

BUILDING INDICATORS FROM STATISTICS TABLES OF UNKNOWN STRUCTURE

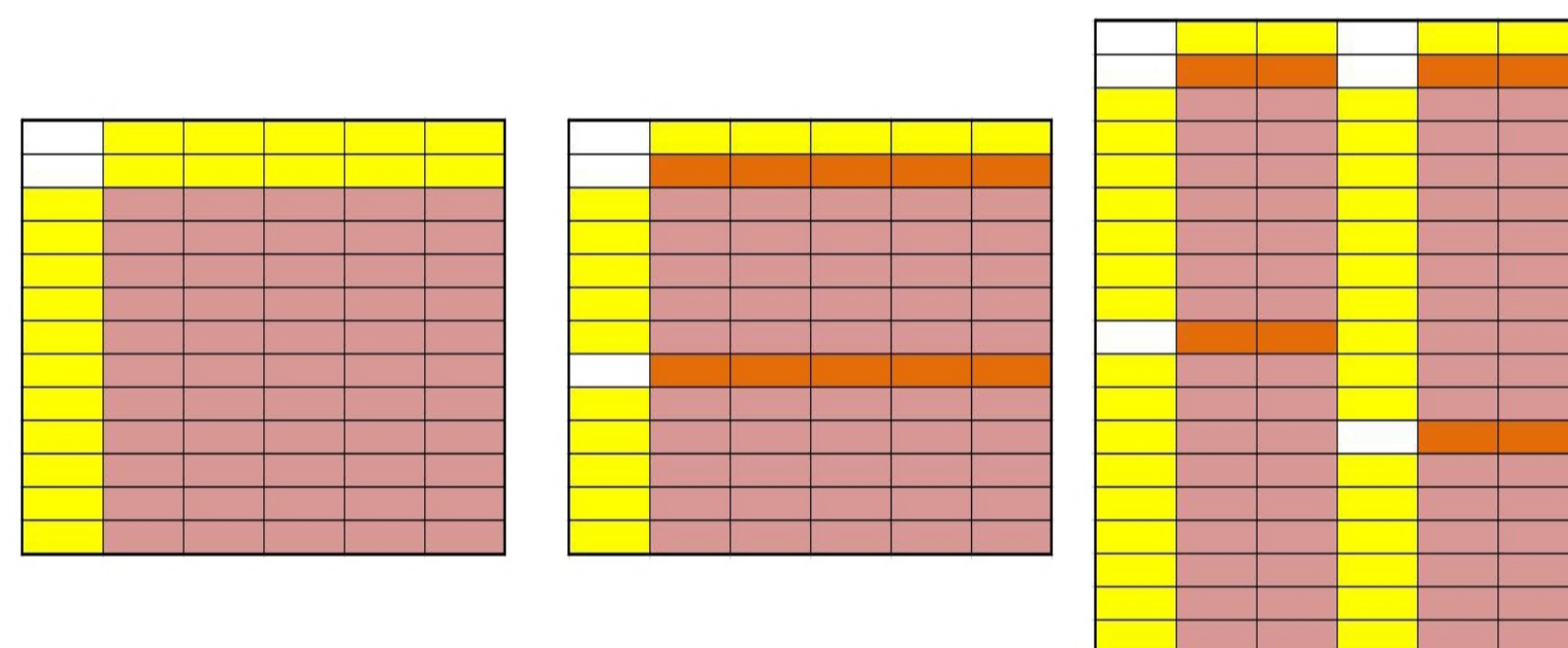
Pavel Kudinov, Dorodnicyn Computing Center of RAS, Moscow, Russia.

INTRODUCTION

At present although there are many authoritative and open sources of statistical data on the web, there is no trusted search engine capable of searching across these sources designed specifically for statistical data. Therefore there is an urgent problem of creating a system for gathering, storing and analyzing statistics from different sources. One of the core features for such system must be outputting a relevant set of indicators by users' queries with links to sources.

DIFFERENT TABLE STRUCTURES

There is no common format of a statistics table. To build an indicator it is necessary to understand underlying table structure. This problem is solved using one of incremental learning algorithms proposed in this system.



