithm must adjust the ranking of relevant sources when a given level of relative overlapping exists; even when the query is broadcasted to all sources, the number of requested documents needs some adjustment; data fusion algorithms must be able to remove replicated document items when joining different result sets.

Our previous work concerning the fusion of different overlapped result sets, showed that, for scenarios in which there are only the rank or similarity of the

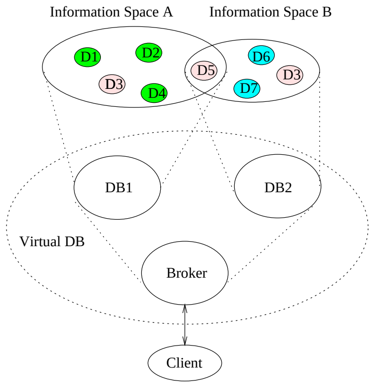


Fig. 1. Data Base overlapping

documents, the use of overlapping information may enable the improvement of fusion results effectiveness [MCF00a].

The work reported in this paper is concerned only with database selection [CPC+99,LCC96,GMT94,Bau97] in database overlapping contexts, produced by uncoordinated replication.

2 Database Overlapping

The overlapping between database X_1 on database X_2 can be defined as

$$O(X_1|X_2) = \frac{|X_1 \cap X_2|}{|X_1|} \tag{1}$$

where $|X_i|$ is the cardinality of the database X_i.

The overlapping of the database X_1 with a set of databases $X_2, X_3, ..., X_N$ is defined as

$$O(X_1|X_2, X_3, ..., X_N) = \frac{|X_1 \cap (X_2 \cup X_3 \cup .... \cup X_N)|}{|X_1|} \tag{2}$$

Note that the one-to-many overlapping is not equal to sum of several one-to-one overlapping, when there is more than one copy of each document, that is $O(X_1|X_1, X_2, ..., X_N) \neq \sum_{k=1}^{N} O(X_1|X_k)$.

In the sequel, the overlapping between databases for each query is named *dynamic overlapping* and the one between the databases document set is the *static overlapping*. In a distributed information retrieval environment, dynamic overlapping is the most important.

For result sets fusion, the detection of the overlapping is done with the actual results sets. In such case, previous overlapping estimation may only prevent unneeded overlapping detection.

However, for database selection purposes, overlapping must be estimated a priori. In such case, it may be relevant for each DB to compute the one-to-many overlapping with best ranked DBs.

At broker level, the information about the content of the different databases is limited. Thus, the real value of overlapping can not be computed, it can be only estimated. For such estimation, only the response for past queries can be used. As stated by [BB98] a careful set of queries may enable the computation of static overlapping among different search engines.

For dynamic overlapping estimation, the broker must have some local database with sample document set from each remote database. The internal identification of documents (replica allowed) must be done using an heuristic to identify the source database. The local database is build using probing methods proposed in [DC98]. The same local database may be used for query expansion. This is a subject for future work.

In this paper, only static overlapping will be considered. As Laboratory environment enables the easy computation of static overlapping, we will substitute the method proposed in [BB98] by database document sets operations.

3 Experimental Scenario

The distributed search engine used was Webtrieve [MCF00b]. Webtrieve system is based on the Vector Space Model and it is composed by four main components: a batch client, a broker, a search engine and a document allocator. The broker is a specialized search engine whose documents are the summaries exported by other search engines (or brokers) combined into an hierarchical information mesh. The broker retrieval strategy is used to select the relevant search engines for a given query. The broker has also fusion and overlapping related features. The STARTS protocol [GCGM+97] is used for communication between different components. The actual prototype is being developed using Java and it is intented for use in Laboratory experiments with TREC [Har93] collections.

Experimental results were achieved by indexing 1991 and 1992 Wall Street Journal documents (TREC disks 1,2), whose statistics are presented in table 1. Webtrieve was used with stemming and stopwords remotion. To simulate the distributed environment, documents were allocated into 10, 20, 40 and 80 different databases using a random allocator with replication. The queries were generated from TREC 151-200 topics.

