Efficient Bug Triaging and Mining of Bug Report

Jyoti Yadav[1], Chetna Chouhan[2]

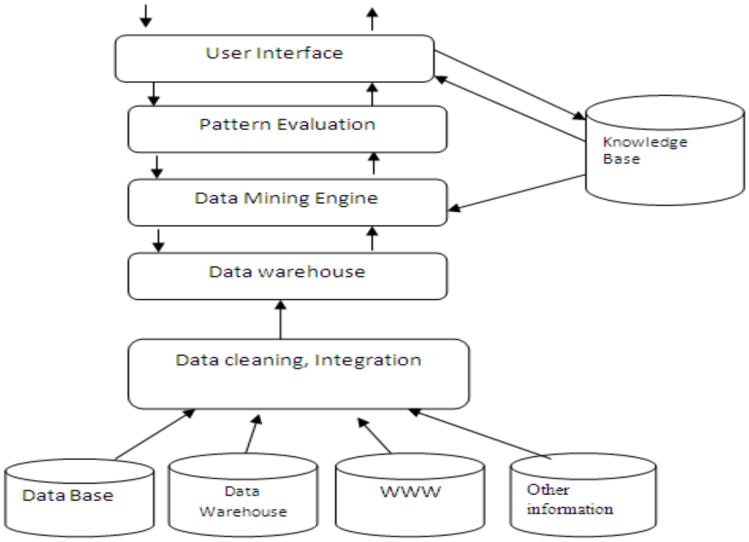
M. Tech Student, Dept of CSE, Mata Raj Kaur Institute of Engineering & Technology, Saharanwas (Rewari),
Maharishi Dayanand University, Rohtak
Assistant Professor, Dept of CSE, Mata Raj Kaur Institute of Engineering & Technology, Saharanwas (Rewari),
Maharishi Dayanand University, Rohtak

ABSTRACT

Large open source software projects receive abundant rates of submitted bug reports. Triaging these incoming reports manually is error-prone and time consuming. The goal of bug triaging is to assign potentially experienced developers to new-coming bug reports. To reduce time and cost of bug triaging, an automatic approach to predict a developer with relevant experience to solve the new coming report. we investigate the use of five term selection methods on the accuracy of bug assignment. In addition, we re-balance the load between developers based on their experience. We conduct experiments on four real datasets. The experimental results show that by selecting a small number of discriminating terms, the F-score can be significantly improved.

1. INTRODUCTION

Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. More precisely, the term refers to the application of special algorithms in a process built upon sound principles from numerous disciplines including statistics, artificial intelligence, machine learning, database science, and information retrieval. Data mining algorithms are utilized in the process of pursuits variously called data mining, knowledge mining, data driven discovery, and deductive learning Data mining techniques can be performed on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data. Some areas of specialty have a name such as KDD (knowledge discovery in databases), text mining and Web mining. Most of these specialties utilize the same basic toolset and follow the some basic process and (hopefully) yield the same product – useful knowledge that was not explicitly part of the original data set. Data mining refers to the process of finding interesting patterns in data that are not explicitly part of the data. The interesting patterns can be used to tell us something new and to make predictions. The process of data mining is composed of several steps including selecting data to analyze, preparing the data, applying the data mining algorithms, and then interpreting and evaluating the results. Sometimes the term data mining refers to the step in which the data mining algorithms are applied. This has created a fair amount of confusion in the literature. But more often the term is used to refer the entire process of finding and using interesting patterns in data. The architecture block diagram of data mining is given below



Architecture of data mining system

Technological Elements of Data Mining

Because of the inconsistent use of terminology, data mining can be called a step in the knowledge discovery process or be generalized to refer to the larger process of knowledge discovery.

Steps in Knowledge Discovery

Step 1: Task Discovery

The goals of the data mining operation must be well understood before the process begins: The analyst must know what the problem to be solved is and what the questions that need answers are. Typically, a subject specialist works with the data analyst to refine the problem to be solved as part of the task discovery step.

