Big & Deep Data Analytics using Statistical Significance: An Introductory Survey

Sourav Dutta

Databases and Information Systems
Max-Planck Institute for Informatics
Saarbrücken, Germany
Email: sdutta@mpi-inf.mpg.de

***Abstract*—The explosion of diverse and rich information sources across the world wide web has fostered the need of extremely efficient approaches for storage, management, and retrieval of such enormous data in the order of hundreds of petabytes. Scalable data mining or extraction of interesting summaries, patterns, and association rules from such huge text and sequence data-stores caters to a multitude of applications, such as search engines, financial modeling, climate monitoring, computational biology, text analysis, and social graph mining to name a few. This necessity has led to the growth of recent research directions in *big data analytics* and *deep learning*.**

Statistical significance attributes the occurrence of an event to chance alone or to the presence of an interesting phenomenon. Such techniques enable the detection of anomalies or deviations from the expected distribution, enabling faster and highly accurate approximate data mining or retrieval by quantization into "normal" or "significant" observational sub-classes. This paper provides a brief survey of interesting recent works and possible future exploratory directions incorporating statistical significance for sub-text mining (in blog analysis, spell checks, etc.), outlier detection, and graph mining in the context of big data analytics.

Keywords–statistical big data analytics; χ^2 significance; text and graph mining; clustering; survey.

I. INTRODUCTION

The surge of information sources and the explosion of data generated world-wide, catering to diverse applications, such as online transactions, financial data, climate systems, computational biology, natural language processing, and social network graphs among others, has necessitated efficient information management and retrieval. This wealth of data provides a rich opportunity to explore, study, and extract interesting user behavioral rules and interactions, natural patterns, or latent semantic models that might provide crucial operational or framework insights not only for the industry (e.g., user-interface design for new online applications, user recommendations, click/shopping behavior, and rule mining), but also for the academia (e.g., pattern analysis, behavior modeling, and algorithm design), and government (e.g., prevention of natural calamities, security, and telecommunications). The study of efficient and scalable analysis and mining of latent structures from such enormous amounts of data has led to the advent of modern research domains, like *Big Data Analytics* [1] and *Deep Learning* [2][3].

Data analytics involve a range of operations such as prediction (user rating of items [4]), extraction of latent patterns (association rules and market basket for online shops [5]), clustering (recommender systems), outlier or anomaly detection (intrusion detection [6]), etc., based on identification of relationships among objects and data observations. Mathematically, statistical significance forms the framework for establishing whether the outcome of an experiment can be ascribed to some latent phenomena affecting the system or to pure chance alone. As such, this enables the quantification of an observation as "interesting" wherein large deviations from the expected cannot be attributed to randomness alone. Detection of such statistically relevant patterns (potentially hidden) using measures such as the *p-value*, *z-score* [7], etc., within a sequence of events indicating the possible existence of hidden parameters and attributes, caters to large modern data mining applications across diverse fields of study.

In this survey paper, we introduce and discuss several state-of-the-art algorithmic approaches, in applications such as *sub-string mining* (text analytics), *motif extraction* (gene mutations in bio-informatics), approximate string matching (spell checks), subgraph mining (social network graph analytics), etc., that involve novel and efficient use of statistical significance in observations for large data analytic purposes.

Roadmap: Section II introduces a background on the measures and computation of statistical significance. Section III presents different approaches to extract statistically significant sub-sequences from an input sequence. Application of such algorithms for text and graph mining in the context of Big Data is next described in Sections IV and V. We also propose several interesting directions of future research in Section VI, while Section VII concludes the paper, followed by an extensive reference of existing literature in this domain.

II. STATISTICAL MEASURES

Statistical methods capture the degree of uniqueness of a pattern and help classify it as "significant" (or not), i.e., depicting a large deviation from the expected analysis, and also inherently take into account the *Bonferroni's Principle* [8], which informally states that the real instances of an event should be considered bogus if the number of such instances are smaller than the expected number of occurrences under a uniform distribution model. We next discuss a few popular statistical analysis measures:

