FACT: A Learning Based Web Query Processing System

Songting Chen[2] *Yanlei Diao*[1] *Hongjun Lu*[1] *Zengping Tian*[2]

[1]Computer Science Department
Hong Kong University of Science and Technology
Hong Kong, China
{diaoyl, luhj}@cs.ust.hk

[2]Computer Science Department
Fudan University
Shanghai, China
{stchen, zptian}@fudan.edu.cn

FACT (Fast and ACcuraTe) is a query processing system aimed at providing users with facilities so that they can get the query results from the Web in a database-like fashion. The system takes user queries in the form of keywords (free text) and returns segments of Web pages that contain the required information. It works as follows. The input from a user is passed to a general-purpose search engine to obtain a set of URLs of Web pages that may contain the required information. The system later locates the query results from the Web pages reachable from these URLs. Since queries expressed in keywords may be not able to express query requirements precisely or may not guarantee the discovery of required information inherently, the user is asked to first browse a few pages, during which the system learns from her/him about the exact query requirements and heuristics of finding the required information through a series of hyperlinks. The system will process the rest URLs and present the results in the form of segments of Web pages to the user.

We call the above approach a *learning-based* approach. As a result of the learning, the system is able to deliver to users the results that better match users' needs in a concise form. The novel features of FACT include:

(1) Different from most search engines that return URLs to users, FACT returns segments of Web pages, which releases users a great deal from the heavy burden of manually browsing pages and locating information.
(2) Though the query is posted in key words, the returned results contain exactly the information that the user is querying for, which may not be explicitly specified in the input query.
(3) The required information is often not contained in the Web pages whose URLs are returned by a search engine. FACT is capable of navigating in the neighborhood of these pages to find those that really contain the queried segments.
(4) The system does not require a prior knowledge about users such as user profiles [1] or preprocessing of Web pages such as wrapper generation [2].

A prototype system has been implemented using the approach. It learns and applies two types of knowledge, *navigation knowledge* for following hyperlinks and *classification knowledge* for queried segment identification. For learning, it supports three training strategies, namely sequential training, random training and interleaved training. *Yahoo!* is currently the external search engine. The URLs of Web pages returned by the external search engine are used in processing.

A range of queries that present different characteristics has been selected to evaluate the system. One query "*room rates of hotels in Hong Kong*" is used to represent the typical queries that have specific targets such as *price* and some indicators (e.g. $). Another query "*admission requirements on graduate applicants*" is selected to represent those queries with more general requirements such as *degree, GPA, test scores*, etc. Such results have larger variance compared with the query for prices. We even tested queries targeted at a concept. An example is "*data mining researcher*", for which the system is expected to extract segments as evidence that a person is a data mining researcher. The returned segments may talk about research interests, research projects or professional activity. It is rather subjective to determine whether a segment should be returned and effective learning is a must.

The preliminary results are encouraging. User query requirements and navigation heuristics can be reasonably well captured and stored in a rather simple form. For queries tested on the system, given a set of about 100 URLs and using interleaved training strategy, users need to browse no more than 10 of them to make the system capable of locating the queried segments or denying the irrelevant Web sites with the correctness rate higher than 80%. With respect to efficiency, to locate a queried segment, the system almost always navigates along the shortest path as human readers can find. In a Web site that does not contain the queried segment, it crawls no more than 2.5 pages to recognize the irrelevancy.

The demonstration will include the following items:
(1) A working system that can accept user queries and present segments as query results.
(2) A set of experiments that are designed to evaluate the system, and compare different implementations, such as knowledge representations and training strategies.

References

[1] R. Armstrong, D. Freitag, T. Joachims, T. Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web. In *Proc. of the 1995 AAAI Spring Symposium on Information Gathering From Heterogeneous, Distributed Environments*, March 1995.

[2] N. Ashish, C. Knoblock. Wrapper Generation for Semi-structured Internet Sources. *SIGMOD Record*, 26(4): 8-15, 1997.