Step 2: Data Discovery

In this stage, the analyst and the end user determine what data they need to analyze in order to answer their questions, and then they explore the available data to see if what they need is available

Step 3: Data Selection and Cleaning

Once data has been selected, it will need to be cleaned up: missing values must be handled in a consistent way such as eliminating incomplete records, manually filling them in, entering a constant for each missing value, or estimating a value. Other data records may be complete but wrong (noisy). These noisy elements must be handled in a consistent way.

Step 4: Data Transformation

Next, the data will be transformed into a form appropriate for mining. Either we have to transform the original data, or the data are supplied in a highly structured format" The process of data transformation might include smoothing (e.g. using bin means to replace data errors), aggregation (e.g. viewing monthly data rather than daily), generalization (e.g. defining people as young, middle-aged, or old instead of by their exact age), normalization (scaling the data inside a fixed range), and attribute construction.

Step 5: Data Reduction

The data will probably need to be reduced in order to make the analysis process manageable and cost-efficient. Data reduction techniques include data cube aggregation, dimension reduction, data compression, and discretization and concept hierarchical generation.

Step 6: Discovering Patterns

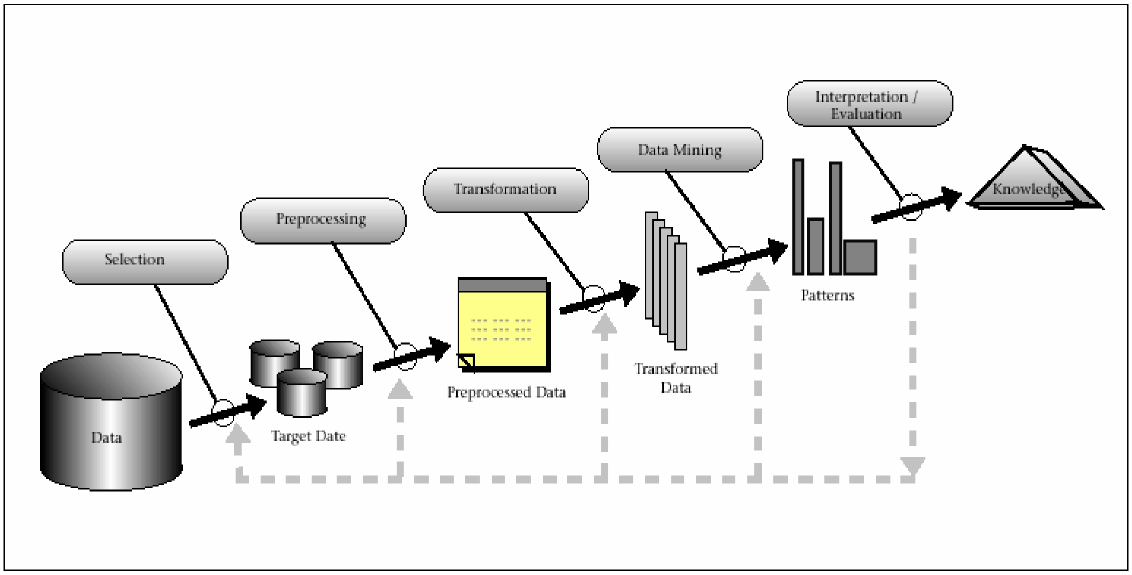
In this stage, the data is iteratively run through the data mining algorithms (see Data Mining Methods below) in an effort to find interesting and useful patterns or relationships. Often, classification and clustering algorithms are used first so that association rules can be applied.

Step 7: Result Interpretation and Visualization

It is important that the output from the data mining step can be "readily absorbed and accepted by the people who will use the results" .Tools from computer graphics and graphics design are used to present and visualize the mined output.

Step 8: Putting the Knowledge to Use

Finally, the end user must make use of the output. In addition to solving the original problem, the new knowledge can also be incorporated into new models, and the entire knowledge or data mining cycle can begin again. The whole process of KDD is given below



Overall Representation of KDD Process

Data mining methods

Common data mining methods include classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.