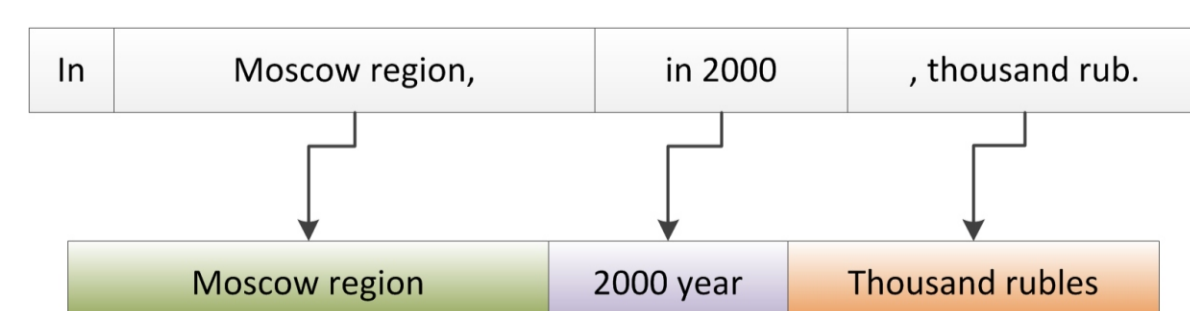
ALTERNATIVE KEYS

There are a lot of alternatives for saying similar things. If system cannot find a key from a table cell, it tries to find it using word synonyms. If there is still no success, a new key is created and there is an ability to mark this key as a synonym for another key.

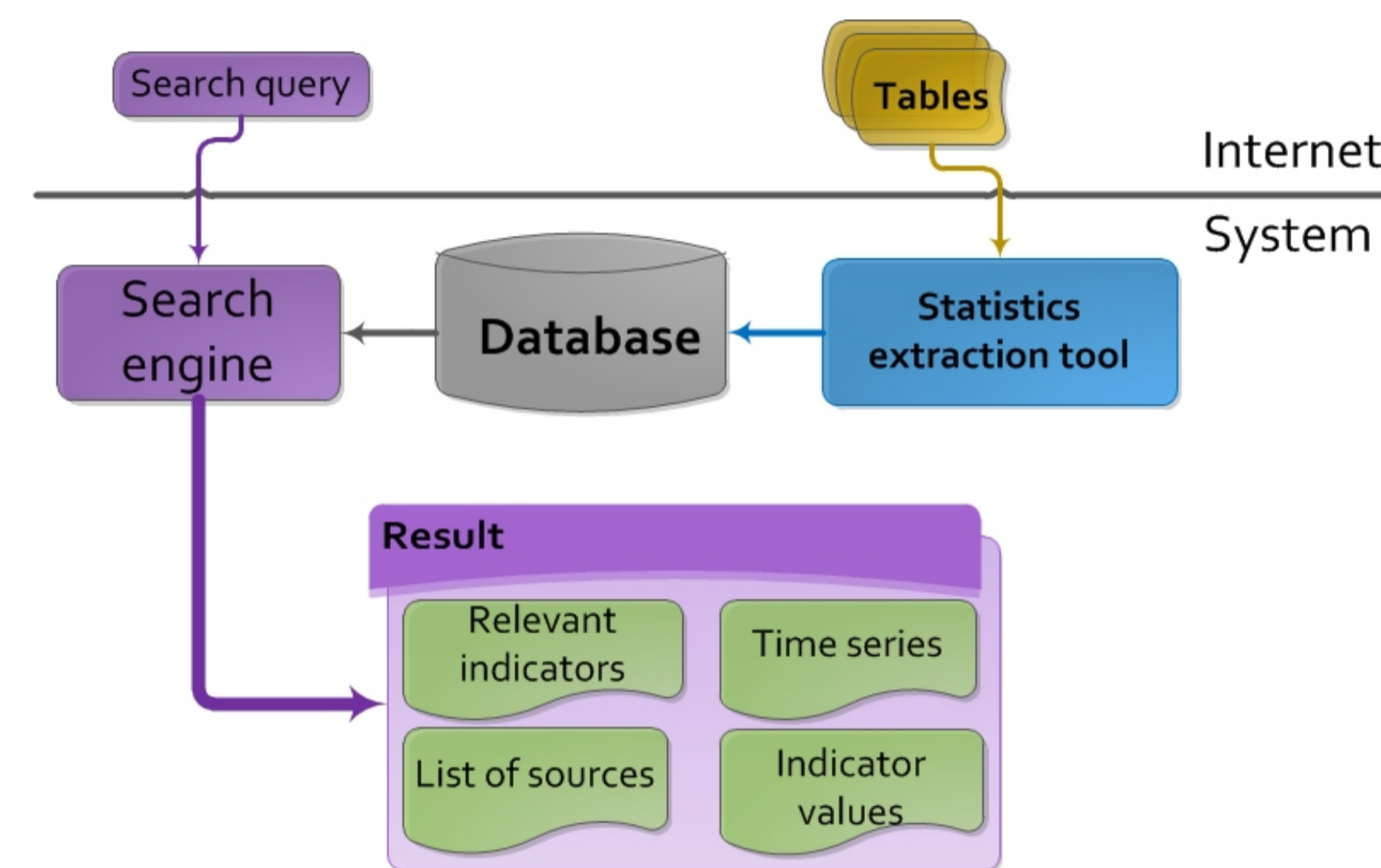


KEY SEARCH

Cells usually contain different types of information and consequently many different keys. A special search procedure is proposed to extract keys of different types from text of the cells. An algorithm for automatic extraction of unknown keys is also proposed.



