An Alternative Approach to Ranking Videos and Measuring Dissimilarity Between Video Content and Titles

Kalp Abhinn Aghada

PES University, Bangalore, India

Abstract

In this work, an alternative statistical approach to video retrieval and ranking by introducing a novel dissimilarity measure is presented. This approach is naturally extrapolated to measure the dissimilarity between a video's audio-visual content and its title, hence aiding in video click-bait detection. The approach is described first, and then preliminary empirical results are stated. Results show that for small data sets, this approach works reasonably well.

Keywords

video retrieval, video ranking, click-bait detection, dissimilarity measure, information retrieval

1. Introduction

Similarity measures like the cosine, Jaccard, and Dice coefficient among others are ubiquitous in retrieving and ranking information including documents and videos [1, 2]. Of these, cosine similarity is a measure that is used significantly in ranking videos [3, 1, 4]. An alternative dissimilarity measure is introduced in this work that ranks videos based on their title, their captions, and their visual content. A search algorithm exemplifies the usage of this measure to retrieve videos. This dissimilarity measure is evaluated for video retrieval on a small-scale dataset and is compared with a state-of-the-art video retrieval model. Empirical results show that this method could serve as a viable alternative to current video retrieval and ranking methods. This is the first contribution of this work.

Misinformation tends to be rife in clickbait content, be it in text, audio, or video format. The ease of consumption of the video format however, makes it easy for misinformation to spread rapidly. Current video clickbait detection methods rely solely on meta-information about the videos [5, 6, 7] like the thumbnails and comments, or individual modalities like audio transcripts [8]. However, it is more often than not, the extent of the dissimilarity between the title of a piece of information and its content that makes it a clickbait [9]. This work extrapolates the introduced dissimilarity measure that is used to retrieve and rank videos and applies it to measure the dissimilarity between the title of a video and its audio-visual content. This extended dissimilarity measure is then used to detect clickbait

ROMCIR 2022: The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2022: the 44th European Conference on Information Retrieval, April 10-14, 2022, Stavanger, Norway

videos on a small-scale dataset and is evaluated. And this is the second contribution of this work. This could potentially contribute towards helping mitigate misinformation and information disorder of the video format.

2. Related Work

Several similarity measures and dissimilarity measures exist that aid in accomplishing document and video ranking like the euclidean distance, cosine similarity [10], overlap coefficient, and dice coefficient among others that can be used in conjunction with vector space based models [11]. Chambon et al. [12] categorize the similarity measures that have been used in conjunction with visual data into five families. A significant number of video retrieval and ranking models utilize cosine similarity. There has been related work done in the field of information retrieval for multi-modal data that utilize these similarity measures, like Miech et al. [13, 4] and the zero-example video retrieval model proposed by Dong et al. (2019) [3] that both use the cosine similarity to perform video retrieval and ranking. Miech et al. [4] in the process of implementing an action localized text-to-video retrieval model, compile 1.22 million instructional videos of which a small sample has been used in this work.

Clickbait detection, albeit being a scantily researched subject a few years back [14], has at this point gained quite a bit of traction. Potthast et al. [14] classify tweets as clickbait by considering titles, linked web pages, and meta-information in conjunction with random forests, logistic regression, and naive Bayes to achieve 0.79 area under the Receiver Operating Characteristic (ROC) curve. Bourgonje et al. [15] classify clickbait news articles using an n-gram based approach. The approach to clickbait detection by Dong et al. [9] comes the closest in methodology to the approach proposed in this work even though they apply clickbait detection to text and not to audio-visual content. They define an article or a piece of text with an accompanying title to be clickbait when its title does not match its content. A similar definition of clickbait has been adopted in this work.

Clickbait detection in videos remains a challenge primarily due to the multi-modal aspect of the data. Shang et al. [5] describe a content-agnostic approach to identify clickbait videos by evaluating user comments instead of evaluating the contents of a video. Zannettou et al. [7] attempt to identify clickbait videos using titles, thumbnails, tags, and other meta-information. Similar methods exist where the meta-information is considered to classify clickbait videos [6].

However, to the best of the author's knowledge, there is no research that takes into account existing audio, visual and titular content of videos together to detect clickbait. The proposed method considers the audio-visual and titular components of videos to firstly, rank videos and secondly, measure dissimilarity between video content and titles, in effect, aiding in detecting clickbait.

