

# Query Log Mining @ HPC-Lab

Fabrizio Silvestri  
High Performance Computing Lab  
ISTI – CNR  
Pisa, Italy

London – May 27-28, 2009

Query Log Analysis: From Research to Best Practice

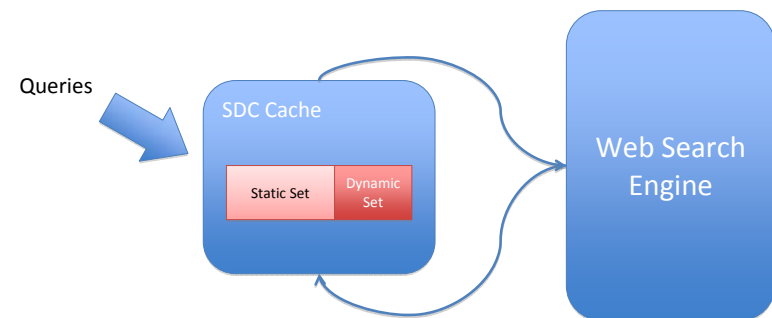
## Introduction

- HPC Lab at ISTI has a background in Parallel and Distributed Computing
- Today’s main research activities are on the following research areas:
  - High Performance Information Retrieval
  - High Performance Data Mining
  - Cloud Computing
- The main research area in which query logs have been heavily used is High Performance IR.

## Past Activities

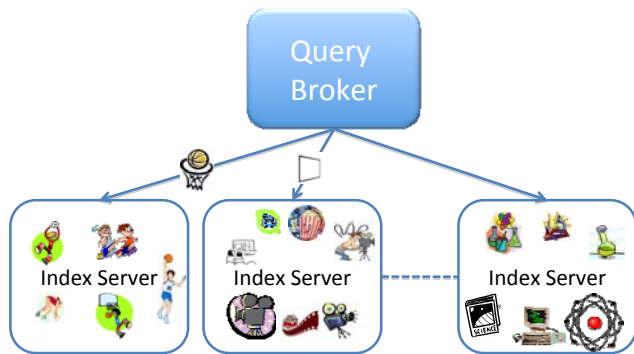
- Caching
  - SDC Policy
- Data Partitioning
  - Document Distribution, Term Distribution
- Query Routing
  - Selection of the “Best” Index Server in a distributed search engine

## SDC Caching



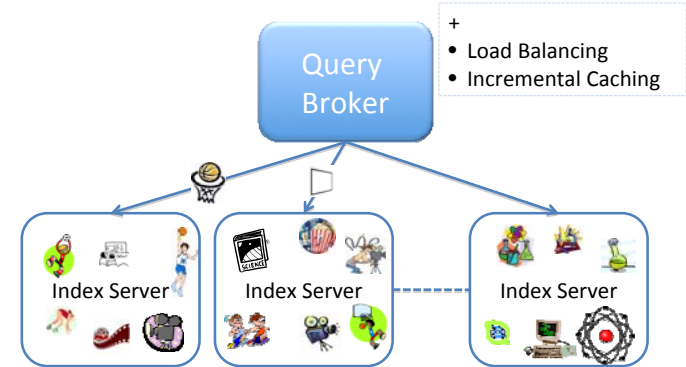
- Tiziano Fagni, Raffaele Perego, Fabrizio Silvestri, Salvatore Orlando. **Boosting the Performance of Web Search Engines: Caching and Prefetching Query Results by Exploiting Historical Usage Data.** *ACM Transactions on Information Systems (TOIS)*. 24, 1 (Jan. 2006), 51-78.
- Ricardo Baeza-Yates, Aristides Gionis, Flavio P. Junqueira, Vanessa Murdock, Vassilis Plachouras, Fabrizio Silvestri. **Design trade-offs for search engine caching.** *ACM Transactions on Web (TWEB)*. 2(4). October 2008.

