To Study the Use of Entropy and the Preprocessing Step in Analyzing Web Users' Behavior and Improving Web Mining

Chandan Kumar[1], Dr. M. M. Rahman[2]

[1]Research Scholar, Department of Computer Science, Patliputra University, Patna-20
[2]Associate Professor, PG Department of Mathematics, A. N. College, Patna

Abstract
The study of entropy's application to online use mining is the main goal of the work. By using the Reference Length approach, entropy provides an additional option for figuring out the ratio of auxiliary pages in the session identification. Two distinct online portals were used for the experiment. The first log file was downloaded via the web portal of a virtual learning environment course. The online gateway with anonymous access provided the second log file. When entropy and sitemap estimations of the ratio of auxiliary pages were compared, it was discovered that entropy could replace the sitemap estimate of the ratio of auxiliary pages with a full-valued substitute in the situation of sitemap abundance

Keywords: data preprocessing; information entropy; web usage mining; session identification; Reference Length

Introduction
Analysis of the behavior of web users is greatly influenced by the preprocessing phase of web usage mining. Many authors offered many solutions and various methods for log files preprocessing. Log files are the corner-stone of analysis of what the users have done on the web portal. The main steps of the web usage data preprocessing are data cleaning, web user identification, session identification, and path completion [1,2]. Each of the phases greatly influences the final results of the analysis. This paper deals with the improvement of data preprocessing of web usage data. It is focused on one particular method of session identification and introduces a novel approach to its usage with the help of entropy. Information theory and entropy were first introduced by C. Shannon [3], and had been used in different fields of informatics. Entropy is used as a measure of disorder, where lower entropy means order, and on the other hand, higher entropy means disorder. Following Shannon's definition [3], entropy can be used as a measure of uncertainty in a data set. It was used as a starting point in creating a novel approach to estimate the ratio of auxiliary pages, which is an important parameter for session identification using the Reference Length method.
The rest of the paper is structured as follows: in Section 2, entropy is presented and the related work of other authors about entropy is summarized. The theoretical and research background is described in **Section 3**. Section 3.1 (Theoretical Background) describes the Reference Length method of session identification. Section 3.2 (Research Background) contains the experiments' results focusing on the influence of sitemap estimation of the ratio of auxiliary pages to the accuracy of the session

identification, especially for Virtual Learning Environment (VLE) portal and portal with anonymous access. Section 4 deals with a description of the introduced approach, methodology, and results of the experiment. Subsequently, the discussion is offered in the last section.

2. Related Work

Entropy comes from the field of thermodynamics [4], and it was used to provide a statement of the second law of thermodynamics on the irreversibility of evolution. It was understood that an isolated system could not pass from a state of a higher entropy to a state of a lower entropy [5]. Entropy was firstly mentioned in the field of information theory by C. Shannon. He used a thought experiment to propose a measure of uncertainty in a discrete distribution [5]. The definition of entropy in information theory is as a degree of disorder or randomness in the system. Based on Shannon's definition in , given a random variable class C with a discrete probability distribution