1. **Classification :**

Classification is composed of two steps: supervised learning of a training set of data to create a model, and then classifying the data according to the model. Some well-known classification algorithms include Bayesian Classification, decision trees, neural networks and back propagation, k-nearest neighbor classifiers, and genetic algorithm

2. **Regression:**

Regression analysis is used to make predictions based on existing data by applying formulas. Using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions.

3. **Clustering :**

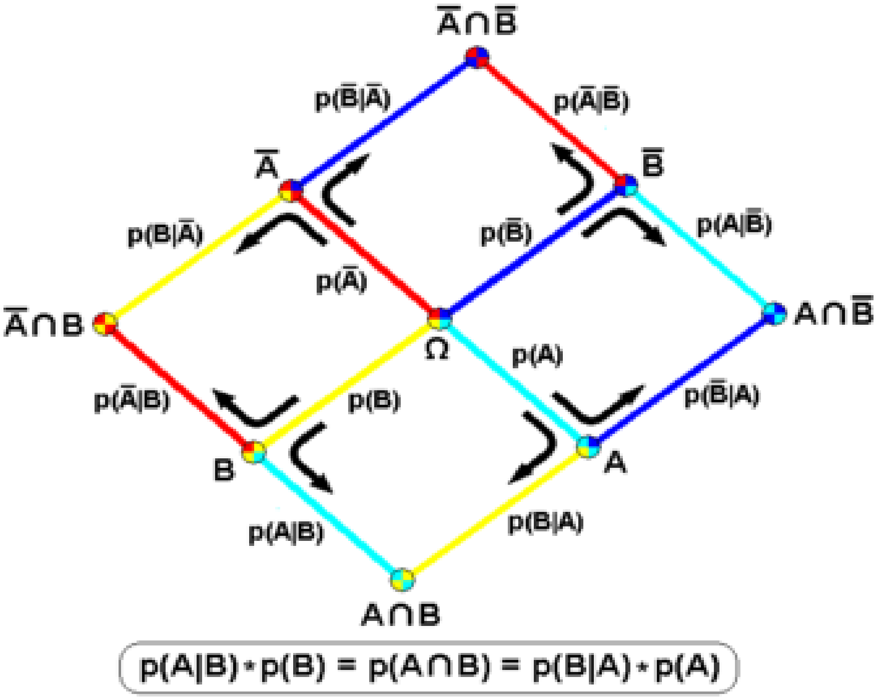
Clustering involves identifying a finite set of categories (clusters) to describe the data. The clusters can be mutually exclusive, hierarchical or overlapping. Each member of a cluster should be very similar to other members in its cluster and dissimilar to other clusters. Techniques for creating clusters include partitioning (often using the k-means algorithm) and hierarchical methods (which group objects into a tree of clusters), as well as grid, model, and density-based methods that subset. It also called characterization or generalization.

4. **Summarization :**

Summarization maps data into subsets and then applies a compact description for that subset. Also called characterization or generalization, it derives summary data from the data or extracts actual portions of the data which “succinctly characterize the contents”.

2. BAYES THEOREM

In probability theory and statistics, **Bayes' theorem** describes the probability of an event, based on conditions that might be related to the event. For example, suppose one is interested in whether Addison has cancer. Furthermore, suppose that Addison is age 65. If cancer is related to age, information about Addison's age can be used to more accurately assess his or her chance of having cancer using Bayes' Theorem.



Visualization of Bayes' theorem by superposition of two decision trees

Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

where A and B are events.

- P(A) and P(B) are the probabilities of A and B without regard to each other.
- P(A | B), a conditional probability, is the probability of A given that B is true.
- P(B | A), is the probability of B given that A is true.