 ISBN: 978-1-61208-423-7

- **p-value:** Given a sample observation O with score $S(O)$, the classical *p-value* of the observation O characterizes the probability that a random sample drawn from the same probability distribution obtains either the same or a greater score [9], i.e. in effect similar to the tail bound analysis. Formally, the underlying *null hypothesis* (H_0) states that the random sample is indeed drawn from identical probability model, while the p-value measure the chance of rejecting H_0 (based on a pre-defined significance level α). Hence, lower the p-value, less likely is it for H_0 being true and hence the observation tends to be significant. The p-value is mathematically represented by the cumulative probability distribution function (cdf) of O as:

$$p - value(O) = 1 - cdf(O) \tag{1}$$

  However, in most scenarios the probability distribution function is hard to estimate or is non-parametric, leading to the enumeration of exponential number of all possible outcomes (along with the associated scores) for accurate p-value computation, making the computation of p-value practically infeasible. To alleviate such problems, *branch-and-bound* techniques have been proposed [10] or other statistical methods are used for asymptotically approximating the p-value in large samples [11].

- **z-score:** The *z-score* or *standard score* [7][9] measures the number of standard deviations by which an observation differs from the mean or expected value under a normal distribution. It is suitable for outlier detection in applications where the data about the entire population (of all possible observations) is known apriori. Otherwise, it is referred to as the *Student's t-measure* when sample based parameters are considered. Mathematically, for an observation O,

$$Z(O) = \frac{O - \mu_O}{\sigma_O} \tag{2}$$

  where μ_O and σ_O are the mean and standard deviation of the population, respectively. The z-score operates only on the mean and variance of the data, ignoring the probability distribution curve at other points [11], thus rendering it less precise than the p-value.

- **Hotelling's T^2 measure:** The T^2 measure provides a generalization of the Student's t-measure by considering a multivariate distribution of the possible outcomes [12]. It considers the difference in the mean of different outcome populations as,

$$T^2 = n(\mathbf{x} - \mu)^T C^{-1}(\mathbf{x} - \mu) \tag{3}$$

  where n is the number of observations, $\mathbf{x}$ is a column vector of observations with corresponding mean μ, and C is the covariance matrix.

- **Log-Likelihood ratio:** The *likelihood ratio* between two models expresses how likely the data fits under one model than the other. The logarithm of this ratio, or the *log-likelihood ratio* (G^2) [9][13] essentially quantifies the deviation of the observed outcome from the expected behavior by using the theoretical distribution with k possible outcomes as,

$$G^2(O) = 2\sum_{i=1}^{k}\left(O_i \ln \frac{O_i}{E_i}\right) \tag{4}$$

  where O_i and E_i represent the observed and expected number of outcomes for the i^{th} possibility, respectively. G^2 is characterized by its *degrees of freedom*; however suffers from logarithmic instability for low (approaching 0) expected or observed values. The log-likelihood ratio can also be approximated using the *Wilk's theorem* [14], which states that as the sample size tends to infinite, the G^2 statistic becomes asymptotically χ^2 distributed (described next). Further, under no parameter assumption, the likelihood ratio demonstrates the best performance as justified by the *Neyman-Pearson lemma* [15].

- **Chi-square (χ^2) measure:** The χ^2 distribution is generally used to model the goodness-of-fit of a set of observations to the null hypothesis model. Although, for small sample sizes, the distribution tends to degenerate to a normal distribution, it provides a good approximation of the p-value in most scenarios [13]. The *Pearson's χ^2 measure* [16] uses the frequency of occurrences of categorical data to fit the observation model to that proposed by theory. The events are assumed to be independent and mutually exclusive. Similar to G^2, the *Pearson's χ^2* is defined as,

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \tag{5}$$

  The chi-square distribution is also characterized by degrees of freedom and is anti-monotonic with the p-value, i.e., larger the deviation from the expected, lower is the p-value; hence greater is the χ^2 value and more significant is the observation. Even for multinomial models, the χ^2 measure approximates the statistical importance more closely than the G^2 measure. Hence, we observe that the *chi-square statistic* provides a good approximation to the p-value by diminishing the probability of type-I errors (false positives) and is widely used to estimate the significance of an event (categorical set).

In the remainder of the paper, we present recently proposed algorithms, for data mining on enormous data stores, utilizing χ^2 and other statistical significance measures to categorize and extract interesting patterns or observations.