$$\{p_i = Pr[C = c_i]\}_{i=1}^{k}, \sum_{i=1}^{k} p_i = 1, \quad (1)$$

where c_i is the ith class, then entropy $H(C)$ is defined as

$$H(C) = -\sum_{i=1}^{k} p_i \log p_i, \quad (2)$$

while the function decreases from infinity to zero and pi takes values from in the range <0, 1> [3,6]. Entropy as a modeling tool was formulated in [8], and it is known as Maximum entropy [9,10]. Other authors have partially used entropy in the field of web usage mining. Kumar et al. [11] implemented an algorithm of semantic-synaptic web mining algorithm that is based on the entropy value and information content. The algorithm deals with clustering of the web page and after that the entropy of web pages is calculated. The algorithm was examined on a large data set of web pages and the results indicate that the web pages with low entropy value generally provide the most relevant data. Liu et al. [12] presented a novel approach to feature selection based on the quality of information. Authors used the maximum-nearest neighbor to generalize Shannon's information theory. The proposed algorithm was verified on a set of datasets of UCI Repository of machine learning databases. The authors compared their algorithm with other popular feature selection of algorithms. The results showed that the proposed approach is more effective than other feature selection of algorithms. Arce et al. [13] presented a heuristic approach based on simulation annealing for the problem of sessionization. The sessionization problem addresses reconstruction of the user sessions. The quality of reconstruction is measured with respect to the power law for the size of the sessions on a web site. Entropy was used to identify the interesting partitions of the log file of a web portal. The authors measured the diversity of a given IP address regarding the entropy of the IP address. Levene and Loizou [14] introduced a novel algorithm for computing the entropy of Markov chain that represents the trail of web navigated pages by a user or a group of users. This allows for the authors to compute the probability of a typical trail, which can be also used to personalize ranking algorithms. The way the algorithm computes the entropy relates with the way users surf the web and how the web log data is collected. Authors also presented an extension of the algorithm to deal with high-order Markov chains of bounded order. Maung and Win [15,16] introduced a new heuristic to test suite reduction by applying entropy gain theory. The use of a large number of test cases from web usage logs is not practical within a time constraint. The algorithm combines the user session data and structural analysis of the examined web site to generate the test suite. The entropy-based reduction method is used to test case reduction for user session based testing. Despite the good empirical results, authors did not compare their approach to current user session based testing techniques. Jin et al.

[17] combined the maximum entropy model for the recommendation system. Their results showed that the recommendation system could achieve better accuracy, than a standard Markov model for page recommendation. It was also showed a better interpretation of web users' navigational behavior. Wang et al. [18] proposed an unified minimax entropy approach to user preference modeling with multidimensional knowledge. Authors used maximum entropy model to learn the check-in preferences of users. Check-in preference is an important component of Point-of-Interest prediction and recommendation. The proposed minimax entropy model is used to estimate the parameters with the preference learning. Ibl and Capek [19,20] used level of uncertainty (entropy) as an indicator for determining the degree of predictability of modelled systems. The authors focused on measuring the uncertainty of a process model that was modelled using stochastic Petri nets. Hui et al. [21] focused on the comparison of the performance of maximum entropy with the Naïve Bayes and Support Vector Machine algorithms, where entropy outperformed all of them. The algorithms were evaluated based on accuracy. The only downside of maximum entropy was its slow running when compared to other algorithms. Erlandsson et al. [22], in their article, showed how association rule learning could be used to predict user participation on social media pages. Data used for the experiment was gathered from Facebook. The results showed that using association rule learning, it is possible to identify influential users and can predict user participation in social media pages. Berezinski et al. [23], dealt with one task of data mining—anomaly detection. The authors presented an alternative entropy-based approach to anomaly detection caused by botnet-like malware. Jozani and Ahmadi [24] explored the ranked set sampling that has many applications in various fields. They have considered the information content of perfect and imperfect ranked set sampling data using the Shannon entropy, Rényi and Kullback-Leibler information measures. The results of their experiments showed desirable properties of ranked set sampling in comparison to commonly used simple random sampling in the context of information theory. Authors in [25,26] have focused on the problem of mining the structure of web site consisting of many hyperlink documents. The authors proposed an entropy-based analysis to analyze the entropy of anchor texts and links. Kao et al. [26] employed the entropy information to analyze the information measures of article sets. Entropy is used also to analyze the behavior on modern online social platforms. Wei and Zhu [27] introduced a cascade detection mechanism based of a web spam on entropy-based outlier mining algorithm. The entropy-based outlier mining algorithm consists of two steps: data discretizing and samples grouping to different sets, both based on entropy. Primarily, Agreste et al. [28] focused on analysis of user behavior on social platform were they created three profiles for each user and compared the profiles between each other using entropy and mutual information. De Meo et al. [29] analyzed the correlation between social and tagging behavior of the users on different social sharing systems.

Methods

This section describes the theoretical and research backgrounds that inspired our experimental direction using the entropy. The theoretical background provides necessary information about the process of session identification and one of its methods—Reference Length—is important for presented research. The research background deals with two previous experiments that served for this research as a starting point in discovery of the entropy-based estimation of a key parameter of the session identification method Reference Length. Experiments are focused on the sitemap estimation of the ratio of auxiliary pages to the accuracy of the session identification especially in the VLE and on portal with anonymous access.