Documents of the original collection were allocated into different databases, using a random allocator with replication. The allocation is explained in the Figure 2. The relation of the overlapping parameter (o) and the number of replicas is given by the expression 3.

Disk	Content	Size Mb	Number of Docs	Word/Doc median	Word/Doc average
1	WSJ,1987-1989	267	98,732	245	434.0
2	WSJ, 1990-1992	242	74,520	301	508,4

Table 1. TREC collections used

$$(1-o)+\sum_{i=1}^{N-1}\frac{q}{i}=1 \tag{3}$$

$p=1-o$ is the probability of a given document not having a copy in the overall system. The probability of a document d to have i copies is $\frac{q}{i}$. Thus, a random X is generated. If $X \leq p$ the document will have 0 copies, if $p < X \leq p+q$ one copy, $p+q < X \leq p+1.5q$ two copies, etc...

$$\text{number of copies} = \begin{cases} 0 & \text{if } X < p \\ 1 & \text{if } p \leq X < p+q \\ 2 & \text{if } p+q \leq X < p+q+\frac{q}{2} \\ \ldots & \ldots \\ i & \text{se } p+\sum_{k=1}^{k=i-1}\frac{q}{k} \leq X < p+\sum_{k=1}^{k=i}\frac{q}{k} \\ \ldots & \ldots \\ N-1 & \text{if } p+\sum_{k=1}^{k=N-1}\frac{q}{k} \leq X < 1 \end{cases}$$

Fig. 2. Random allocation model with replication using an overlapping parameter ($o = 1-p$);

A variant of this method with several classes of overlapping has been also introduced. The aim is to create a scenario where a given level of variance is ensured for the overlapping of different databases. There are several classes of databases with a given overlapping level and without any inter-class overlapping.

The awkwardness of the method comes from the fact that is not possible predict the number of copies and the location of a document, even knowing the overlapping level and the number of BDs. However, bear in mind that the goal is the simulation of an uncoordinated replication environment.

The naming of different scenarios used is webtXX.YY The YY is the value of overlapping parameter used (5%, 10%, 15%, 20%, 25% and 30%) and XX is the number of databases for each scenario. Thus, the scenario webt80.15 is the one with 80 DBs and 15% of overlapping parameter.

4 Baselines for overlap scenarios

Metrics used for selection evaluation are precision, recall and alternative recall[CPC+99]. These metrics are defined using the RBR (Relevance Based Ranking) baseline, that is the number of relevant documents in each database for a given query. In this scenario, the used baseline must be modified with overlapping information. So instead of RBR, the used baseline is derived from the equivalent one (ORBR [1]) using overlap information. *ORBR* are the relevant documents in DB_i not occuring on best ranked databases. When database document sets are disjoints, the ORBR and RBR are identical.

Figure 3 shows the best possible alternative recall using these baselines for webt80.30 DB scenario. To see the difference between the baselines (ORBR and RBR) the optimal alternative recall was computed. As shown in figure 3, the ORBR optimal alternative recall is higher than the RBR one, due to the existence of replicated documents. n is the number of considered databases.

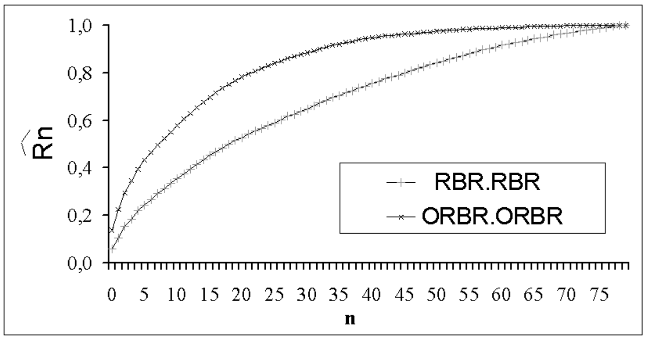


Fig. 3. Baselines using ORBR and RBR

5 Using overlapping information

The goal of our next experiment, using webt10.15 scenario, was to verify wether or not it is possible to consider the overlapping information into DB selection ranking using the inaccurate values given by the static overlapping. Our aim is to answer the question of wether or not is possible to have an overlapping factor α_{ij}, function of the static overlapping with best ranked DBs, and get a best correlation with optimal selection.

