Astronomical spectral database of active galactic nuclei

P. Wasiewicz[a], K. Hryniewicz[b], P. Gajewski[c]

[a]Institute of Electronic Systems, Warsaw University of Technology
[b]N. Copernicus Astronomical Centre, Warsaw, Poland
[c]The Warsaw School of Information Technology

ABSTRACT

Although recent years bring massive astronomical surveys which have been extensivly searched, there are still many mysteries burried in the data. We attempt to extract objects with untypical emission lines. Especialy those with with weak and absent emission but without significant absorption. For that purpose we created database which contains quasars spectra for a quick access and peaks detection code in R environment what we describe in this article.

Keywords: active galaxies, quasars, data analysis, line identification, database, R environment, advanced sql

1. INTRODUCTION

Active galactic nuclei (AGN) are one of the most distant objects of the Universe we observe. They are so far away from us that their light travels to us for bilions of years. It means that we see snapshot of the young Universe - few hundred milions years old. Astronomical objects of that time were much more massive than nowadays galaxies in our Local Group. Researchers reveal that for high luminous quasars responsible are massive galactic cores which outshines bilions bright stars of its host galaxy. In bright galactic centres reside super-massive black holes (SMBH) with mass of dozen milions to bilions solar masses. Around black holes attracted matter forms a disc.[1] In general proces of infalling mass in a gravitational field is called accretion. Accretion onto black holes is the most efficient energy generation process (far more efficient than nuclear fission and fussion reactions). In addition densities range is so huge that it makes Universe extraordinary laboratory but the only thing we can do is to passively observe it and deduce on that basis.

So distant objects like quasars are affected by cosmological effects. Edwin Hubble observed redshift in galaxies spectra and interpreted it as escape of galaxies (see Fig 1). Whole spectrum is shifted towards longer wavelenghts. For instance light emitted as a blue change its color and could become red or even infrared. But following works have explained that redshift of distant objects is caused by expansion of the Universe which mimics escape of galaxies and essentialy makes space-time streching together with every object within it.

Nowadays, massive digital sky surveys are developed and release publicly. Those databases containing about milions of objects. As a case study we initialy focused on the SDSS quasar spectra catalogue[2] and in this paper we show preliminary results.

Our aim is to find interesting celestial objects among tens of thousands of recorded raw data. If we have knowledge of the celestial bodies with specific patterns of visible spectrum, we can discover an object property that has not yet been described.

In order to compare active galactic nuclei spectra we utilized row-oriented PostgreSQL queries together with column-oriented parallel processing of dataframes with R environment double precision accuracy.

Further author information: (Send correspondence to Piotr Wasiewicz)
P. Wasiewicz, e-mail: pwasiewi@elka.pw.edu.pl
K. Hryniewicz, e-mail: krhr@camk.edu.pl

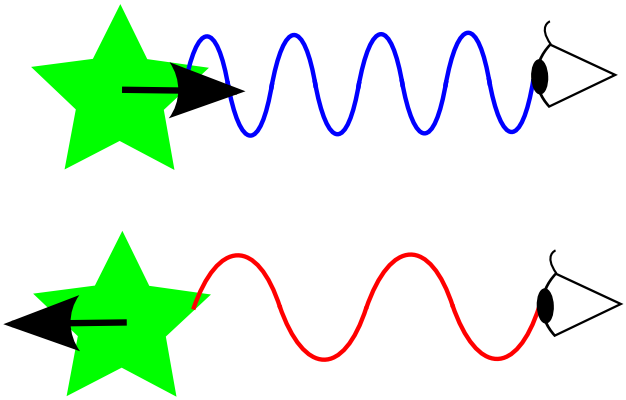

Figure 1. Kinematical redshift (lower) or blueshift (upper). Effects are caused by motion of an emitting source or an observer. In contrary to cosmological redshift which originate in stretched space-time during expansion of the Universe.

