Project Report

# SAMPLE SIZES FOR QUERY PROBING IN DISTRIBUTED INFORMATION RETREIVAL

December 6, 2014

Rahul Ramakrishna
University of Massachussets Amherst
Dept Of Computer Science
rahulram@cs.umass.edu

# 1 Introduction

In Distributed Information Reterival systems, information is held in separate collections, which might be in different physical locations or on separate servers. The query is first passed to a central broker. The broker then sends this query to all or some of the servers. The cost of networking during query execution seems an overhead, escepially if the query is sent to servers which dont have similar collections.

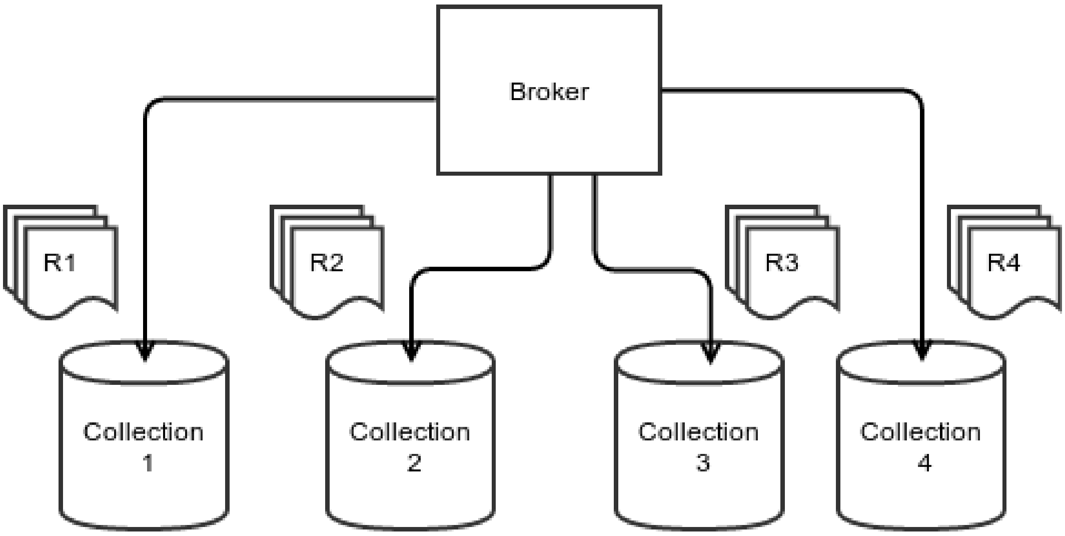

Figure 1: DIR System.

Thus, DIR has 2 major issues to be resolved.

1. Selection of a particular collection
2. Merging of Results.

In this report, we are mainly focusing on selection of collections. We are focusing on how to characterize a particular collection such that, the broker can decide on which servers to send the queries to further probe for results. We further explore 2 main sampling techniques which are used to represent collection of sets.

# 2 Query Probing

In non-cooperative environments like distributed systems. We dont get index information from each of the collections that are being fetched. Instead, its the brokers that constantly probe into the collection by sending artificial queries in random order and evaluate returned answers. The answers are called `probes` and the process is called `query` probing. For example, if we send a series of single probe queries such as ”books”, ”football”, ”ibm” and the following number of answers are returned: 1000, 200, 450 . From the results we can guess that, the documents are more likely to talk about Books and IBM rather than football.

The following is the algorithm initially proposed by Callan et al [1], which iteratively discovers the language model of the collections in non-cooperative environments.The language model is updated according to the new terms found in the retrieved documents. The next probe queries are selected from the obtained language model. Probing continues until a stopping criterion is met.

**Algorithm 1** Query Sampling

```
 1: procedure QS
 2:     Select an Initial Query Term
 3:     Run Query on the IR System
 4:     Reterive Top N Documents as Result.
 5:     Update the Resource Description based on characterstics of retreived documents.
 6: loop:
 7:     for words,freq in Documents
 8:     Update Learned Resource Descriptions.
 9:     if $StopCriteria() = Yes$ then
10:         break;
11:     else
12:         Goto 3
```

Callan et al suggested that assigning $N = 4$ and $StopCriteria = 75$ Docs i.e examining a total of 300 documents will be a good representation of the collection. The algorithm has many open ended choices which needs to be tuned. For example, how to select query terms and how to select documents to examin per query and most importantly *when to stop sampling*. In the subsequent sections we will evaluate sampling techniques mentioned by Callan et al and adaptive query probing techniques by Milad et al [2].

# 3 Sampling Techniques

## 3.1 Static Sampling using Ctf

Use sections and subsections to organize your document. LaTeX handles all the formatting and numbering automatically. Use ref and label for cross-references — this is Section 3, for example.

$$Ctf = \frac{\sum_{i\epsilon V'} ctf_i}{\sum_{i\epsilon V} ctf_i}$$

### 3.2 Adaptive Query Probing

Use `tabular` for basic tables. You can upload a figure (JPEG, PNG or PDF) using the files menu. To include it in your document, use the `includegraphics` command (see the comment below in the source code).

Cosine $tf.idf \geq \gamma$ Where $(\gamma)$ is a threshold value.

$$Recall(s, \gamma) = \frac{\text{No of Significant terms in sample}}{\text{Total No of Significant terms}}$$

$$Term = \begin{cases} Significant & \text{if } tf.idf \geq \gamma \\ \textit{Not Significant} & \text{otherwise} \end{cases}$$

### 3.3 Mathematics

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_i^n X_i$$

denote their mean. Then as $n$ approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

## 4 Measure of Effectiveness

You can make lists with automatic numbering ...

### 4.1 Comparing Ctf Ratios

### 4.2 Significant terms

### 4.3 Result Evaluation

... or bullet points ...

- Like this,
- and like this.

## 5 References

1. Query Based Sampling of Text Databases. Jamie Callan, Margaret Connell
2. Sample Sizes for Query Probing in Uncooperative Distributed Information Retrieval. Milad Shokouhi, Falk Scholer, and Justin Zobel.