In this experiment, we considered as overlapping factor the value

[1] Overlap RBR

$$\alpha_{ij} = (1 - \frac{R_{ij} - 1}{N - 1} * O_j) \quad (4)$$

where

R_{ji} is the rank of DB_j for Q_i
(in this case a value between 1 and 10)

N the actual number of DBs (10)

O_j static overlapping of DB_j with all other DBs (from 0.0 to 1.0)

The formula 4 is a linear indication of the overlapping using the one-to-all overlapping and the ranking of the database in the selection.

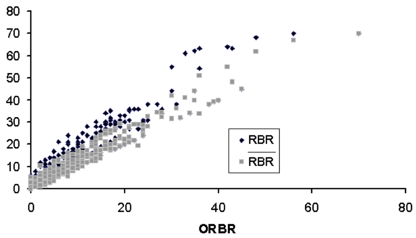


Fig. 4. Correlation between RBR and $\overline{RBR}$ versus ORBR

The queries were automatically generated using the TREC topics 151-200. For each query, the number of relevant documents in each DB (RBR_{ij}) as computed as well as the number of relevant documents not occurring at best ranked DB ($ORBR_{ij}$).

This procedure was compute for all queries and all DBs using value $\overline{RBR_{ij}} = \alpha_{ij} * RBR_{ij}$. Moreover for all these values compute the correlation between $(RBR_{ij}, ORBR_{ij})$ and $(\overline{RBR_{ij}}, ORBR_{ij})$. The gotten results, RBR_{kj} versus $ORBR_{kj} = 0,9432$ and $\overline{RBR_{kj}}$ versus $ORBR_{kj} = 0,9769$, give a good indication about the use of overlapping information by the selection algorithm. The correspondent dispersion diagram is shown in Figure 4.

In the following experiments, the above overlapping factor was adjusted to introduce the maximum overlapping with best ranked DBs as the minimum value and, of course, the overall overlapping as the maximum. The formula used was the one presented in 5. The selection algorithm used was CORI[LCC96,FPV98] with parameters b=0.75 and k=200. In this way, the CORI rank for each database is adjusted with overlapping factor.

$$\alpha_{ij} = \begin{cases} 1 & \text{if } R_{ij} = 1 \\ (1 - O_{max(j)} + (O_j - O_{max(j)}) * (\frac{R_{ij}-2}{N-2} * O_j) & \text{if } R_{ij} > 1 \end{cases} \quad (5)$$

Where O_j is the one-to-all overlapping, $O_{max(j)} = max\{O(DB_j/DB_k) : R_{ik} < R_{ij}\}$ and O_j is the one-to-all overlapping DB_j. This gives the maximum one-to-one overlapping with best ranked DBs.

6 Experimental results

The first issue was the study of the results given by the used allocation method. We aim to find, for a given overlapping parameter used, what is the overlapping for the different databases. Also we want to find out what happens when the number of DBs increases. This statistics, showed in table 2, may help to understand the performance of selection results. The number of documents increases with the overlapping parameter and decreases with the number of DBs. This means that the size of databases is the same for all scenarios. The global and one-to-one overlapping increases with the number of databases and with the overlapping parameter. The values of these types of overlapping has a correlation value of 0,99.

