Project Report

SAMPLE SIZES FOR QUERY PROBING IN DISTRIBUTED INFORMATION RETREIVAL

December 5, 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.

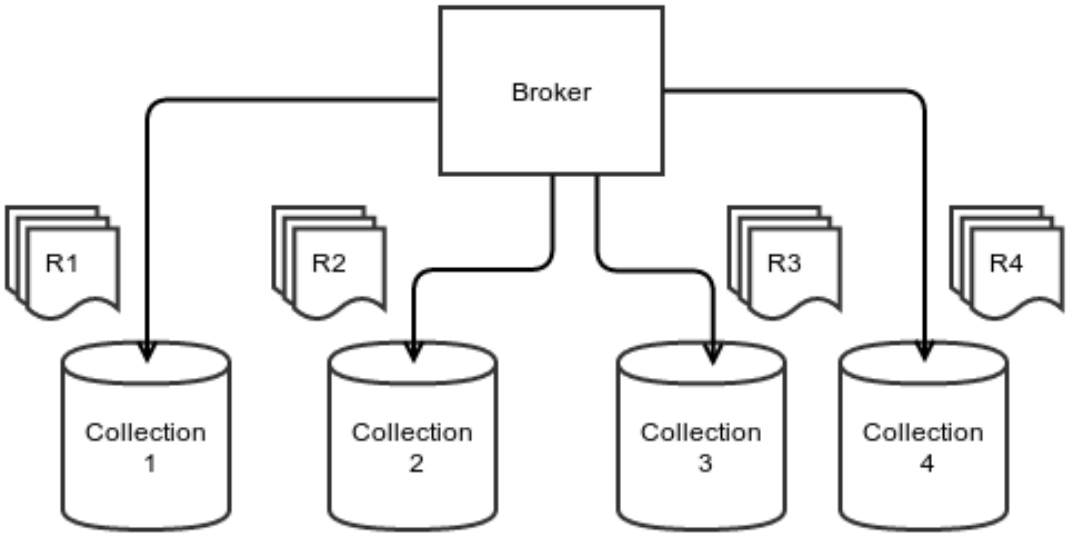

Figure 1: DIR System.

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 Sampling Techniques

2.1 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 .

2.2 Static Sampling using Ctf

Use `section`s and `subsection`s to organize your document. LaTeX handles all the formatting and numbering automatically. Use `ref` and `label` for cross-references — this is Section 2, for example.

2.3 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).

2.4 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 $\mathcal{N}(0, \sigma^2)$.

3 Measure of Effectiveness

You can make lists with automatic numbering ...

... or bullet points ...

- Like this,
- and like this.