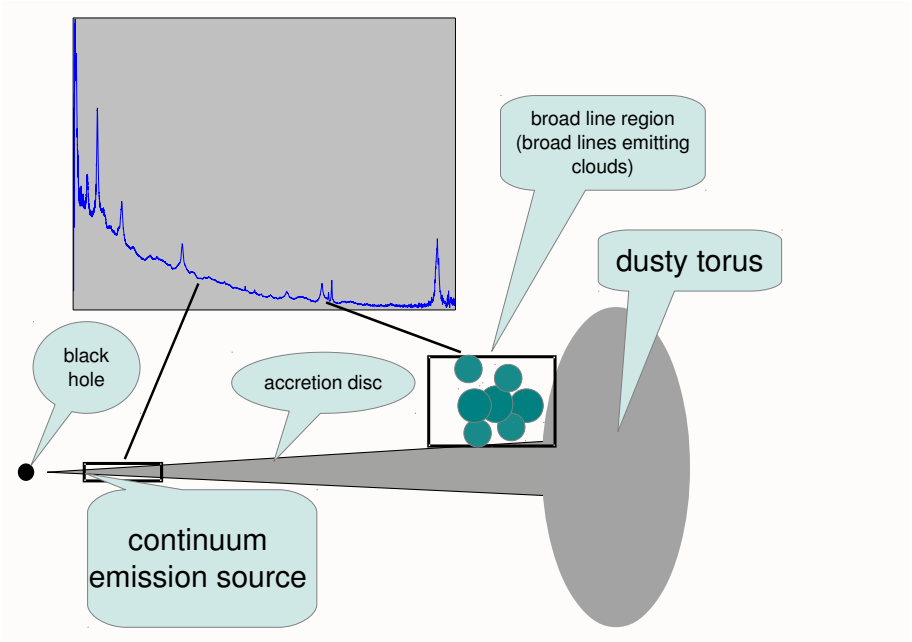

Figure 2. Quasar structure with plotted typical spectrum. Marked regions are responsible for main spectral features.

2. ACTIVE GALACTIC NUCLEI

Simple schematic of an active galactic nuclei is shown in Fig. 2 together with the mean spectra.[3] However, despite many hypotheses we still do not know exact, detailed geometry of the medium producing emission and absorption lines. It would be useful to analyse new data to test present hypotheses. Recent research in galaxy evolution suggests that AGN has important feedback on its host galaxy but also influences inter galactic medium.

As it is shown in Fig. 2 continuum emission originates in the inner parts of the accretion discs. Disc matter emits thermal radiation seen in the IR, optical and UV bands. For higher accretion rates we sometimes observe relativistic beam of matter in the form of jet which could be perpendicular to the accretion disc. On higher distances from SMBH in the accretion disc atmosphere dust grains are formed and escape from the disc in the form of dust driven wind. In this outflow broad emission lines (BEL) originates.[4]

Signature of AGNs are prominent emission lines on top of the localy power-law-like continuum emitted in the IR/optical/UV continuum bands. Continuum can have different slope because of different parameters of the AGN and its host galaxy. We can also differentiate between objects with broad (BEL) or narrow emission lines (NEL), and broad (BAL)[5] or narrow absorption lines (NAL), or objects with jets (BL Lacs).