III. Mining Statistically Significant Strings

Consider, an automated temperature monitoring system (e.g., in industrial combustion chambers) composed of interconnected sensors or a computer server sniffing the network for possible intrusion attacks involving real-time decision based on a sequence of observed events. Such scenarios require the detection of certain "important" events pre-defined as trigger points (e.g., temperature increasing beyond threshold, etc.). However, for many data analytic settings, such as financial modeling, stock prediction, gene mutation characteristics, etc., apriori categorization of events as normal or otherwise is not

possible. Hence, significant pattern detection (using statistical significance) over a sequence of observables, like telecommunication traffic [6][17], time series transactions [18], and others [19][20][21] have been studied.

In this context, we now study the problem of extracting the statistically most significant sub-sequence from an event stream; and provide insights into the working of novel algorithms and theoretical bounds present in the literature. We later show that several other data analytic settings are systematically mapped to solving this central problem, stated formally as:

Problem Statement: Given a sequence of length l composed of event symbols s_i taken from a finite alphabet set Σ with cardinality m, let p_{σ_i} denote the associated probability of occurrence of σ_i such that $\sum p_{\sigma_i} = 1$. For θ_{σ_i} observed number of occurrence of event σ_i, we need to efficiently compute the sub-sequence demonstrating the maximum *chi-squared* value or maximum deviation from the normal.

For the remaining discussion in this section, we use the following example: Consider the input event sequence $I = \{1, 0, 0, 0, 1, 1\}$ of length $l = 6$ and alphabet set $\Sigma = \{0, 1\}$ of size $m = 2$. Assume the probability distribution of the events to be $p_0 = 0.9$ and $p_1 = 0.1$.

1) **Naïve Algorithm:** The simplest approach to identify interesting patterns involve the brute-force extraction of all the sub-sequences present in the input, trivially compute the individual χ^2 score, and finally return the *top-k* events (using a heap data structure) demonstrating the maximum chi-squared value. However, it suffers from $O(l^2 m)$ (quadratic) computation cost for sequence length l and event cardinality m, making it infeasible for real-time large data analytics.
   In our example, the sub-sequences and their corresponding χ^2 values (using Eq. (5)) are 1 (9), 0 (0.11), 10 (3.5), 11 (18), and so on and so forth.
2) **Blocking Algorithm:** To reduce the practical running of the naïve algorithm, [22] proposed partitioning the input symbols sequence into *blocks* consisting of adjacent identical symbols, and each block being replaced by only one of its symbols. Hence, our example input is modified to $I' = \{1, 0, 1\}$.
   The naïve algorithm was then executed on this "block"-ed input to obtain the top-k significant sub-sequence. Although the theoretical complexity remained the same, significant gains in run-time were shown. Interestingly, for binary symbol settings it was proven that the most significant sub-sequence starts and ends with same event symbol.
3) **Local Maxima Approach:** A combination of blocking and the use of local maxima (based on chi-square scores) was presented in [23][24]. The input events were read serially and the positions of local maxima (based on the chi-square measure) were stored as *begin* candidates. The local maxima found in our running example are $\{1\}$ and $\{0, 0, 0, 1, 1\}$. Similarly, the input sequence is reverse and the local maxima is re-computed for *end* candidates extraction ($\{1, 0, 0, 0\}$ and $\{1, 1\}$).
   The Cartesian product of the positions of the $begin$ and end candidates is then considered for finding the maximum significant locality, and corresponding χ^2 value computed. The candidate positions were conjectured to be necessary and sufficient to find the global maxima, and the subsequent pruning of the search space reduced the run-time of the procedure. Alternatively, a linear time probabilistic approach was also proposed based on the positions by treating the 2 categories of candidates separately. This $O(lm)$ algorithm was shown to attain 90% accuracy under different empirical settings; making it applicable for big data scenarios with slight error-tolerance demanding high run-time efficiency, such as transaction management, spurious web clicks, etc.
4) **Bernoulli Modeling:** Although the above approaches provided significant run-time efficiency, they suffered from a worst case complexity similar to that of the naïve algorithm. The theoretical complexity of a deterministic approach for significant sub-sequence mining was recently reduced to $O(l^{3/2} m)$ [25] with high probability. The proposed algorithm considered the events to be generated from a *memoryless Bernoulli* model and explored possible candidates with all possible lengths of sub-sequences for maximizing the χ^2 value for a chain of events.
5) **Motif Discovery:** The extraction of significant sub-structures and their interactions have also been exhaustively studied in the domain of bio-informatics for DNA sequencing, protein structure interactions, gene mutations, etc., [26][27] and is referred to as the *motif* discovery problem. Several online tools, such as $WebMOTIFS$ [28], $CompleteMotifs$ [29], etc., have also been designed to offer complete frameworks for motif discovery, scoring, analysis, and visualization. However, the underlying methodology in such methods remains the same, i.e., extracting and exploring the cause of statistically significant observations and structures. However, for modeling statistical significance of events in higher dimensional (matrices, tensors, etc.) settings prevalent in biological domains, generally the *log-likelihood* measure under a Poisson distribution is used [30].

