Mining Query Logs

(Tutorial Proposal)

Salvatore Orlando, Fabrizio Silvestri

December 22, 2008

1 Goal

Web Search Engines (WSEs) have stored in their query logs information about users since they started to operate. This information often serves many purposes. The primary focus of this tutorial is to introduce to the discipline of query log mining. We will show its foundations, by giving a unified view on the literature on query log analysis, and also present in detail the basic algorithms and techniques that could be used to extract useful knowledge from this (potentially) infinite source of information. Finally, we will discuss how the extracted knowledge can be exploited to improve different quality features of a WSE system, mainly its effectiveness and efficiency.

2 Description

Web search engines (WSEs) are queried by users to satisfy their information need. WSEs have stored in their logs information about users since they started to operate. This information often serves many purposes. The primary focus of this tutorial is to introduce to the discipline of query mining by showing its foundations and by analyzing the basic algorithms and techniques that could be used to extract useful knowledge from this (potentially) infinite source of information. We will show how search applications may benefit from this kind of analysis by analyzing popular applications of query log mining and their influence on user experience. We will conclude the tutorial by, briefly, presenting some of the most challenging current open problems in this field.

The first part of the tutorial will be devoted to introduce basic data mining techniques and tasks, such as clustering, classification, and association rules. Many of these techniques have been utilized to mine Web usage data, thus extracting from logs actionable knowledge, like patterns and models. We will, furthermore, show modern "*ad-hoc*" techniques designed to address typical problems when dealing with such an impressive amount of WSE query logs: noise removal and query result unbiasing. The first problem is related to removing those queries that do not carry too much information, the second one deals with the analysis of click data that keeps into account the position of the result clicked (i.e. it keeps into account the fact that people tends to click the first two or three results).

We will show how search applications can benefit from this kind of analysis by analyzing popular applications of query log mining and their influence on user experience. In addition, we will review some of the most recent results in this field where techniques enhancing both effectiveness and efficiency of WSE system are proposed.

Regarding effectiveness of WSEs:

> Previously submitted queries represent a very important mean for enhancing effectiveness of search systems. Query logs keep track of information regarding interaction between users and the search engine. Sessions, i.e. the sequence of queries submitted by the same users in the same period of time, can be used as a way for deriving recurring query patterns used, for instance, to

give a user query suggestions, thus improving the precision of her/his search. Click-through data is, usually, the main mean for capturing users' relevance feedback information. All in all, every single kind of user action (also, for instance, not clicking on a result) can be exploited to derive aggregate statistics which are very useful for the optimization of search engine effectiveness.

Regarding efficiency of WSEs:

> "*The scale and complexity of Web search engines, as well as the volume of queries submitted every day by users, make query logs a critical source of information to optimize precision of results and efficiency of different parts of search engines. Features such as the query distribution, the arrival time of each query, the results that users click on, are a few possible examples of information extracted form query logs. The important question to consider is : can we use, exploit, or transform this information to enable partitioning the document collection and routing queries more efficiently and effectively in distributed Web search engines?* [1]"

This means that dealing with efficiency in Web search engines is as important as it is dealing with user preferences and feedback to enhance effectiveness. Literature works show that usage patterns in WSE logs can be exploited to design effective methods for enhancing both effectiveness and efficiency in different directions.

Finally, the last part of the tutorial will, briefly, go through some of the most challenging current open problems in this field.

3 Relevance of the Topic to the IR Community

WSEs are part of the more general class of Information Retrieval (IR) systems. A search engine is, in fact, not very different from a "classical" IR system. The uncertainty in users' intent is present in WSEs as well as in IR systems. Unlike old-fashioned IR systems, though, Web IR systems can rely on the availability of a huge amount of usage information stored in query logs. Therefore, query log analysis connects to IR in many different ways. For example, the exploitation of the knowledge contained within past queries helps to improve the quality (both in terms of effectiveness and efficiency) of a WSE.