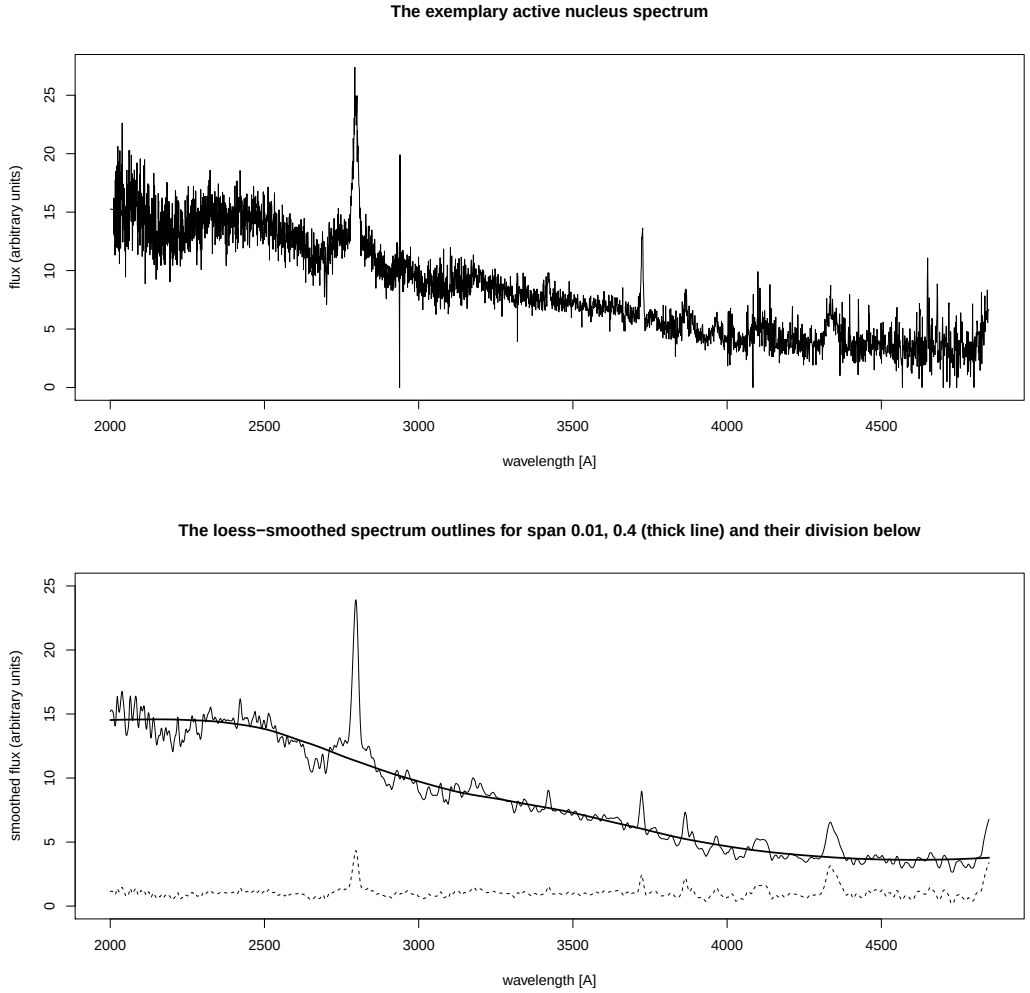

Figure 3. The active nucleus spectrum and its smoothed copies. Smoothing beside noise kills artifacts of partialy removed atmospheric emission. What left are AGN emission lines on top of pseudo continuum.

When some matter obscuring emitting source we see absorption troughs which can partialy vanishes emission lines. There are few mechanisms which makes typicaly strong lines weak in particular AGNs. That situation is interesting since allow us in principle probing extreme cases — objects in early evolution phase (very rare), quasars with high accretion rates, jet relativistic beaming, BALs. First two classes are not well described, so finding more their members together with statistical analysis is step forward in resolving their mystery.[6,7] In particular statistical aproach helps estimate lifetime of a given state or observing angle. Selection of the object can be done by identification of peaks or lack of peaks on the expected position. Additional detection or lack of detection of negative peaks confirms or exclude absorptions. Exemplary typical quasar spectra is shown in Fig. 3. On the second subfigure we illustrated process of removing pseudocontinuum made of the sum of accretion disc continuum and emission bands of hydrogen and iron. In Fig. 4 we plot emission lines peaks detection.

3. DATABASE METHODOLOGY

A database is an integrated collection of information in such a way that it can easily be accessed, managed, and updated. A Database Management System (DBMS) is a software package with computer programs, which control and manage the database. In one view, databases can be classified according to types of their content: bibliographic, full-text, numeric, and images. In another one, databases can be ordered by their orga-

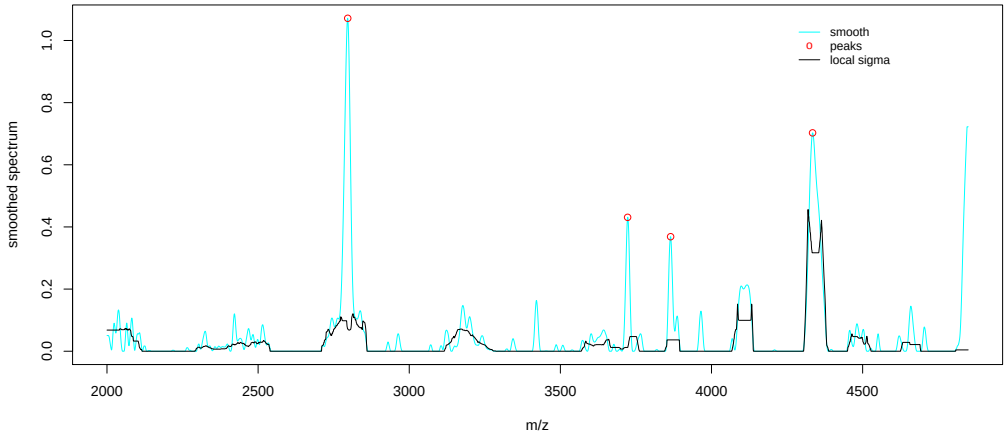

Figure 4. The active nucleus spectrum peaks detection.

nizational properties: relational row-oriented (PostgreSQL), column-oriented (MonetDB), distributed (Hbase), object-oriented.