Interestingly, the probabilities of occurrences of the different events can be considered as a combination of different distribution functions, and hence the above algorithms provide a generic framework for diverse model working scenarios. An exhaustive performance comparison of the proposed methods with real as well as synthetic datasets can be found in [25]. It was shown that the Bernoulli modeling and the Local Maxima approaches deterministically obtained the sub-string with the maximum χ^2 value with at least $3\times$ improvement in run-time. Although, the probabilistic algorithm (using Local Maxima) ran faster than the other, the accuracy was observed to vary from $80 - 90\%$ under various data inputs.

IV. Approximate Text Matching

Natural language processing (NLP) and text mining applications extract patterns of words and sentiment usage from blogs, twitter posts, articles, etc., to obtain behavioral rules of users for varied mining tasks, such as recommending product advertisements, studying the veracity of information, prevailing public sentiments, or security measures. Further, several applications involving auto-suggest, text correction, spell checks, and web search require robust approximate text matching [31]

to report documents or resources similar to the user query. Traditional methods employ *Levenshtein distance* [32] and other similarity metrics (e.g., Jaro-Winkler, cosine, etc.) to obtain the closest match, but suffer from high computation complexity – quadratic in the query length for pair-wise similarity computation – for a large dictionary of vocabulary. Several approaches for reduce the complexity involving indexing schemes [33], variable length n-grams [34], along with dynamic programming based filtering techniques [35] was proposed to partially solve the scalability challenge.

To alleviate the above problems, we now discuss a recent *approximate text matching* algorithm using statistically significant sub-sequence mining (of Section III) based on *n-grams* with 1-sliding window protocol [31]. The algorithm proposed a unique mapping of tri-grams present in the document texts onto symbols based on the degree of its matching with triplets present in the query. The similarity between two 3-grams was pre-defined into 4 hierarchical classes, and the probabilities of occurrences of the symbols correspondingly computed (assuming an independent and uniform distribution on the alphabet set). The intuition was to transform the document into a symbol sequence (based on triplet similarity) and thus closely matching words or phrases would lead to multiple adjacent trigram matches represented by high similarity symbols (having low probability) in the documents leading to a high χ^2 value. The probabilistic linear-time local maxima based sub-sequence mining approach (described in Section III) was used on the modified documents to extract the approximately matched texts with efficient run-time complexity.

For example, consider a document $D = abcdef$ and a query $Q = bcde$ with alphabet set $\Sigma = \{a, b, c, d, e, f\}$; where the triplets in D (namely, abc, bcd, cde, and def) are matched with those in Q (namely, bcd and cde). For simplicity, assume an exact match of a 3-gram in D with a triplet in Q to be represented by symbol 1, or by 0 otherwise. Hence, depicting D by similarity symbols, we obtain $D' = 0110$. Observe that the probability of exact triplet match (symbol 1) is very small, and hence the sub-sequence 11 of D' (representing $bcde$ in D) providing the highest χ^2 value is extracted as the most statistically significant string (i.e., best approximate matching to the query Q).

The proposed algorithm [31] is linear in run-time (efficiently bypassing the expensive edit distance computations) and hence provides real-time characteristics applicable to the scenario of big data. Further, it was shown to be $7\times$ faster compared to the naïve algorithm while attaining similar accuracy in results.

V. Sub-Graph Mining

The popularity of social networking communities provide large graphical network structures containing hidden or latent patterns for user-user interaction, influence, and behavior. Efficient mining of association rules from such huge network structures caters to enormous research interest in the multimedia and advertisement domains for collaborative based applications, product recommendations, etc. Similarly, analytics based on *belief propagation* [36], effect of influence, recommendations, and community detection on hugely connected graphs involve efficient and scalable sub-graph mining procedures. Analysis of computer network structures to identify security weak-points and other connectivity problems, along with road networks, etc., also involve mining of network graphs, albeit at varying operational scales. The use of graph mining is also pertinent in computational biology for detecting hidden structural patterns in protein-protein interactions and their associated effects.