Theoretical Background

In the process of session identification, it is important to divide the user's visits into sessions. The session is characterized by an activity of one user in a certain time on the web portal [30]. Issues with session identification can be solved by time-oriented heuristics, structure-oriented heuristics, and navigation-oriented user session identification. Time-oriented heuristics *h1* and *h2* create sessions based on a time window, for example, 30 or 10 min [1,31]. Structure-oriented heuristic, such as *h-ref*, identifies new sessions based on another parameter, such as a field referrer, where if the URL is not followed by the referrer, it becomes a new session [1,31,32]. Navigation-oriented methods assume that two sets of transactions (auxiliary-content and content-only) can be formed. In experiment [33], based on the sitemap the assumption about the ratio of auxiliary pages estimated for the session identification using the Reference Length method, was made. The Reference Length method falls into this category. It is based on the assumption that the amount of time, that the user spends on a page depends on whether the page is classified as an auxiliary or content page [1,32,33,34,35]. Figure1 shows a histogram depicting the distribution of the variable *RLength* representing the time spent on the pages of a particular portal. We assume that the variance of times spent on the auxiliary pages is small because the user "only" passes through the pages to his/her search target. The auxiliary page shapes the left part of the graph. The length of the time spent on content pages has a higher variance and shapes the right part of the graph. The proposed work differs from the above works since it is oriented on a specific method of session identification, namely the Reference Length method. The aim of this research is to find an alternative estimation of the ratio of auxiliary pages used in Reference Length method.

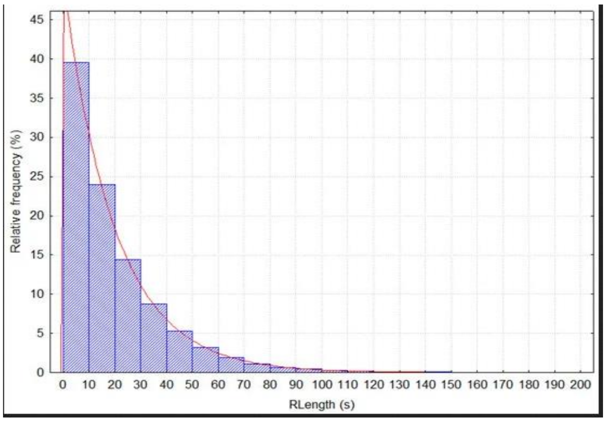


Figure 1. Distribution of the variable *RLength*.

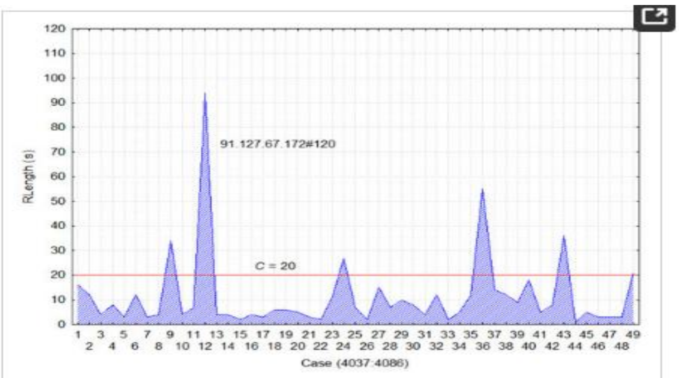


Figure 2. Reference Length method.

Research Background

Two previous experiments [33,35] are closely related to the aim of this paper. We focused on the influence of estimation of the ratio of auxiliary pages to the accuracy of the session identification. These experiments provide the reader with a solid review of the issues that are related to the estimate of the ratio of auxiliary pages in the process of session identification in the context of Web Usage Mining. Both experiments [33,35] dealt with the analysis of actionable (useful), trivial and inexplicable sequence rules [36]. The first experiment [35] dealt with the analysis of several data preprocessing techniques for session identification and path completion. Authors compared eight files in different stages of data preprocessing (Figure 3). Using the STATISTICA Sequence, Association, and Link Analysis were extracted sequence rules from frequent sequences for each examined file (Figure 3). Authors expected that the identification of sessions using the Reference Length method, estimated from a sitemap would have a significant impact on the quantity of extracted rules. This was not proved as session identification using the Reference Length method based on sitemap did not have an impact on the quantity of extracted rules in case of files without and with path completion. On the other hand, an assumption concerning the impact of the method based on sitemap on increasing the portion of useful rules was proved. It was proved that the use of the Reference Length method with the estimate of the ratio of auxiliary pages from the sitemap, has significantly increased the number of useful sequence rules found.