SYSTEM HIGH-LEVEL ARCHITECTURE



STATISTICS EXTRACTION TOOL

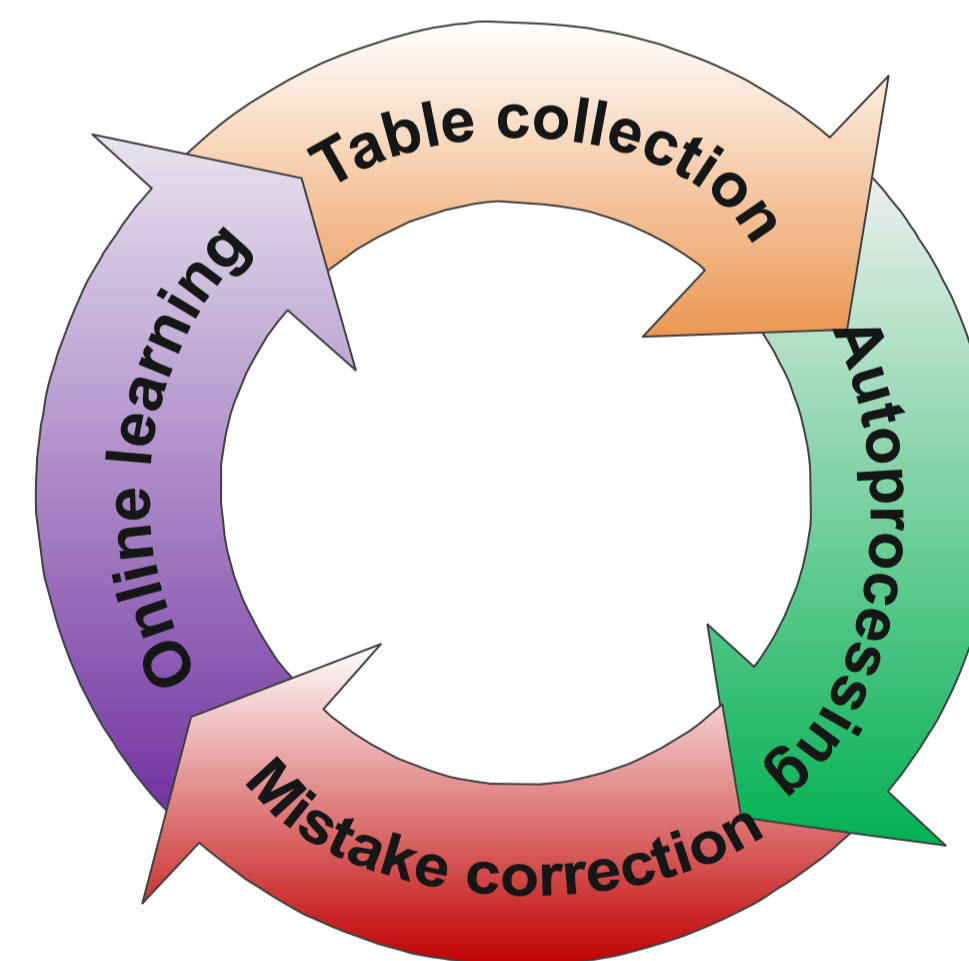


TABLE COLLECTION

There exists a rich arsenal of methods for extraction of tables from documents of widespread computer formats nowadays. Lines in tables can be both continuous and in the form of text pseudographics which was used in the epoch of printing machines. In the current research we consider marked up tables, e.g., of HTML or LaTeX format, which can be acquired using above mentioned methods from various sources including the Internet and printed collections.