Unfortunately, no polynomial time solution exists for the *graph isomorphism problem* and thus the similarity between two graph (with vertex and edge labels) for huge structures is computationally infeasible. Hence, traditional sub-graph mining involve a threshold based frequent pattern search with intelligent indexing schemes, and correspondingly approximate similarity computation to an input query [37]. Extraction and indexing of individual sub-structures of graphs such as k-length cliques for aggregated query reporting (via merging) was proposed in [38], providing a divide-and-conquer strategy using smaller sub-graphs as the working model. However, such methods involve complicated pre-processing stages and expensive merge step at query time.

The use of statistical significance for mining connected subgraphs from vertex labeled graphs was recently studied in [39]. Based on the vertex labeling (for example, discrete set of biochemical entities ranging from molecules to genes [40]), the input graph was *compressed* using rule-based edge and vertex fusion (*contracting edges*) to form a smaller super-graph. The super-graph enabled a faster run-time complexity and was shown to preserve certain properties of the original graph (such as connected sub-graphs, etc.) along with preservation of 96% of the optimal χ^2 value. For each vertex in the super-graph, its *z-score* is computed using the weighted average of the neighborhood attributes (vertex label symbols, edge weights, etc.), thus modeling the structure of the current sub-graph under consideration.

Detection of *spatial outliers* is then performed by combining the individual z-scores, and a *chi-squared* based statistical score is computed from the multi-dimension z-scores to obtain a contiguous region with high significance (i.e., connected sub-graph outlier). This approach provided a framework for generic outlier detection for vertex labeled graphs with discrete as well as continuous labels. The approach was shown to provide an analysis of statistically significant connected sub-graphs (specifically, outliers) within large social networks, such as Orkut, DBLP, etc. within 3 hours considering continuous vertex labels.

VI. Open Directions of Research

The intelligent mapping of various data mining problems to statistical significance computations in the above applications have led to a reduction in run-time with high accuracy of results, forming the basic strategy for tackling queries on huge data stores. Hence, we observe that pattern mining using statistical significance holds potential for efficiently handling Web-scale data for diverse applications. In this section, we discuss a few further directions of research involving data significance as applied to real-time mining tasks.

- **Clustering:** Clustering involves the task of grouping together items depicting similar attributes. The analysis of clusters and its use thereof for recommendation, collaborative filtering, etc., forms a basic approach in data mining and information retrieval. However, certain scenarios such as relief-help distribution, traffic

 ISBN: 978-1-61208-423-7

congestion, social community popularity, etc., require detection of only the top-k clusters based on cardinality. They depict the most "crucial" areas and help resource concentration for better management. Hence, end-to-end clustering in such scenarios provides an inefficient approach.
However, the modeling of search space into k-dimensional matrix structures, and corresponding mapping of data points (represented by symbols with associated probability of occurrence) onto the cells for statistical significance computation (where more symbols generate more significance) might provide an alternative shortcut to intelligently tackle such huge data volumes. Further, the early and efficient identification of the most populous clusters and their analysis with *centrality measures*, such as *Katz centrality* [41] might help in faster epidemic control. The real-time nature of such approaches would help combat decision delays in situations of calamity.

- **Sub-graph Matching:** The problem of sub-graph isomorphism search has myriad applications for graph classification, electronics circuits, and protein interactions to name a few. However, finding sub-graph isomorphism is NP-hard; leading to the proposal of pruning-rule based approaches [42] and *combine and permute* indexing strategies [43] for approximate sub-graph matching to a query graph.
Similar to the approximate text matching, neighboring edge and vertex locality based sub-graph matching using χ^2 significance score might be performed by mapping vertices to the symbols based on their degree of similarity to structures in the query graph. This would enable efficient approximate sub-graph isomorphism for analyzing social community and other huge network graphs. Generalizing such approaches to graph mining and connected component association provides further research interest.

- **Skyline Queries:** Skyline involves the ranking of search results based on user-define preferences using the *Pareto dominance* criteria. However, the computation of relationship for every item-item pair provides a bottleneck for scalability of such methods. Hence, the expensive computation of skyline queries have been reduced by a number of caching approaches and efficient indexing structures, such as *closed skycubes* [44]. However, similar to the clustering approach, the encoding of data points in each dimension (user preference) to matching symbols, and the corresponding significance computation promises to capture the data points respecting the user constraints (at least in most of the specified preferences). Use of pruning mechanisms based on the significance score and extraction of top-k results might provide significant run-time and storage improvements in such scenarios.