Structured Query Language (SQL) is a database computer language designed for managing data in relational database management systems (RDBMS) and based upon relational algebra and calculus is very popular, easy to use and row-oriented.

R is a column-oriented environment[8] for statistical computing and graphics, where most calculations are made on arrays, in particular matrices and it has a well-developed, simple, effective programming language which includes e.g. conditionals, loops, user-defined functions, packages from the Comprehensive R Archive Network (CRAN) and enables an effective data handling and storage facility.

Apache Hadoop is a initiated and led by Yahoo! distributed multihost Java framework implementing Google map reduce paradigm, which enables applications written in different programming and script languages such as Java, Python, Perl, Bash to work with thousands of nodes and petabytes of data and can improve database performance by hot-plug adding the another server node to a cluster. Hbase is a distributed base and runs on top of HDFS (Hadoop Distributed Filesystem) providing fault tolerant storing huge data BigTable-like capabilities for Hadoop. Hive is a data warehouse infrastructure based directly on Hadoop with a sql-like language HiveQL. Hadoop, Hive, Hbase is an open-source panacea for data processing over 1 TB of data. If we extend the cluster to the size of several dozen typical servers (at least five hosts), computing power will be much higher than the best-designed high-end PostgreSQL machine.

4. DATABASE FIRST APPROACH

For this research the data set of 60 thousand active galactic nuclei spectra was obtained from Sloan Digital Sky Survey (SDSS) online database.[9] Every object has its typical parameters e.g. its name, its kind, z - shift of wavelengths (Fig. 1), its place in an observatory plate and additionally light spectrum of about 3500 wavelengths.

In the typical relational database such as PostgreSQL[10] there is a limit for a column number (200-2000 columns - their number depends on their type). Above 3510 parameters can not be put in table columns. Two main tables were created: the first one called object with ten basic parameters and the second one called wave for above 3500 additional parameters placed in rows with the appropriate object id number (the foreign key from the table object) and the its own index within the given object.

All operations like z-shift of wavelengths (lambda), smoothing spectrum values (val) by moving averages, counting differences between spectrum values shifted by 1, finding maximum value within the given window, and finally finding peaks and comparing them between four models and objects were calculated with the help of SQL

language e.g. sql query finding maximum values of smoothed earlier object within the given window is placed below:

```
SELECT DISTINCT p.id_object, p.id_lambda, p.val
FROM (SELECT w.id_object as id_object, w.id_lambda as id_lambda, w.val as val,
      max(neigh.val) OVER (PARTITION BY w.id_object, w.id_lambda) AS local_max
      FROM smooth_object w
      JOIN smooth_object neigh ON w.id_object = neigh.id_object
      AND neigh.id_lambda >= (w.id_lambda - 10)
      AND neigh.id_lambda <= (w.id_lambda + 10)
      ORDER BY 1,2) p
WHERE p.local_max=p.val order by 1,2;
```

After making necessary indexes in all tables (indexes for columns from queries), computation time decreased by at least 100 times e.g every 100 objects were processed in about 4 minutes (60000 objects in about 40 hours and they used 60GB disc space together with indexes). Despite still quite a long computation time, the prepared and tested SQL-based system is ready for further implementations in more advanced database-like, data warehouse environments (MonetDB, Infobright, LucidDB) with parallelized SQL queries (Hive) and huge data tables (Hbase).

5. DATABASE SECOND APPROACH

Our second database approach was made in column-oriented R environment.[8,11] Hbase main advantage is its „unlimited" column number. Before moving to Hadoop-based database we first modified our PostgreSQL base.

Connection between R and PostgreSQL was made with a help of RPostgreSQL package, which enables storing data frames of maximum 1000 columns of a numerical type double. The table object remained unchanged. The table wave was partitioned into 60 tables and changed from row-oriented to column-oriented and stores in one column one object spectrum. Thus, one wave part table contains 1000 columns with 1000 objects. Such part tables can be transfered into R data frames and treated parallely with the help of foreach package.

The original spectrum was smoothed with a help of loess.smooth built-in R function as is seen in Fig. 3. With its span parameter different outlines can be acquired e.g. with a span equal to 0.01 a smoothed moving-average-like curve is obtained, which is divided by values of regression-like outline (a span equal to 0.4). Division result is depicted as lowest outline in Fig. 3 and after subtraction of 1 its upper part above 0 is used for detecting upper peaks as is shown in Fig. 4 and its lower part below 0 absolute values for detecting lower spectrum disturbances describing BAL or NAL behaviours. Function isPeak from package PROcess was utilized in the mentioned process.