Figure 3. Application of data preparation to the log file of Virtual Learning Environment (VLE).

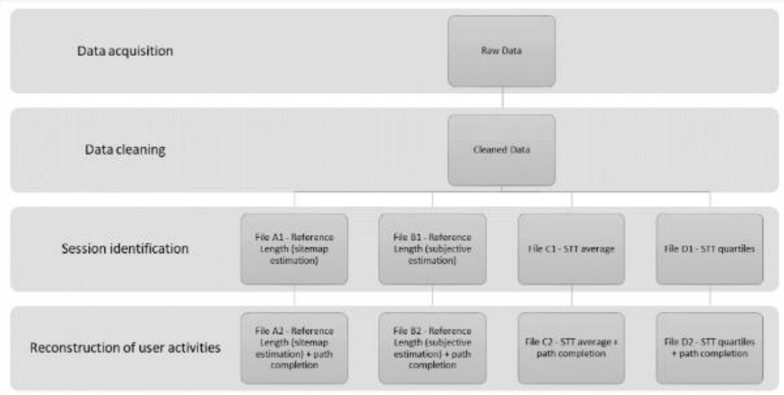
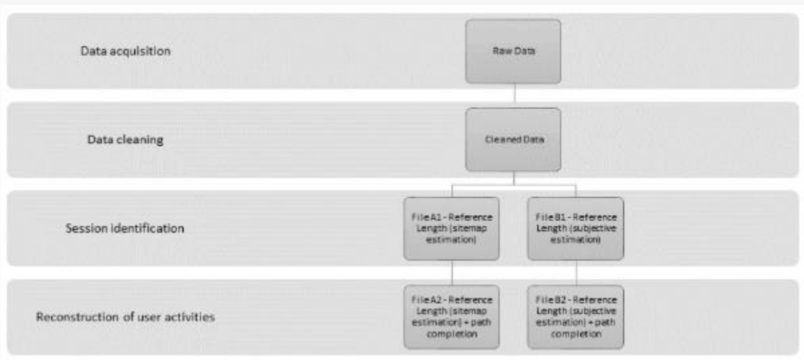


Figure 4. Application of data preparation to the log file of web portal.



A closer look at the results (Table 1) shows that the files without path completion (Files A1 and B1) contain almost identical rules, except four rules (5%) for the file with the subjective estimation of the ratio of auxiliary pages (B1). In case of files with path completion (Table 2) was proved a statistically significant difference in 16 new rules (almost 21%) in favor of file with subjective estimation (B2).

Table 1. Crosstabulations of File A1 × File B1.

A1\B1	0	1	∑
0	39 50.00%	4 5.13%	43 55.13%
1	0 0.00%	35 44.87%	35 44.87%
∑	39 50.00%	39 50.00%	78 100.00%
McNemar (B/C)	*Chi-square* = 2.25000; *df* = 1; *p* = 0.134		

Table 2. Crosstabulations of File A2 × File B2.

A2\B2	0	1	∑
0	0 0.00%	16 20.51%	16 20.51%
1	1 1.28%	61 78.21%	62 79.49%
∑	1 1.28%	77 98.72%	78 100.00%
McNemar (B/C)	*Chi-square* = 11.52941; *df* = 1; *p* = 0.00069		

Table 3. Crosstabulations: Incidence of rules × Types of rules: File A2.

A2\Type	Useful	Trivial	Inexplicable
0	0 0.00%	6 14.63%	10 37.04%
1	10 100.00%	35 85.37%	17 62.96%
∑	10 100%	41 100%	27 100%
Pearson	*Chi-square* = 7.97115; *df* = 2; *p* = 0.019		
Con. Coef. C	0.30450		
Cramér's V	0.31968		

Based on the contingency coefficients (Coef. C, Cramér's V), which represent the degree of dependency between two nominal variables, there is a moderate dependency among the portion of useful, trivial, and inexplicable rules, and their occurrence in the set of discovered rules of files without path completion (A1: 0.40, B1: 0.37) separately. Besides, the contingency coefficient is statistically significant. The obtained results for the files with path completion (A2, B2) were more interesting. A moderate dependency (0.30) was found among the portion of useful, trivial, and inexplicable rules, and their occurrence in the file A2 (Table 3). The coefficient value for the incidence of rules and types of rules for the file B2 was approximately 0.11, where 1 represents perfect dependency and 0 means independence, which means that there is only a small dependency for the file B2 (Table 4). Also, the contingency coefficient is not statistically significant (Table 4). The file B2 contained the most inexplicable rules, but the portion of useful rules was the same for all of the files.