Additionally, theoretical analysis of algorithms, under the ambit of statistical significance testing approach, to derive performance bounds for varying probability distributions of symbols also provides a pertinent area of future research across different communities.

VII. Conclusion

This paper presented an introductory survey of recent algorithmic trends in the applicability of classical *statistical significance* testing for the domain of big data analytics and deep mining from varied and huge data sources. We initially provide a brief background of the statistical measures commonly used and then discuss state-of-the-art approaches based on the Pearson's χ^2 measure (and others) to efficiently solve graph mining, text analysis, and approximate matching problems, among others. The use of statistical significance for mining tasks to extract interesting patterns and rules across diverse domains such as computational biology, social networks, etc., has been shown to provide enhanced accuracy, run-time, and scalability performance compared to state-of-the-art methods. We also enumerated a few possible exciting further research directions involving graph isomorphism and clustering, based on the statistical significance of observations.

Acknowledgment

The author would like to thank the *Google European Doctoral Fellowship* for financially supporting this work.

References

[1] J. Manyika et al., "Big Data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., June 2011.

[2] J. Dean, "Large Scale Deep Learning," Keynote at CIKM (research.google.com/people/jeff/CIKM-keynote-Nov2014.pdf), 2014, retrieved: June 7, 2015.

[3] L. Deng and D. Yu, "Deep Learning: Methods and Trends," Foundations and Trends in Signal Processing, vol. 7, no. 3-4, 2013/14, pp. 197–387.

[4] A. Moreno et al., "Hybrid Model Rating Prediction with Linked Open Data for Recommender Systems," Communications in Computer and Information Science, vol. 475, 2014, pp. 193–198.

[5] H. Aguinis, L. Forcum, and H. Joo, "Using Market Basket Analysis in Management Research," Journal of Management, vol. 39, no. 7, 2013, pp. 1799–1824.

[6] N. Ye and Q. Chen, "An anomaly detection technique based on chi-square statistics for detecting intrusions into information systems," Quality and Reliability Engineering International, vol. 17, no. 2, 2001, pp. 105–112.

[7] M. Regnier and M. Vandenbogaert, "Comparison of statistical significance criteria," Journal of Bioinformatics and Computational Biology, vol. 4, no. 2, 2006, pp. 537–551.

[8] J. Leskovec, A. Rajaraman, and J. Ullman, Mining of Massive Datasets. Infolab: Stanford University (http://www.mmds.org/), 2015, retrieved on: June 7, 2015.

[9] T. Read and N. Cressie, Goodness-of-fit statistics for discrete multivariate data. Springer, 1988.

[10] G. Bejerano, N. Friedman, and N. Tishby, "Efficient exact p-value computation for small sample, sparse and surprisingly categorical data," Journal of Computational Biology, vol. 11, no. 5, 2004, pp. 867–886.

[11] S. Rahmann, "Dynamic programming algorithms for two statistical problems in computational biology," in Workshop on Algorithms in Bioinformatics (WABI), 2003, pp. 151–164.

[12] H. Hotelling, "Multivariate quality control," Techniques of Statistical Analysis, vol. 54, 1947, pp. 111–184.

[13] T. Read and N. Cressie, "Pearson's χ^2 and the likelihood ratio statistic G^2: a comparative review," International Statistical Review, vol. 57, no. 1, 1989, pp. 19–43.

[14] S. S. Wilks, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," The Annals of Mathematical Statistics, vol. 9, 1938, pp. 60–62.

[15] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 231, no. 694706, 1933, pp. 289–337.

 ISBN: 978-1-61208-423-7

[16] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philosophical Magazine Series 5, vol. 50, no. 302, 1900, pp. 157–175.

[17] R. Goonatilake, A. Herath, S. Herath, S. Herath, and J. Herath, "Intrusion detection using the chi-square goodness-of-fit test for information assurance, network, forensics and software security," Journal of Computing Sciences, vol. 23, no. 1, 2007, pp. 255–263.

[18] R. Povinelli, "Identifying temporal patterns for characterization and prediction of financial time series events," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 2, 2003, pp. 339–352.

