KEYRY: A Keyword-based Search Engine over Structured Sources based on a Hidden Markov Model

Sonia Bergamaschi[1] , Francesco Guerra[1] , Silvia Rota[1], Yannis Velegrakis[2]

[1]Università di Modena e Reggio Emilia, Italy
[2]Università di Trento, Italy

Introduction

Keyword searching is becoming the de-facto standard for information searching, mainly due to its simplicity. The existing techniques for keyword searching over structured sources heavily rely on an a-priori instance-analysis that scans the whole data instance and constructs some index which is later used during run time to identify the parts of the database in which each keyword appears. This limits the application of these approaches to only cases where direct a-priori access to the data is possible.

Objectives

We aim to study a new technique for keyword search over relational databases that do not rely on the knowledge of the database instance, but uses only information extracted from the database schema.

Modeling Keyword search with a HMM

The matching between keywords and database terms is modeled by using a HMM, i.e. a stochastic finite state machine where the states are hidden variables. The user keywords are the observable part of the process, while the database terms are the unknown variables that have to be inferred.

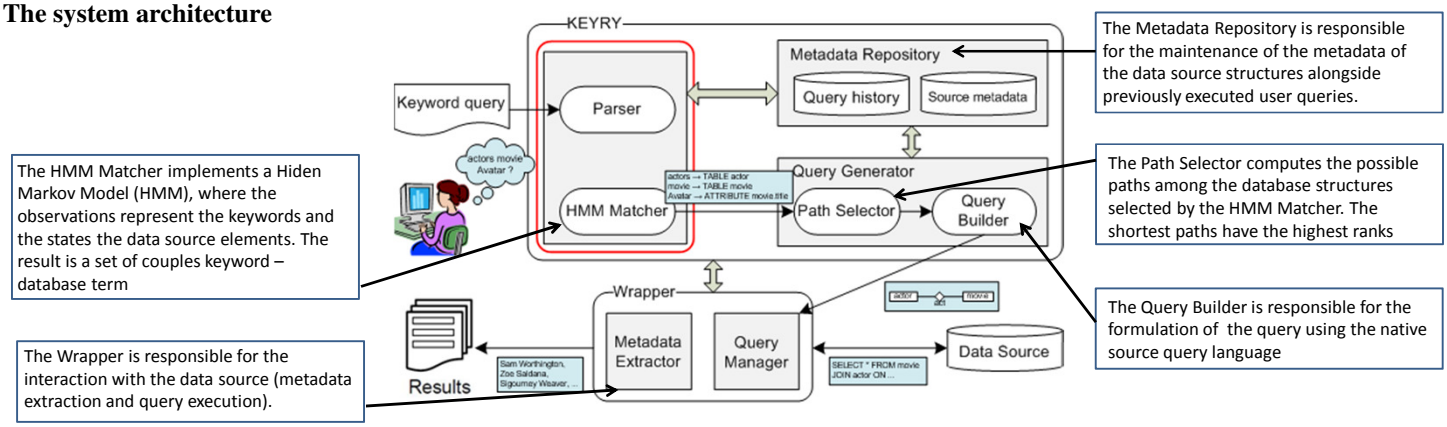
Setting the initial HMM parameters:

Transition probabilities: are computed using heuristic rules that take into account the semantic relationships that exist between the database terms (aggregation, generalization and inclusion relationships).

Emission probabilities: are computed on the basis of similarity measures.

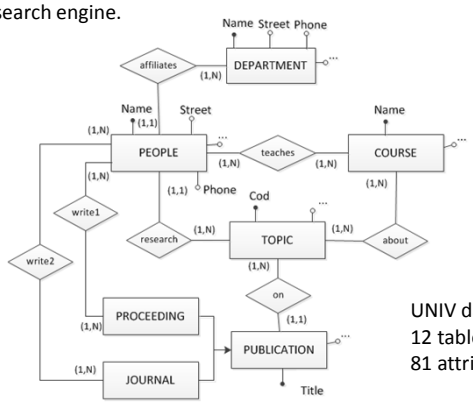
Initial state probabilities: are estimated by means of the scores provided by the HITS algorithm.

The system architecture



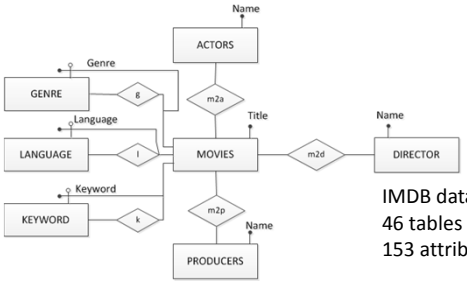
Demo Scenarios

There are two scenarios used in the demo: *IMDB, the Internet Movie DataBase*, and *UNIV,* a database used in our faculty for managing courses and professors. We will demonstrate: 1. Keyword search is possible even without prior access to the data instance; 2. Using a HMM is a successful approach in generating SQL queries, that are good approximations of the intended meaning of the user keyword queries; 3. We will illustrate how the previous queries may be used for training the search engine.



UNIV database:
12 tables
81 attributes

UNIV database: partial ER schema



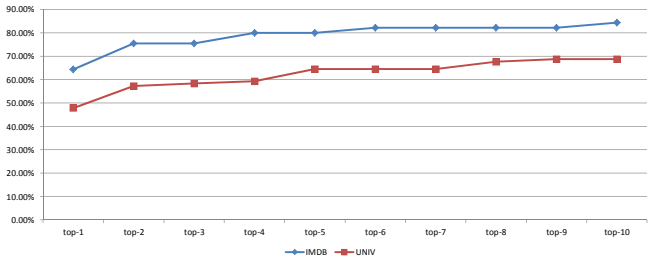
IMDB database:
46 tables
153 attributes

IMDB database: partial ER schema

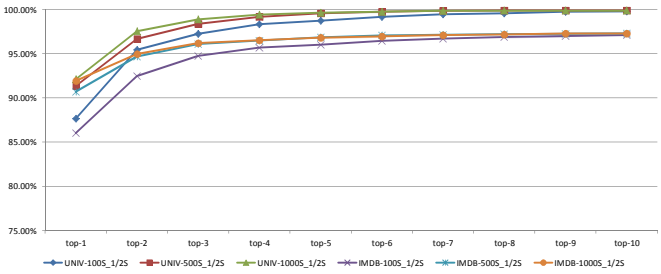
Approach effectiveness

Goal: to evaluate how many times the correct result expected by the user was part of the results returned by the system.

1. Without Training



2. With Training



Note that, in the above Figure 1, a label of a series with the format ###S_1/X means that there is a supervised initial training set of ### queries and after there is user feedback for 1 out of X queries. For the evaluation we adopted a 10-fold cross validation approach, where each fold is composed of 10000 keyword queries. See details in S. Rota, S. Bergamaschi, F. Guerra: The List Viterbi training algorithm and its application to Keyword Search over Databases, CIKM 2011.

For details: http://www.dbgroup.unimore.it/keymantic
Contact: francesco.guerra@unimore.it