3. OBJECTIVE OF PROPOSED WORK

The main purpose of research is Mining the Bug Repository perform efficient bug Triaging using Bayes theorem. The objectives of proposed work are

1. To determine and use knowledge that is enclosed in a document collection as a complete, extracting necessary information from variety of different sources and document collections.
2. Text Mining lets executives inquire questions of their text base resources rapidly extract information and find answer they never expected.
3. "Preprocessing" the text to refine the documents into structured format.
4. Reducing the results into new practical size.
5. Mining the compacted data with traditional data mining methods.

4. RESEARCH METHODOLOGY USED IN

Three main steps are used in this paper.

1. The first step is to obtain a labeled BR data set that contains textual descriptions of bugs and labels to indicate whether a BR is an SBR or An NSBR. The labeled BR Data set is required for building and evaluating our natural language predictive model.

2. The second step is text mining of bug summary to extract useful knowledge from it. Find the frequency of words in summary, remove stop words from the text summary and apply the stemming process on preprocess summary.
3. The third step is to know the type of bugs generated on the basis of the frequent terms generated from the bug data base and are used by the developers to plan the code in the future project.

In order to carry out the first step, Bayes theorem is used. Bayes theorem is based on the supervised learning. A probabilistic approach is used in this theorem and it is very simple to implement and explanation of the bayes theorem.

Obtaining Textual Description (Br/Sbr/Nsbr)

The first step is to obtain a labeled BR(Bug Report) data set that contains textual descriptions of bugs and labels to indicate whether a BR is an SBR (Security bug report) or An NSBR(not-security bug report). The labeled BR Data set is required for building and evaluating our natural language predictive model.

Textual Description	SITE1[1]	SITE2[2]	SITE 3[3]	SITE 4[4]
BR[1]	Textual Description[1,1]	Textual Description[1,2]	Textual Description[1,3]	Textual Description[1,4]
SBR[2]	Textual Description[2,1]	Textual Description[2,2]	Textual Description[2,3]	Textual Description[2,4]
NSBR[3]	Textual Description[3,1]	Textual Description[3,2]	Textual Description[3,3]	Textual Description[3,4]

BR,SBR,NSBR report containing textual description

Obtaining Textual Frequency Count (Br/Sbr/Nsbr)

The second step is text mining of bug summary to extract useful knowledge from it. Find the frequency of words in summary, remove stop words from the text summary and apply the stemming process on preprocess summary.

Frequency word	SITE1[1]	SITE2[2]	SITE 3[3]	SITE 4[4]
BR[1]	Frequency of words [1,1]	Frequency of words [1,2]	Frequency of words [1,3]	Frequency of words [1,4]
SBR[2]	Frequency of words [2,1]	Frequency of words [2,2]	Frequency of words [2,3]	Frequency of words [2,4]
NSBR[3]	Frequency of words [3,1]	Frequency of words [3,2]	Frequency of words [3,3]	Frequency of words [3,4]

Chart representing the structure of frequency of word count

Frequency Word	SITE1[1]	SITE2[2]	SITE 3[3]	SITE 4[4]
BR[1]	123	170	156	132
SBR[2]	100	150	130	120
NSBR[3]	126	190	160	140