[19] M. Kaboudan, "Genetic programming prediction of stock prices," Computational Economics, vol. 16, no. 3, 2000, pp. 207–236.

[20] A. Denise, M. Regnier, and M. Vandenbogaert, "Assessing the statistical significance of overrepresented oligonucleotides," in WABI, 2001, pp. 537–552.

[21] I. Kuznetsov and S. Rackovsky, "Identification of non-random patterns in structural and mutational data: the case of Prion protein," in CSB, 2003, pp. 604–608.

[22] S. Agarwal, "On Finding the most statistically significant substring using the chi-square measure," Master's thesis, Indian Institute of Technology, Kanpur, 2009.

[23] S. Dutta and A. Bhattacharya, "Most Significant Substring Mining Based on Chi-Square Measure," in PAKDD, 2010, pp. 319–327.

[24] A. Bhattacharya and S. Dutta, "Mining Statistically Significant Substrings Based on the Chi-Square Measure," in Pattern Discovery and Sequence Mining: Applications and Studies. IGI Global, 2011.

[25] M. Sachan and A. Bhattacharya, "Mining Statistically Significant Substrings using the Chi-Square Statistic," VLDB, vol. 5, no. 10, 2012, pp. 1052–1063.

[26] P. Ng, "Statistical Significance for DNA Motif Discovery," Ph.D. dissertation, Cornell University, 2011.

[27] D. Lovell, "Biological Importance and Statistical Significance," Journal of Agricultural and Food Chemistry, vol. 61, no. 35, 2013, pp. 8340–8348.

[28] K. A. Romer, G. R. Kayombya, and E. Fraenkel, "WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches," Nucleic Acids Research, vol. 35, 2007, pp. 17–20.

[29] L. Kuttippurathu et al., "CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments," Bioinformatics, vol. 27, no. 5, 2011, pp. 715–717.

[30] M. Frith, J. Spouge, U. Hansen, and Z. Weng, "Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences," Nucleic Acids Research, vol. 30, no. 14, 2002, pp. 3214–3224.

[31] S. Dutta, "MIST: Top-k Approximate Sub-String Mining using Triplet Statistical Significance," in ECIR, 2015.

[32] V. Levenshtein, "Binary Codes capable of Correcting Deletions, Insertions, and Reversals," Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707–710.

[33] D. Fenz, D. Lange, A. Rheinlnder, F. Naumann, and U. Leser, "Efficient Similarity Search in Very Large String Sets," in Springer, 2012, pp. 262–279.

[34] C. Li, B. Wang, and X. Yang, "VGRAM: Improving Performance of Approximate Queries on String Collections using Variable-length Grams," in VLDB, 2007, pp. 303–314.

[35] D. Deng, G. Li, J. Feng, and W. Li, "Top-k string similarity search with edit-distance constraints," in ICDE, 2013, pp. 925–936.

[36] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding Belief Propagation and Its Generalizations," in Exploring Artificial Intelligence in the New Millennium. Morgan Kaufmann Publishers Inc., 2003, pp. 239–269.

[37] C. Jiang, F. Coenen, and M. Zito, "A Survey of Frequent Subgraph Mining Algorithms," The Knowledge Engineering Review, vol. 28, no. 1, 2013, pp. 75–105.

[38] L. Zhu, W. Ng, and C. J., "Structure and Attribute index for approximate graph matching in large graphs," Information Systems, vol. 36, 2011, pp. 958–972.

[39] A. Arora, M. Sachan, and A. Bhattacharya, "Mining Statistically Significant Connected Subgraphs in Vertex Labeled Graphs," in SIGMOD, 2014, pp. 1003–1014.

[40] C. You, L. Holder, and D. Cook, "Temporal and structural analysis of biological networks in combination with microarray data," in CIBCB, 2008, pp. 62–69.

[41] L. Katz, "A New Status Index Derived from Sociometric Index," Psychometrika, vol. 18, 1953, pp. 39–43.

[42] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 10, 2004, pp. 1367–1372.

[43] W. Han, J. Lee, and J. Lee, "$Turbo_{ISO}$: Towards UltraFast and Robust Subgraph Isomorphism Search in Large Graph Databases," in SIGMOD, 2013, pp. 337–348.

[44] C. Raïssi, J. Pei, and T. Kister, "Computing Closed Skycubes," VLDB, vol. 3, 2010, pp. 838–847.

 ISBN: 978-1-61208-423-7