	5		**10**		**15**		**20**		**25**		**30**	
	$\overline{D}$	ΔD	$\overline{D}$	ΔD	$\overline{D}$	ΔD	$\overline{D}$	ΔD	$\overline{D}$	ΔD	$\overline{D}$	ΔD
10	20026	103,2	22882	154,4	25570	116,6	28324	75,4	31077	183,8	33852	172,4
20	10904	90,9	13216	106,8	15548	103,0	17932	97 ,9	20186	120,6	22252	144,5
40	6302	61,4	8222	63,6	10331	85,6	12334	70, 6	14244	69,9	16275	82,0
80	3853	52,3	5537	57,6	7314	64,8	9031	72,1	10777	79,1	12535	87,3
	$\overline{O_g}$	ΔO_g	$\overline{O_g}$	ΔO_g	$\overline{O_g}$	ΔO_g	$\overline{O_g}$	ΔO_g	$\overline{O_g}$	ΔO_g	$\overline{O_g}$	ΔO_g
10	17,93	0,14	32,13	0,21	42,56	0,23	51,16	0,18	58,31	0,23	64,24	0,22
20	24,65	0,24	41,09	0,28	52,78	0,26	61,43	0,29	67,88	0,28	73,19	0,31
40	34,84	0,40	52,68	0,41	64,49	0,38	72,02	0, 33	77,23	0,35	81,39	0,31
80	46,74	0,56	64,85	0,50	74,88	0,42	80,89	0,4 2	84,94	0,33	87,97	0,30
	$\overline{O}$	ΔO	$\overline{O}$	ΔO	$\overline{O}$	ΔO	$\overline{O}$	ΔO	$\overline{O}$	ΔO	$\overline{O}$	ΔO
10	9,13	0,14	16,28	0,21	21,63	0,23	26,05	0,18	29,55	0,23	32,66	0,22
20	11,81	0,24	19,93	0,28	25,50	0,26	29,95	0,29	33,13	0,28	35,55	0,31
40	16,94	0,40	25,34	0,41	31,27	0,38	35,17	0,33	37,63	0,35	39,92	0,31
80	22,52	0,56	31,51	0,50	36,49	0,49	39,22	0,46	41,59	0,40	43,07	0,40

Table 2. Average and standard deviation for the number of documents ($\overline{D}$ and ΔD) one-to-ne overlapping ($\overline{O}$ and ΔO) and one-to-all overlapping ($\overline{O_g}$ and ΔO_g) on several scenarios

Table 2 gives the starting point for most of the scenarios used on this paper. The numbers in first row are the overlapping parameter used (5%, 10%, 15%, 20%, 25% and 30%) and the number in the first column is the number of databases for each scenario. Thus, the scenario webt80.15 is the one with 80 DBs and 15% of overlapping parameter.

From these scenarios, the databases from webt80.?? scenarios were mixed to build the scenario webt80.xx ($\overline{O} = 9,57$, $\Delta O = 11,73$, $\overline{O_g} = 93,05.$, $\Delta O_g = 2,32$). The idea is to increase variance on the overlapping values between different databases.

Databases from scenarios webt10.??, webt20.??, webt40.?? and webt80.?? were also mixed to build the scenario webt80.zz ($\overline{O} = 8,12$, $\Delta O = 6,9$, $\overline{O_g} = 99,26$, $\Delta O_g = 0,30$). This second scenario includes databases with more differences on the number of documents. To have an alternative scenario with larger variance on the overlapping, the allocation method with overlapping classes was used to generate the scenario webt80.var ($\overline{O} = 13.98$, $\Delta O = 16.12$, $\overline{O_g} = 58.27$, $\Delta O_g = 22,85$). To generate such scenario, two classes were used both with 40 databases and overlapping parameter of 5% and 30%.

6.1 Static versus dynamic overlapping

In the next experiment , the analyzed factor was the overlapping of several result sets returned by different DBs. In distributed information retrieval, this is the looked forward estimation . With this purpose, the used queries were derived from TREC 151-160 topics. The idea is to broadcast the queries to the existing databases and compute the overlapping of the returned result sets. The used scenario was webt10.15 to reduce the number of results. The query overlapping was compared with static overlapping and the with average overlapping of all the queries. The DB 0 was used as example.

The static one-to-one overlapping has values between 21,35 % and 22,18% with a maximum variation of 0,83%. The query one-to-one overlapping of DB 0 has values between 9,79% and 30,56%. The global overlapping has static value of 42,96% and query value between 31,94 % (query 154) and 44,4 % (query 157). The query one-to-one overlapping for DB 0 has 0,9 correlation with static overlapping and 0,92 with the average overlapping of all queries.

The botton line is that static overlapping gives only a good indication of possible query overlapping values.

6.2 Selection performance results

The presented selection results are precision, recall and alternative recall for the scenarios webt80.xx, webt80.zz and webt80.var presented in Figures 5 and 6. These results show that the overlapping factor improves the effectiveness of the results in all scenarios. In all situations the effectiveness improvement shows itself in the first half of selected databases, which is precisely when it is more relevant.