Chart representing actual Frequency of word according to data

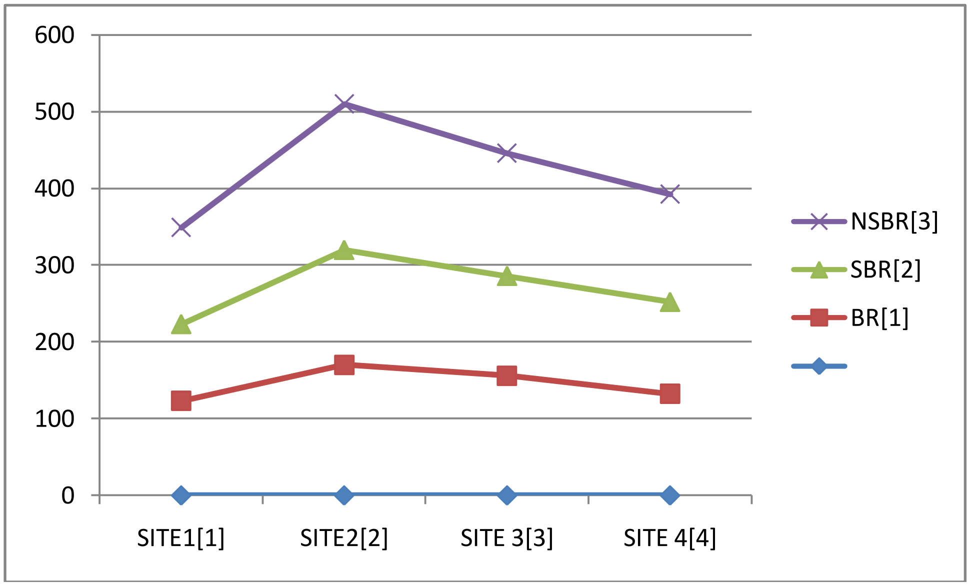


Fig . Comparative analysis of frequency count of BR,SBR,NSBR

OBTAINING TYPE OF BUGS (BR/SBR/NSBR)

The third step is to know the type of bugs generated on the basis of the frequent terms generated from the bug data base and are used by the developers to plan the code in the future project.

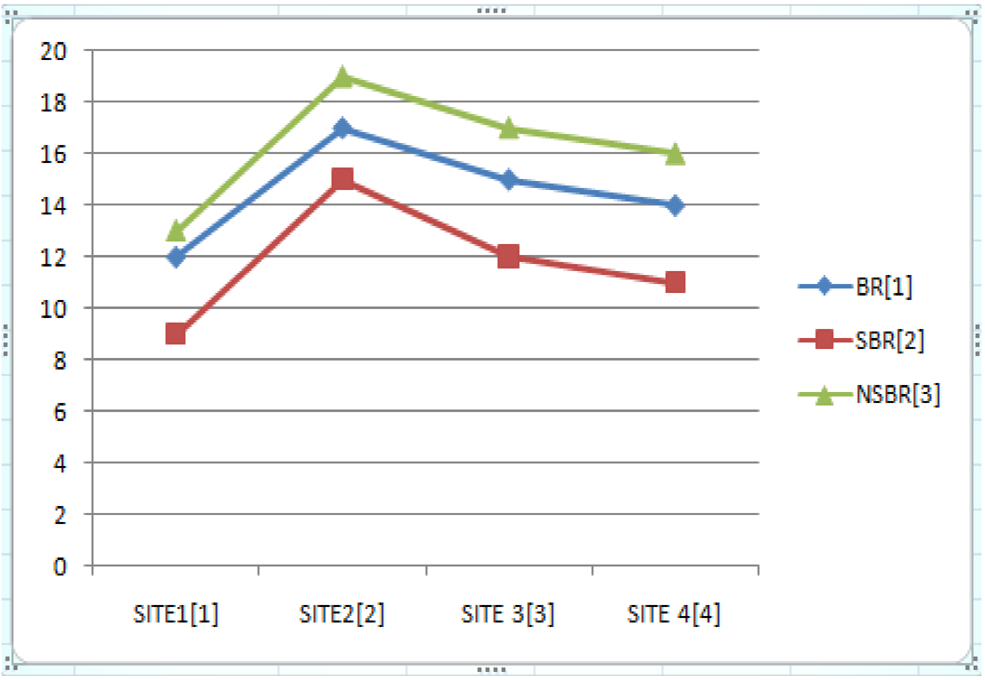
Bugs	SITE1[1]	SITE2[2]	SITE 3[3]	SITE 4[4]
BR[1]	Bugs[1,1]	Bugs [1,2]	Bugs [1,3]	Bugs [1,4]

SBR[2]	Bugs [2,1]	Bugs [2,2]	Bugs [2,3]	Bugs [2,4]
NSBR[3]	Bugs [3,1]	Bugs [3,2]	Bugs [3,3]	Bugs [3,4]