All 60000 object spectra were stored in PostgreSQL database and searched for significant peaks with a use of R functions in about 10 hours with the help of quad core processor 2.8 GHz and parallelized loops (packages doMC and foreach). The occupied disc space was decreased from 60GB to 6GB.

6. SUMMARY

We successfully developed AGN spectra comparison software environment, in the first step made of typical relational database and sophisticated SQL queries, in the second one consisted of column-oriented storage and processing in parallelized R environment. We decreased by 10 times occupied disc space (due to 1000 column tables) and by 4 times computation time (due to four core parallelization).

In future we will accelarate computing with nvidia cuda technology in one direction, with multihost hadoop environment in the second direction.

ACKNOWLEDGMENTS

KH was partly supported by grant NN203 380136 of the Polish State Committee for Scientific Research.

Funding for the Sloan Digital Sky Survey (SDSS) and SDSS-II[9] has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, and the Max Planck Society, and the Higher Education Funding Council for England.

The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, The University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington. The SDSS Web site is http://www.sdss.org/.

REFERENCES

[1] Shakura, N. I. and Sunyaev, R. A., "Black holes in binary systems. Observational appearance.," *Astronomy & Astrophysics* **24**, 337–355 (1973).

[2] Schneider, D. P., Richards, G. T., Hall, P. B., Strauss, M. A., Anderson, S. F., Boroson, T. A., Ross, N. P., Shen, Y., Brandt, W. N., Fan, X., Inada, N., Jester, S., Knapp, G. R., Krawczyk, C. M., Thakar, A. R., Vanden Berk, D. E., Voges, W., Yanny, B., York, D. G., Bahcall, N. A., Bizyaev, D., Blanton, M. R., Brewington, H., Brinkmann, J., Eisenstein, D., Frieman, J. A., Fukugita, M., Gray, J., Gunn, J. E., Hibon, P., Ivezić, Ž., Kent, S. M., Kron, R. G., Lee, M. G., Lupton, R. H., Malanushenko, E., Malanushenko, V., Oravetz, D., Pan, K., Pier, J. R., Price, T. N., Saxe, D. H., Schlegel, D. J., Simmons, A., Snedden, S. A., SubbaRao, M. U., Szalay, A. S., and Weinberg, D. H., "The Sloan Digital Sky Survey Quasar Catalog. V. Seventh Data Release," *Astronomical Journal* **139**, 2360–2373 (June 2010).

[3] Richards, G. T., Hall, P. B., Vanden Berk, D. E., Strauss, M. A., Schneider, D. P., Weinstein, M. A., Reichard, T. A., York, D. G., Knapp, G. R., Fan, X., Ivezić, Ž., Brinkmann, J., Budavári, T., Csabai, I., and Nichol, R. C., "Red and Reddened Quasars in the Sloan Digital Sky Survey," *Astronomical Journal* **126**, 1131–1147 (Sept. 2003).

[4] Czerny, B. and Hryniewicz, K., "The origin of the broad line region in active galactic nuclei," *Astronomy & Astrophysics* **525**, L8+ (Jan. 2011).

[5] DiPompeo, M. A., Brotherton, M. S., and De Breuck, C., "Very Large Telescope Spectropolarimetry of Broad Absorption Line QSOs," *Astrophysical Journal Suplement* **193**, 9–+ (Mar. 2011).

[6] Hryniewicz, K., Czerny, B., Nikołajuk, M., and Kuraszkiewicz, J., "SDSS J094533.99+100950.1 - the remarkable weak emission line quasar," *Montly Notices of the Royal Astronomical Society* **404**, 2028–2036 (June 2010).

[7] Wu, J., Brandt, W. N., Hall, P. B., Gibson, R. R., Richards, G. T., Schneider, D. P., Shemmer, O., Just, D. W., and Schmidt, S. J., "A Population of X-ray Weak Quasars: PHL 1811 Analogs at High Redshift," *ArXiv e-prints* (Apr. 2011).

[8] R Development Core Team, *R: A Language and Environment for StatisticalComputing*. R Foundation for Statistical Computing, Vienna, Austria (2010).

[9] Abazajian, K. and et al, "The seventh data release of the sloan digital sky survey," *Astrophysical Journal Supplement Series* **182**, 543–558 (June 2009).

[10] PostgreSQL Global Development Group, "Postgresql documentation." `http://www.postgresql.org/docs/` (2011).

[11] Venables, W. N. and Ripley, B. D., [*Modern applied statistics with S*], Springer-Verlag, New York, 4th ed. (2002).