## Data Partitioning



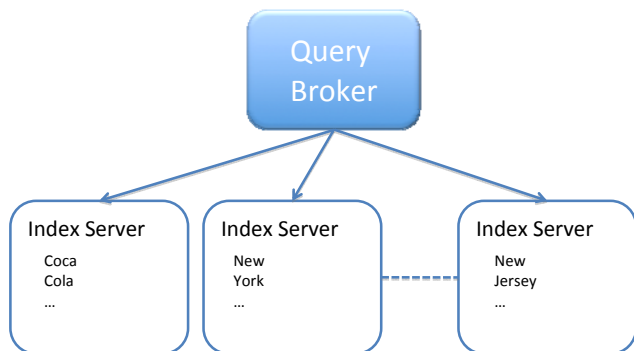
- Diego Puppini, Raffaele Perego, Fabrizio Silvestri, Ricardo Baeza-Yates. **Tuning the Capacity of Search Engines: Load-driven Routing and Incremental Caching to Reduce and Balance the Load.** To appear in *ACM Transactions on Information Systems (TOIS)*.

## Query Routing: Collection Prioritization



- Diego Puppini, Raffaele Perego, Fabrizio Silvestri, Ricardo Baeza-Yates. **Tuning the Capacity of Search Engines: Load-driven Routing and Incremental Caching to Reduce and Balance the Load.** To appear in *ACM Transactions on Information Systems (TOIS)*.

## Intelligent Term Partitioning



- Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri. **Mining Query Logs to Optimize Index Partitioning in Parallel Web Search Engines.** In *Proceedings of The 2nd International Conference on Scalable Information Systems (INFOSCALE 2007)*.

## Ongoing Activities

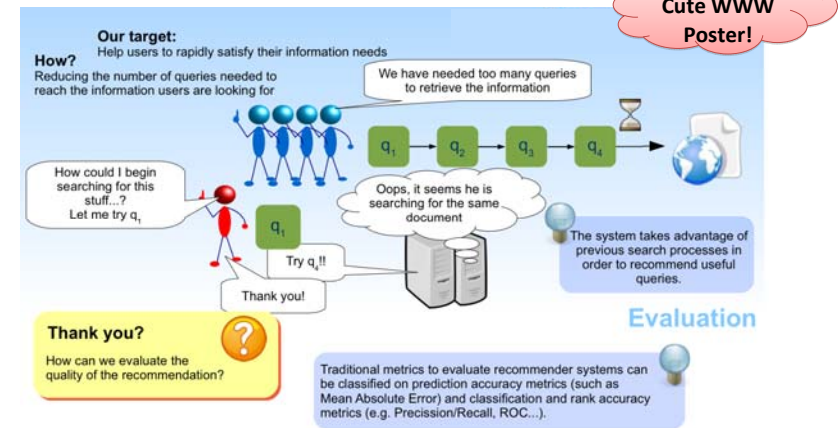
- Caching
  - Similarity-based caching
- Query Recommendation
  - Query Shortcuts
- Indexing
  - “Intelligent” (i.e. Query Log Driven) Index Organization

# Similarity Cache



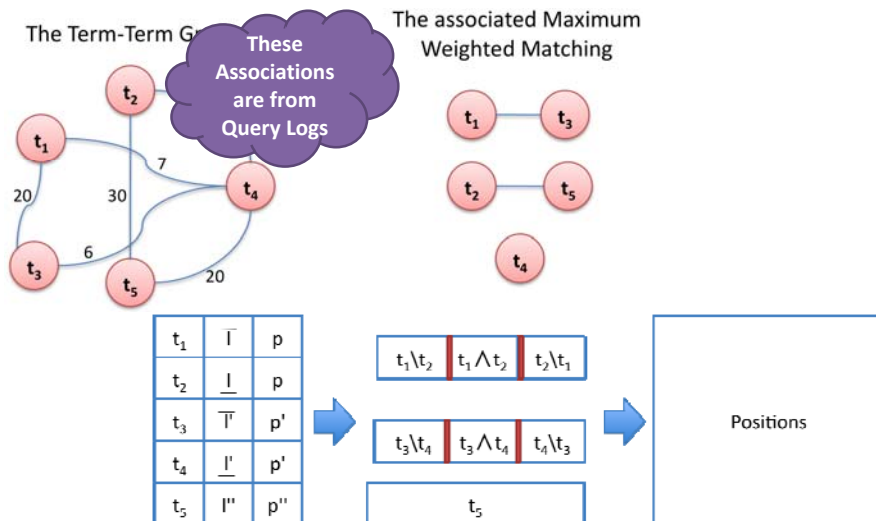
- Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fausto Rabitti. **Caching Content-based Queries for Robust and Efficient Image Retrieval**. In Proceedings of «EDBT '09: the twelfth International Conference on Extending Database Technology 2009».
- Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fausto Rabitti. **A Metric Cache for Similarity Search**. In « 6th Workshop on Large-Scale Distributed Systems for Information Retrieval ». October 26-30, 2008. Napa Valley, California, USA.