MISTAKE CORRECTION

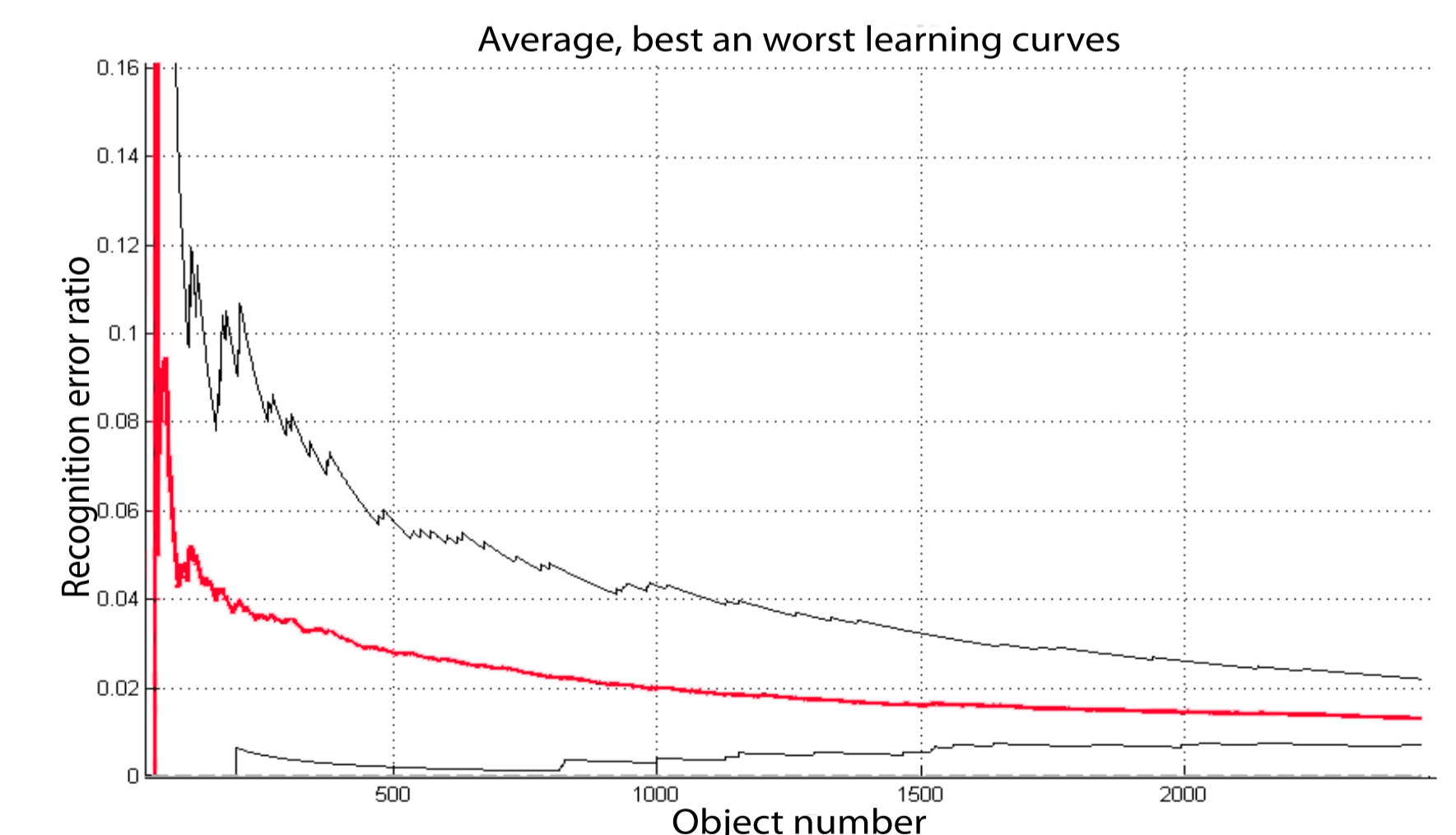
There is a special human interface that allows adding precedents to the system for each recognition problem. Learning sample is filled up on basis of these new precedents. An operator initializes a learning sample by processing manually several tables on system startup. Classification algorithms are applied for new tables after learning on small sample. After that operator browses table processing results and corrects probable mistakes. Operator's corrections are used to generate new object samples that are used for incremental online learning.

ONLINE LEARNING

Online-learning is a process of two alternating steps. The first one is classification and the second one is learning. At the first step an unknown object X is classified using existing algorithm A. Note, that this algorithm might have made a mistake. As soon as the right answer for this object is appeared, it is used to adjust (learn) this algorithm. After that we receive new version of algorithm which doesn't make mistake on object X.

Table processing consists of many steps, some of them are online-learning tasks. After a table has been processed, an operator has an ability to check and correct the results. System adjusts it's algorithms to make it possible not to make the same mistakes.

In the graphic below a learning curve of one of the learning tasks is shown. It is a decreasing curve and you can see that the recognition error ratio tends to 1%.



CONCLUSION

When creating unified repository of socio-demographic and economic data there arises a problem of its continuous replenishment by tables received from different sources and having heterogeneous structure and format. Main problems of statistics table processing are discussed and a semi-automated solution based on online learning is proposed in this work.