This improvement is more evident when the variance of the overlapping between databases increases. In such case the overlapping factor may change the rank of the database in the selection results. Even for the cases where variance values are less noticeable, the results are the quite similar using (or not) the proposed overlapping factor.

7 Related Work

The work reported here is based on the CORI database selection algorithm [CPC+99,LCC96]. The intended goal is to study the CORI performance on database overlapping context and to check if its effectiveness can be improved using overlap information. For evaluation purposes, the used metrics [CPC+99] are refined using overlap information.

Another important related work is [ZS99] which proposes a replica architecture for distributed information retrieval, in which a replica selection algorithm enables user queries load balancing, between original and replica servers, without degradation of retrieval effectiveness . The replica selection algorithm is also CORI based . The aim of this work is not to propose a replication architecture. Rather it uses replica distribution information to prevent expected selection algorithm performance degradation when there is some level of overlapping between the eligible databases. The main difference comes from the coordinated and uncoordinated replication scenarios.

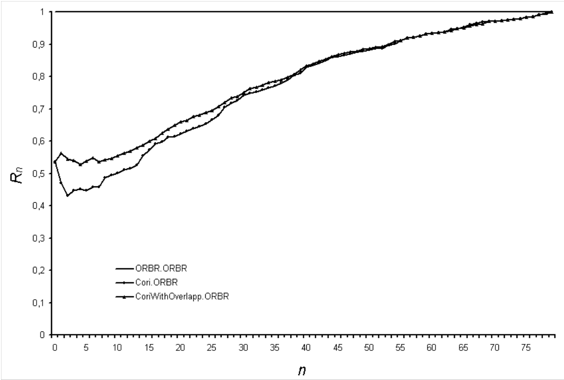
Important work already exists to prevent the return of replicated document itens to the end user, remotion of duplicates in information dissemination services and to prevent copyright violation[SGM94,BDM95].

However the closest work the one presented here, it is that of [BB98] for measuring the relative size and overlap of public WWW search engines. Its main goal is to enable a third party evaluation of public search engines claimed merits. The statistical technique used is based on random queries.

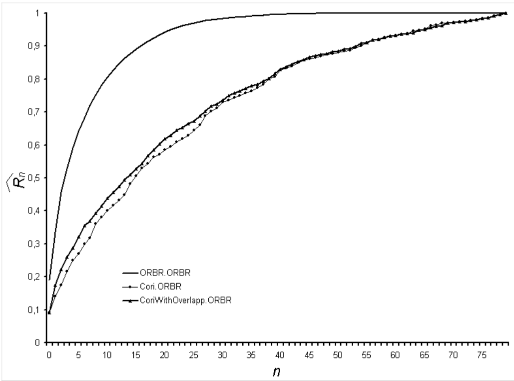
In this work, database overlapping estimation capabilities are intended to enhance meta-crawler or brokers behaviour when queried about target database contents. Thus, a large number of databases can be involved and the available information about each one may also differ. [BB98] uses query overlapping to measure static relative and size and overlapping. In this paper, static overlapping is only used as an indicator of query overlapping. However since, we believe that query overlap can have a better estimation, it will be replaced by more suitable techniques.

8 Conclusions and Further Work

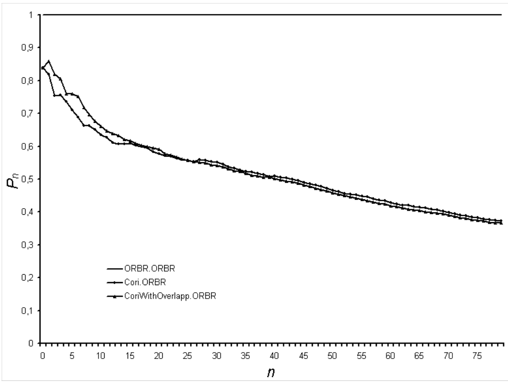
This work shows strong indication that even using one-to-one and one-to-all static overlapping, it is possible to improve selection results performance in uncoordinated database overlap contexts. The experimental results show that static overlapping is quite different from query overlapping and can be used only as an indicator of its value. As more variance exists on different databases overlap, the use of the overlapping information becomes more important.