Table 4. Crosstabulations: Incidence of rules × Types of rules: File B2.

B2\Type	Useful	Trivial	Inexplicable
0	0 0.00%	1 2.44%	0 0.00%
1	10 100.00%	40 97.56%	27 100.00%
∑	10 100%	41 100%	27 100%
Pearson	*Chi-square* = 0.91416; *df* = 2; *p* = 0.633		
Con. Coef. C	0.10763		
Cramér's V	0.10826		

Methodology

The log files used in the experiment were extracted from the VLE portal and web portal with anonymous access. Research methodology proceed from the results of the above research articles The experiment is comprised of the following steps:

Data acquisition- obtaining the log file and sitemap and defining the observed variables in the log file to obtain the necessary data (IP address, date and time of access, URL address, etc.).

Data cleaning-removing unnecessary data, such as access to images, styles, etc., and removing the accesses of robots of search engines.

- User identification—based on IP address and User Agent.
- Data transformation and sequence identification—creating time variable Unix Time from date and time of access and creating variable Length based on time window that the user would most likely spend on a web portal.
- Creating a data matrix from the log file with unique web portal pages and corresponding time spent on the page by the users.
- Calculating Relative Mean Time spent by the users for each web portal page from the Length variable.

- Calculating Entropy from Relative Mean Time for each web portal page and Average Entropy and Quartiles Entropy for the whole web portal.
- Estimate the ratio of auxiliary pages based on entropy—dividing the web portal pages to auxiliary and content based on the entropy and Average/Quartiles Entropy.
- Estimate the ratio of auxiliary pages from the sitemap—the sitemap consists of the variables URL and Referrer, where the number of auxiliary pages corresponds to the number of unique referring pages in the used web portal.
- Draw a comparison of the results of ratio of the auxiliary pages estimation by various techniques.The experiment expectations are in finding an alternative method to estimate the ratio of auxiliary pages in case the sitemap is missing or is inaccurate.

Results

The experiment was conducted separately for log files of two different web portals. The first log file was received from the virtual learning environment portal. This type of portal represents web portals with the need of users' login and cannot be accessed anonymously. The second log file represents a web portal with anonymous access, in our case, it is our university portal. Both log files were preprocessed using standard web usage data preprocessing techniques, as in The preprocessed log files were imported into the database separately. Several calculations, involving entropy, were conducted with the aim to find a way to distinguish auxiliary pages from content pages based on the length of time, which the user spent on each site. It resulted in the creation of an algorithm that could be able to calculate entropy for a specific page based on a random variable *RLength*, representing the length of the time spent on each web page of the portal. With the use of the algorithm, the variable *Relative Mean Time* was created, representing the time which the user spent on the page. From the variable *Relative Mean Time*, *Entropy* was derived from the individual page, and a new data matrix was created (Table 7) containing the *Entropy* of each page. Subsequently, it was calculated the average length of all accesses on the web portal, and it served as the cut-off value of time that divides the pages into auxiliary and content pages. The reason why calculated entropy was higher than 1 was a greater number of examined categories. All of the pages with a higher *Entropy* than the *Average Entropy* of the whole portal, will be classified as auxiliary pages. On the other hand, pages with a smaller *Entropy* than *Average Entropy*, will be classified as content pages**.**

References

1. Cooley, R.; Mobasher, B.; Srivastava, J. Data preparation for mining world wide web browsing patterns. *Knowl. Inf. Syst.* 1999, *1*, 5–32. [Google Scholar] [CrossRef]
2. Munk, M.; Kapusta, J.; Švec, P. Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. *Procedia Comput. Sci.* 2010, *1*, 2273–2280. [Google Scholar] [CrossRef]
3. Shannon, C.E. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 2001, *5*, 3–55. [Google Scholar] [CrossRef]
4. Clausius, R. On the Motive Power of Heat, and on the Laws which Can be Deduced from it for the Theory of Heat. In *Annalen der Physik*; Dover: Mineola, NY, USA, 1960. [Google Scholar]
5. Holzinger, A.; Hörtenhuber, M.; Mayer, C.; Bachler, M.; Wassertheurer, S.; Pinho, A.J.; Koslicki, D. On Entropy-Based Data Mining. In *Interactive Knowledge Discovery and Data Mining in*

