Summary
This document reports the work done for the Yahoo! Faculty Research Grant I received in December 2007 from Yahoo!. I visited the Yahoo! Research lab in Barcelona for two weeks in October 2008. This report reports on: how we spent the grant, our research on decentralized search, a new master course on distributed data processing at the University of Twente, open source software resulting from the project, and Ph.D student interns from Twente at Yahoo.

1. The Hadoop cluster
The full grant of $25,000 (about €16,000 at the time) was spent in December 2008 on a cluster of 16 machines, that mirrors the Yahoo Hadoop infrastructure. We use the PowerEdge R200 from Dell: dual core Intel Xeon E3110 64 bit processors with 6MB Cache. Each machine has 8GB main memory (4 times 2GB DDR2), and two hard disks, one of 750 GB is used for the Hadoop distributed file system, the other, 150 GB, is used for the operating system. The machines run Suse 10.3, and Hadoop 0.19.2, see Figure 1.

Figure 1: The Yahoo Hadoop cluster at the University of Twente

2. Decentralized Search
As a direct result of the Faculty Grant, we published two papers that analyse aspects of decentralized search: 1) Result merging using (only) search engine result pages, and 2) Peer selection using so-called discriminative keys and query adaptive indexing. Merging search results from different servers is a major problem in Distributed Information Retrieval. We used Regression-SVM and Ranking-SVM which would learn a function that merges results based on information that is readily available: i.e. the ranks, titles, summaries and URLs contained in the results pages. By not downloading additional information, such as the full document,
we decrease bandwidth usage. CORI and Round Robin merging were used as our baselines; surprisingly, our results show that the SVM-methods do not improve over those baselines [3].

In the second study, we look at collection selection using highly discriminative keys and query-driven indexing as part of a distributed web search system. The approach is evaluated on different splits of the TREC WT10g corpus. Experimental results show that the approach outperforms a Dirichlet smoothing language modeling approach for collection selection, if we assume that web servers index their local content [4]. Partly as a result of this work, we received a prestigious national Vidi grant to continue the research on decentralized search.

3. Teaching
We developed a new course in the programme of the Master Computer Science at the University of Twente: Distributed Data Processing using MapReduce. The course was run from November 2009 to February 2010. Students learn to specify algorithms using map and reduce steps and to implement these algorithms using Hadoop. The course introduces Yahoo's Pig Latin as an example of high-level languages on top of MapReduce. The course consists of lectures and practical assignments. Students solved lab exercises on the Hadoop cluster of machines in order to get hands-on experience and solve real-life large-scale problems. Examples of lab exercises are: anchor text extraction, computing the web link graph, computing PageRank, and running a search experiment by linear scanning the data. The Pig Latin exercises were done on a large Netflow log. After successful completion of the course, the students were able to:

- Disect complex problems in algorithms that use map and reduce steps,
- Specify these algorithms in a functional language such as Haskell,
- Implement these algorithms using the Hadoop framework,
- Specify relational queries using Pig Latin.

![Figure 2: MIREX at SourceForge](image-url)
4. MIREX
We reimplemented the results of the student assignments to show that even a small cluster like the Yahoo Twente cluster, is able to run large scale information retrieval experiments. We extracted anchor text for 0.5 billion English web pages of the ClueWeb09 web crawl, and run a simple but effective TREC experiment [1], [2]. The source code for the experiments is open, and available via SourceForge at http://mirex.sourceforge.net (see Figure 2). The anchor texts too are available to the research community, and have been used by several researchers that use the ClueWeb09 web crawl.

5. Interns at Yahoo
Three Ph.D students from the University of Twente were interns at Yahoo! Research, Barcelona between 2007 and 2010: Henning Rode, Pavel Serdyukov, and Claudia Hauff. Their work resulted in joint papers with researchers at Yahoo! [5], [6], [7], [8], [9].

Conclusion
I am grateful and honoured to have received a Yahoo faculty grant. Many thanks to Ricardo Baeza-Yates, Hugo Zaragoza, Roelof van Zwol, Vanessa Murdock, and many others at Yahoo! Research Barcelona for their support in making the cooperation between Yahoo and the University of Twente a success.

Publications