(a) xx Reccal

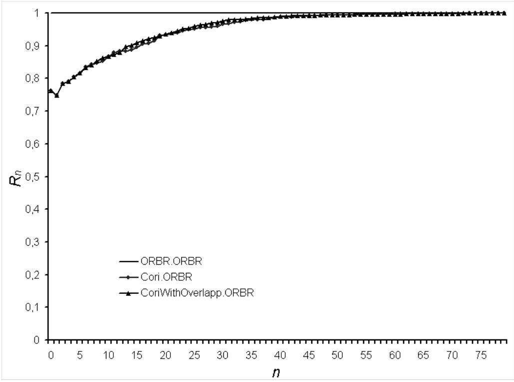


(b) xx Alternative Reccal

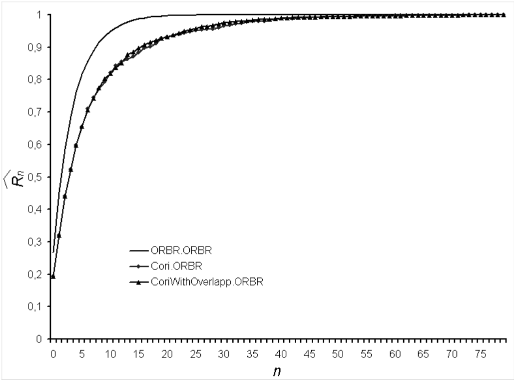


(c) xx Precision

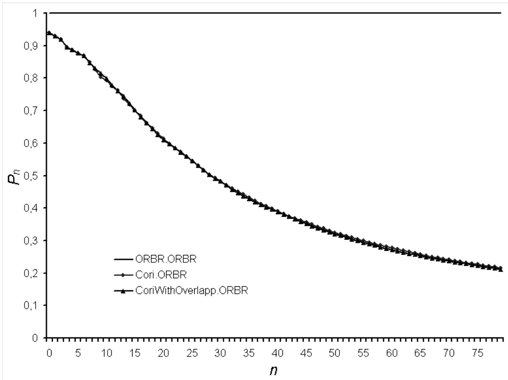
Fig. 5. Effectiveness results for webt80.xx scenario



(a) zz Reccal

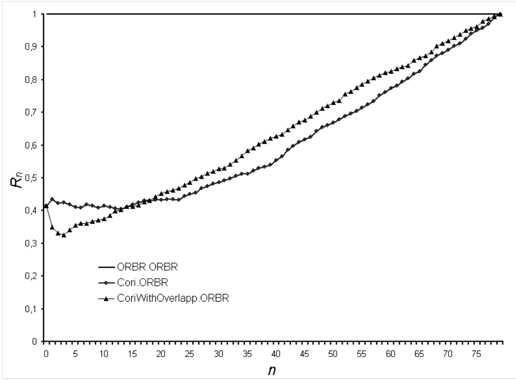


(b) zz Alternative Reccal

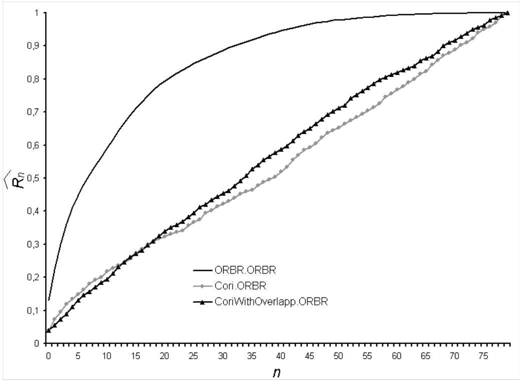


(c) zz Precision

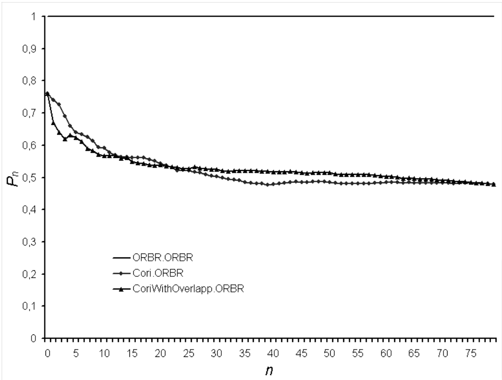
Fig. 6. Effectiveness results for webt80.zz scenario