# Query Shortcuts



- Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Vreixo Formoso, Raffaele Perego and Fabrizio Silvestri. **Search Shortcuts: Driving Users Towards Their Goals**. In *Poster Proceedings of the 18 International World Wide Web Conference*. April 20th - 24th, 2009. Madrid, Spain.
- Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Vreixo Formoso, Raffaele Perego and Fabrizio Silvestri. **Search Shortcuts Using Click-Through Data**. In *Proceedings of The 1st Workshop on Web Search Click Data*, held in conjunction with WSDM 2009. February 9, 2009. Barcelona, Spain.

# “Intelligent” (i.e. Query Log Driven) Index Organization

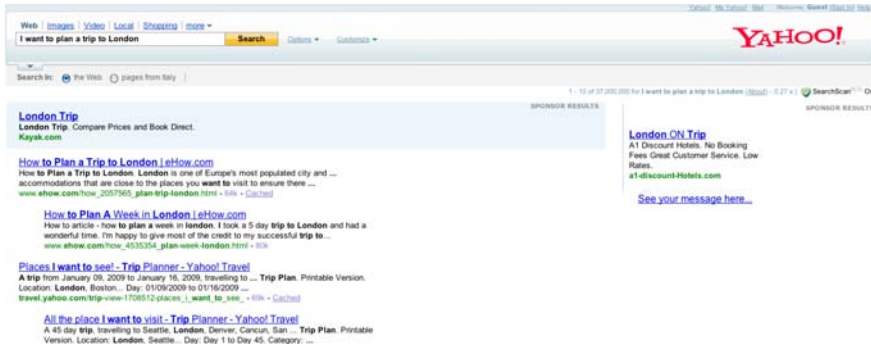


# Future/Planned Activities

- Searching for Human Activities
- ...

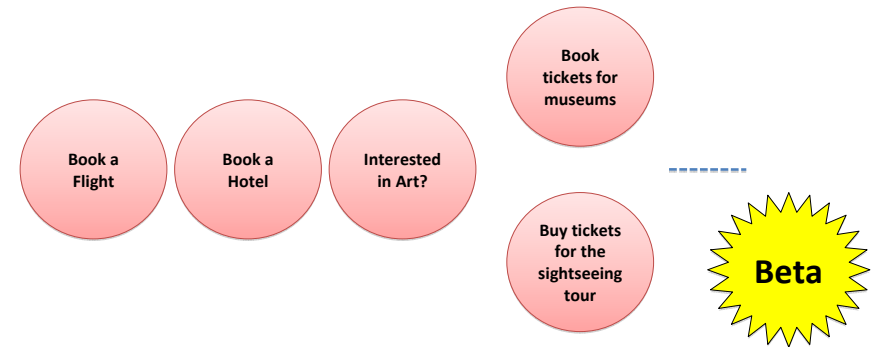
# Searching for Human Activities

- Query: I want to plan a trip to London
- Answer:



# Searching for Human Activities

- Query: I want to plan a trip to London
- Answer:



# What Kind of Query Logs?

- Large
  - Otherwise results could be biased
- Clickthrough data
  - To infer implicit feedback
- Multilingual
  - To be able to validate h
- Long term
  - Spanning multiple mon
  - information on how per
  - activities through query
- Publicly available!!!
  - Reproducible results



# Acknowledgments

- Thanks to the various people in my group that in these years has contributed to the research activity presented:
  - Ranieri Baraglia
  - Tiziano Fagni
  - Claudio Lucchese
  - Salvatore Orlando
  - Raffaele Perego
  - Diego Puppini

## Questions?



- Commercial:

Fabrizio Silvestri. **Mining Query Logs: Turning Search Usage Data into Knowledge.**

*Foundations and Trends in Information Retrieval. To Appear.*