Biomedical Informatics: State-of-the-Art and Future Challenges; Holzinger, A., Jurisica, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 209–226. ISBN 978-3-662-43968-5. [Google Scholar]

6. Lima, C.F.L.; de Assis, F.M.; de Souza, C.P. A Comparative Study of Use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection. In Proceedings of the 13th International Conference on Intelligent Data Engineering and Automated Learning, Natal, Brazil, 29–31 August 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 492–501. [Google Scholar]
7. Arora, P.N. On the Shannon measure of entropy. *Inf. Sci.* 1981, *23*, 1–9. [Google Scholar] [CrossRef]
8. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* 1957, *106*, 620. [Google Scholar] [CrossRef]
9. Karmeshu, J. (Ed.) *Entropy Measures, Maximum Entropy Principle, and Emerging Applications*; Springer: Berlin/Heidelberg, Germany, 2003; ISBN 3540002421. [Google Scholar]
10. Harremoeës, P.; Topsøe, F. Maximum Entropy Fundamentals. *Entropy* 2001, *3*, 191–226. [Google Scholar] [CrossRef]
11. Kumar, S.; Abhishek, K.; Singh, M.P. Accessing Relevant and Accurate Information using Entropy. *Procedia Comput. Sci.* 2015, *54*, 449–455. [Google Scholar] [CrossRef]
12. Liu, J.; Lin, Y.; Lin, M.; Wu, S.; Zhang, J. Feature selection based on quality of information. *Neurocomputing* 2017, *225*, 11–22. [Google Scholar] [CrossRef]
13. Arce, T.; Román, P.E.; Velásquez, J.; Parada, V. Identifying web sessions with simulated annealing. *Expert Syst. Appl.* 2014, *41*, 1593–1600. [Google Scholar] [CrossRef]
14. Levene, M.; Loizou, G. Computing the Entropy of User Navigation in the Web. *Int. J. Inf. Technol. Decis. Mak.* 2003, *2*, 459–476. [Google Scholar] [CrossRef]
15. Maung, H.M.; Win, K. An Efficient Test Cases Reduction Approach in User Session Based Testing. *Int. J. Inf. Educ. Technol.* 2015, *5*, 768–771. [Google Scholar]
16. Maung, H.M.; Win, K. Entropy Based Test Cases Reduction Algorithm for User Session Based Testing. In *Advances in Intelligent Systems and Computing, Proceedings of the Genetic and Evolutionary Computing (GEC 2015), Yangon, Myanmar, 26–28 August 2015*; Springer: Cham, Switzerland, 2015; Volume 388, pp. 365–373. [Google Scholar]
17. Jin, X.; Zhou, Y.; Mobasher, B. A maximum entropy web recommendation system. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD' 05), Chicago, IL, USA, 21–24 August 2005; ACM Press: New York, NY, USA, 2005; p. 612. [Google Scholar]
18. Wang, J.; Li, M.; Han, J.; Wang, X. Modeling Check-in Preferences with Multidimensional Knowledge: A Minimax Entropy Approach. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM' 16), San Francisco, CA, USA, 22–25 February 2016; pp. 297–306. [Google Scholar]
19. Ibl, M.; Čapek, J. Measure of Uncertainty in Process Models Using Stochastic Petri Nets and Shannon Entropy. *Entropy* 2016, *18*, 14. [Google Scholar] [CrossRef]
20. Ibl, M.; Čapek, J. A Behavioural Analysis of Complexity in Socio-Technical Systems under Tension Modelled by Petri Nets. *Entropy* 2017, *19*, 572. [Google Scholar] [CrossRef]