Moreover, some of the most important results, presented in important venues like SIGIR, the search track of WWW, WSDM, etc., deal with the topics covered by this tutorial.

4 Format

The tutorial will be divided into four parts. An introduction showing results on statistical and data mining analyses of user querying activities stored in query logs. The second part will be focused on how the knowledge extracted from query logs can be used to enhance the WSE effectiveness. In particular we will show techniques for query expansion, personalization and query suggestion, and applications of query log analysis to learning to rank techniques. The third part will review research works aimed at enhancing the performance of a search engine. In particular caching and partitioning techniques for distributed search engines will be reviewed. The last part of the tutorial will go through some of the most recent results in the field of query log analysis: eye-tracking-based analysis, computational advertisements, etc.

In details this is the proposed table of contents:

- Introduction:
    - The nature of Queries
    - User Actions
- Data Mining Techniques for Query Log Mining

  - “Classical” Data Mining Tasks
  - New Mining Tasks for Query Logs
    - Unbiasing the Click Distribution
    - Techniques for Removing Noise from Query Logs
- Enhancing Effectiveness of Search Systems:
  - Query Expansion
  - Query Suggestion
  - Personalized Query Results
  - Learning to Rank: ranking SVM
  - Query Spelling Correction
- Enhancing Efficiency of Search Systems:
  - Caching
  - Index Partitioning and Querying in Distributed Web Search Systems
- New Directions:
  - Eye tracking
  - Web Search Advertisement
  - Time-series Analysis of Queries

5 Presenters Biography

Salvatore Orlando is an associate professor at the Department of Computer Science, University Ca’ Foscari of Venice, and a research associate at ISTI - C.N.R. of Pisa. In 1985 he received a laurea degree cum laude in Computer Science from the University of Pisa, and a PhD in Computer Science from the same University in 1991. Then he was a post-doc fellow of the HP laboratories, and a post-doc fellow of the University of Pisa. In 1994 he joined as an assistant professor the Ca’ Foscari University of Venice, where since 2000 he has been an associate professor. His research interests include the design of efficient and scalable solutions for various data/Web mining techniques and information retrieval problems, distributed and P2P systems for information discovery, parallel and distributed systems, parallel languages and programming environments. Salvatore Orlando has published over 100 papers on international journals and conferences on several subjects, in particular on parallel processing, data and Web mining, and information retrieval. He co-chaired the 10th EuroPVM/MPI03 Conference, and the 8th SIAM Workshop on High Performance and Distributed Mining (HPDM’05). He has served on the Program Committees of many international conferences, among which Siam Data Mining Conference (SDM), European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Int. Conf. on Computational Science (ICCS), Int. Conf. on Scalable Information Systems (INFOSCALE), ACM Symposium on Applied Computing (SAC), Euro PVM/MPI, CCGRid.

Fabrizio Silvestri is currently a Researcher at ISTI - CNR in Pisa. He received his Ph.D. from the Computer Science Department of the University of Pisa in 2004. His research interests are mainly focused on Web Information Retrieval with particular focus on efficiency related problems like caching, collection partitioning, distributed IR in general.

In his professional activities Fabrizio Silvestri is member of the Program committee of many of the most important conferences in IR as well as organizer and, currently, member of the steering committee, of the workshop Large Scale and Distributed Systems for Information Retrieval (LSDS-IR). He has more than 40 publications on the field of efficiency in IR. In particular, in these last years his main research focus is on query log analysis for performance enhancement of web search engines. In the topic of the tutorial, Fabrizio Silvestri has written recently a survey paper for the journal Foundations and Trends in Information Retrieval, and has given a keynote speech at the LA-Web 2008 conference with a talk entitled "Past Searches Teach Everything: Including the Future!

References

[1] Ricardo Baeza-Yates, Carlos Castillo, Flavio Junqueira, Vassilis Plachouras, and Fabrizio Silvestri. Challenges in distributed information retrieval. In *International Conference on Data Engineering (ICDE)*, Istanbul, Turkey, April 2007. IEEE CS Press.