(a) var Reccal



(b) var Alternative Reccal



(c) var Precision

Fig. 7. Effectiveness results for webt80.var scenario

With an accurate estimator for one-to-one and one-to many overlapping, computed for each query, the improvement on selection effectiveness will be made more evident.

The results presented on last section are promising and suggest that future work needed. We need to improve its effectiveness by the use of probing methods to construct a local sample database to estimate the dynamic overlapping for a given query.

References

[Bau97] Christoph Baumgarten, *A Probablistic Model for Distributed Information Retrieval*, Proc. of 20th ACM/SIGIR Conference on Research and Development in information Retrieval, Seattle, WA (Philadelphia PA, USA), 1997, pp. 258–266.

[BB98] Khrisna Bharat and Andrei Broder, *A technique for measuring the relative size and overlap of public search engines*, Computer Networks and ISDN Systems (1998), no. 30, 379–388.

[BDM95] Sergey Brin, James Davis, and Hector Garcia Molina, *Copy Detection Mechanism for Digital Documents*, Maio 1995.

[CPC+99] James C.French, Allison L. Powell, Jamie Callan, Charles L. Viles, Travis Emmit, kevin J.Prey, and You Mou, *Comparing the Performance of Database Selection Algorithms* , Proc. of 22th ACM/SIGIR Conference on Research and Devolopment in information Retrieval, Berkeley, CA, USA (Berkeley,CA, USA), 1999, pp. 238–245.

[DC98] Aiqun Du and Jamie Callan, *Probing a Collection to Discover Its Language Model*, Tech. report, CS, Univ. of Massachusets, 1998.

[FPV98] James C. French, Allison L. Powell, and Charles L. Viles, *Evaluating Database Selection Tecnhniques: A Testbed and Experiment* , Proc. of 21th ACM/SIGIR Conference on Research and Development in information Retrieval, Melbourne, Australia, Agosto 1998.

[GCGM+97] L. Gravano, K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke, *STARTS, Stanford Protocol Proposal for Internet Retrieval and Search*, Tech. report, DL Project, Stanford University, Jan 1997.

[GMT94] Luis Gravano, Hector Garcia Molina, and Anthony Tomasic, *The Efectiveness of Gloss for the Text Database Problem*, Fevereiro 1994.

[Har93] Donna Harman, *Overview of the First Text REtrieval Conference*, Proc. of 16th ACM/SIGIR Conference on Research and Development in information Retrieval, Pittsburgh, PA USA, Julho 1993.

[LCC96] Z. Lu, J. Callan, and W. B. Croft, *Applying Inference Networks to Multiple Collections Searching*, Tech. report, CS, Univ. of Massachusets, 1996.

[MCF00a] Joaquim Macedo, António Costa, and Vasco Freitas, *Fusion of Overlapped Results Sets*, Research and Advanced Technology for Digital Libraries, 4th European Conference, Lisboa, Lectures Notes in Computer Science, Setembro 2000.

[MCF00b] Joaquim Macedo, António Costa, and Vasco Freitas, *Webtrieve: A Testbed Systems for Distributed Information Retrieval* , The Proceedings of 3rd International Conference of Asian Digital Libraries, Seoul, Korea, Dezembro 2000.

[SGM94] Naraynan Shivakumar and Hector Garcia-Molina, *SCAM: A Copy Detection Mechanism for Digital Documents*, Tech. report, Electrical Engineering Department, Stanford University, 1994, Accessible by WWW.

[ZS99] Z.Lu and S.McKinley, *Partial Replica Selection Based on Relevance for Information Retrieval* , Proc. of 22th ACM/SIGIR Conference on Research and Development in information Retrieval, Seattle, WA (Berkeley,CA, USA), 1999, pp. 246–256.