21. Wang, H.; Wang, L.; Yi, L. Maximum Entropy framework used in text classification. In Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen, China, 29–31 October 2010; pp. 828–833. [Google Scholar]
22. Erlandsson, F.; Bródka, P.; Borg, A.; Johnson, H. Finding Influential Users in Social Media Using Association Rule Learning. *Entropy* 2016, *18*, 164. [Google Scholar] [CrossRef]
23. Bereziński, P.; Jasiul, B.; Szpyrka, M. An Entropy-Based Network Anomaly Detection Method. *Entropy* 2015, *17*, 2367–2408. [Google Scholar] [CrossRef]
24. Jozani, M.J.; Ahmadi, J. On uncertainty and information properties of ranked set samples. *Inf. Sci.* 2014, *264*, 291–301. [Google Scholar] [CrossRef]
25. Kao, H.-Y.; Chen, M.-S.; Lin, S.-H.; Ho, J.-M. Entropy-based link analysis for mining web informative structures. In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM' 02), McLean, VA, USA, 4–9 November 2002; pp. 574–581. [Google Scholar]
26. Kao, H.-Y.; Lin, S.-H.; Ho, J.-M.; Chen, M.-S. Mining Web Informative Structures and Contents Based on Entropy Analysis. *IEEE Trans. Knowl. Data Eng.* 2004, *16*, 41–55. [Google Scholar] [CrossRef]
27. Wei, S.; Zhu, Y. Cleaning Out Web Spam by Entropy-Based Cascade Outlier Detection. In Proceedings of the Database and Expert Systems Applications, Lyon, France, 28–31 August 2017; Springer: Cham, Switzerland, 2017. [Google Scholar]
28. Agreste, S.; De Meo, P.; Ferrara, E.; Piccolo, S.; Provetti, A. Analysis of a Heterogeneous Social Network of Humans and Cultural Objects. *IEEE Trans. Syst. Man Cybern. Syst.* 2015, *45*, 559–570. [Google Scholar] [CrossRef]
29. De Meo, P.; Ferrara, E.; Abel, F.; Aroyo, L.; Houben, G.-J. Analyzing user behavior across social sharing environments. *ACM Trans. Intell. Syst. Technol.* 2013, *5*, 14. [Google Scholar]
30. Patil, P.; Patil, U. Preprocessing of web server log file for web mining. *World J. Sci. Technol.* 2012, *2*, 14–18. [Google Scholar]
31. Spiliopoulou, M.; Mobasher, B.; Berendt, B.; Nakagawa, M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS J. Comput.* 2003, *15*, 171–190. [Google Scholar] [CrossRef]
32. Kapusta, J.; Munk, M.; Drlík, M. Cut-off time calculation for user session identification by reference length. In Proceedings of the 2012 6th International Conference on Application of Information and Communication Technologies (AICT 2012), Tbilisi, Georgia, 17–19 October 2012. [Google Scholar]
33. Munk, M.; Benko, L'.; Gangur, M.; Turčáni, M. Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E M Ekon. Manag.* 2015, *18*, 144–159. [Google Scholar]
34. Munk, M.; Benko, L'. Improving the Session Identification Using the Ratio of Auxiliary Pages Estimate. In Proceedings of the Mediterranean Conference on Information & Communication Technologies (MedICT), Saidia, Morocco, 7–9 May 2015; Volume 381, pp. 551–556. [Google Scholar]
35. Munk, M.; Drlik, M.; Benko, L.; Reichel, J. Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE Access* 2017, *5*, 8989–9004. [Google Scholar] [CrossRef]
36. Berry, M.J.A.; Linoff, G.S. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*; John Wiley & Sons: Hoboken, NJ, USA, 2004. [Google Scholar]

37. Benko, L'.; Reichel, J.; Munk, M. Analysis of student behavior in virtual learning environment depending on student assessments. In Proceedings of the 13th International Conference on Emerging eLearning Technologies and Applications (ICETA 2015), Stary Smokovec, Slovakia, 26–27 November 2015; pp. 33–38. [Google Scholar]
38. Kapusta, J.; Munk, M.; Drlík, M. Analysis of Differences between Expected and Observed Probability of Accesses to Web Pages. In *Lecture Notes in Computer Science, Proceedings of the Computational Collective Intelligence. Technologies and Applications, Seoul, Korea, 24–26 September 2014*; Hwang, D., Jung, J., Nguyen, N.-T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8733, pp. 673–683. ISBN 978-3-319-11288-6. [Google Scholar]
39. Kapusta, J.; Munk, M.; Drlík, M. Identification of Underestimated and Overestimated Web Pages Using Pagerank and Web Usage Mining Methods. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9240. [Google Scholar]