Chart representing the structure of bugs

Bugs	SITE1[1]	SITE2[2]	SITE 3[3]	SITE 4[4]
BR[1]	12	17	15	14
SBR[2]	9	15	12	11
NSBR[3]	13	19	17	16

Chart representing the bugs values



Comparative analysis of bugs count of BR,SBR,NSBR

CONCLUSION

In bug site, bug reports are organized in the form of different attachments and attachments are grouped into General, Commit, Build, Test, Fix Entries category. According to us, attachments of General category are relevant for classification purpose. General category attachments contain information which is available before the bug is analyzed, tested and fixed by the developer. General category attachments are further divided into Description, Crash info, Decode file, Event log, Email, Static analysis etc attachments.

Knowledge that is enclosed in a document collection as a complete, extracting necessary information from variety of different sources and document collections is determined and used.

Text base resources questions may inquired and using Text Mining and information is extracted rapidly that is not expected. To refine the document into structured format "Preprocessiong" of text is made. The results Reduced into new practical size. Traditional data mining methods are used to mine compacted data.

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data. Then bug information is extracted by analyzing the attachments and irrelevant attachments are discarded. For example, Static analysis and Email information etc. are discarded from General category. The information is retrieved from the bug site in html format; html tags are then removed to get individual paragraphs. Information is then statically analyzed to find some pattern for automatic feature extraction. Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behaviour and cannot be reproduced on a new sample of data and bear little use.

FUTURE SCOPE

The advent of laptops, palmtops, cell phones, and wearable computers is making ubiquitous access to large quantity of data possible. Advanced analysis of data for extracting useful knowledge is the next natural step in the world of ubiquitous computing. Accessing and analyzing data from a ubiquitous computing device offer many challenges.

UDM (UBIQUITOUS DATA MINING) introduces additional cost due to communication, computation, security, and other factors. So one of the objectives of UBIQUITOUS DATA MINING is to mine data while minimizing the cost of ubiquitous presence. Human-computer interaction is another challenging aspect of UDM. Visualizing patterns like classifiers, clusters, associations and others, in portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments. Data management in a mobile environment is also a challenging issue. Moreover, the sociological and psychological aspects of the integration between data mining technology and our lifestyle are yet to be explored. The key issues to consider include theories of UDM, advanced algorithms for mobile and distributed applications, data management issues, mark-up languages, and other data representation techniques; integration with database applications for mobile environments, architectural issues: (architecture, control, security, and communication issues), specialized mobile devices for UDM, software agents and UBIQUITOUS DATA MINING (Agent based approaches in UDM, agent interaction---cooperation, collaboration, negotiation, organizational behavior), applications of UBIQUITOUS DATA MINING (Application in business, science, engineering, medicine, and other disciplines), location management issues in UBIQUITOUS DATA MINING and technology for web-based applications of UDM.

REFERENCE

[1]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

[2]. "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[3]. Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[4]. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[5]. Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 9781558604896. Thus, data mining should habe been more appropriately named "knowledge mining from data," which is unfortunately somewhat long

[6]. OKAIRP 2005 Fall Conference, Arizona State University About.com: Datamining

[7]. Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.

[8]. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research **11**: 2533–2541. the original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons.

[9]. Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, and Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.
[10]. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, and Knowledge Discovery: An Introduction".
[11]. Data Mining. KD Nuggets. Retrieved 30 August 2012., Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
[12]. "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.
[13]. Scholar: Top publications - Data Mining & Analysis". Google Scholar.
[14]. Proceedings, International Conferences on Knowledge Discovery and Data Mining, ACM, New York.
[15]. SIGKDD Explorations, ACM, New York.
[16]. Piatetsky-Shapiro (2002) KDnuggets Methodology Poll
[17]. Shapiro (2004) KDnuggets Methodology Poll
[18]. Gregory Piatetsky-Shapiro (2007) KDnuggets Methodology Poll