3. Proposed Search Methodology

Videos are ranked by comparing the input query with their audio, visual and titular data separately. Hence, a dissimilarity measure accounting for this would depend on object and audio recognition from videos. A few important definitions pertaining to this are stated before introducing the dissimilarity measure.

3.1. Object Recognition

In this work, the YOLOv3 [16] implementation of ImageAI [17] trained on the Microsoft COCO data set [18] is used to detect a class of 80 everyday objects from videos. In theory, any object detection model can be used to perform retrieval with the proposed dissimilarity measure. The ratio of the number of times each object is detected to the total number of times all objects are detected is calculated for each video. The set of these ratios pertaining to all objects for the k^{th} video will be referred to hereon as C_k and the set of objects for the k^{th} video as O_k.

3.2. Audio Recognition and Query Processing

In this work, the pretrained English speech model of CMU Sphinx [19, 20] is used to convert natural language from the audio feed of videos to text. In theory, any speech recognition model can be used. The set of tokens pertaining to the k^{th} video will be referred to hereon as S_k and the set of words in the title of the k^{th} video as T_k. The term frequency for each $word$ in S_k and T_k is calculated and represented as $tf(word, S_k)$ and $tf(word, T_k)$ respectively.

The input query is tokenized and stop words are removed. This processed input query will be referred to hereon as q. Based on whether q is to be compared with the title, the captions or the content of the k^{th} video, $q \cap T_k$, $q \cap S_k$ and the ratios from C_k that pertain to $q \cap O_k$ give the required matches respectively.

3.3. Dissimilarity Measure

The proposed novel dissimilarity measure is a variadic function f, defined by Equation 1.

$$f(x_1, x_2, x_3, ..., x_n) = \frac{\sqrt{\frac{\sum (x_i - \frac{\sum x_i}{n})^2}{n}}}{(\frac{\sum x_i}{n})^{n-1}} + \frac{1}{\frac{\sum x_i}{n}} \tag{1}$$

Where,

- $x_i \in C_k \ \forall i \in q \cap O_k$ and $n = |q \cap O_k|$, or
- $x_i = tf(i, S_k) \ \forall i \in q \cap S_k$ and $n = |q \cap S_k|$, or
- $x_i = tf(i, T_k) \ \forall i \in q \cap T_k$ and $n = |q \cap T_k|$.

Depending on whether the query is to be compared with the visual content, the audio content, or the title of the k^{th} video respectively.

The lesser the value of this function, the better the match between the input query and the video for which the value is being calculated. Algorithm 1 described in the next section exemplifies the usage of this dissimilarity measure. It returns the video with the lowest value for f among all videos present in a database, no matter what the absolute numerical value of that number may be. This dissimilarity measure combines measures of spread and mean, such that the score from the measure is lower, or better, when the spread is low and the mean is high. This implies, that in the instance of video retrieval, if there exist matches for a query in the audio and the visual content of a video, but if either the spread of the normalized matches is too high, or the mean of the normalized matches is too low, then the video will rank poorly as should be expected. In contrast, if the spread of the normalized matches is sufficiently low, and the mean of the normalized matches is sufficiently high, a video will rank favourably in response to the query. An example of the search methodology in action is given in section 3.5.

3.4. Search Algorithm

Algorithm 1 is a one-pass algorithm with time complexity of $O(n)$, where n = number of videos. In terms of the set of all words in the captions for the k^{th} video S_k and the input query q, the time complexity is $O(n \times |q \cap S_k| \times |q|)$.

Algorithm 1 shows a blueprint for video retrieval using the dissimilarity measure f.

Algorithm 1 Algorithm blueprint for video retrieval using f

Input: $D = \{set\ of\ all\ visual\ data\ or,\ set\ of\ all\ captions\ or,\ set\ of\ all\ titles\}$, q = set of all words in tokenized input query, $E_k = \{O_k\ or,\ S_k\ or,\ T_k\}$ where $k \in$ set of all videos

Output: R = search output

```
for entry in D do
    for word in q ∩ E_entry do
        Calculate C_entry or tf(word, entry)
    Calculate f.
    if Calculated f is lesser than previous f then
        Update R with video corresponding to entry
if (R != empty) then
    Return R
```

3.5. Search Methodology in Action

An example of the search methodology in action can be demonstrated with the help of Figure 1. The three rectangles on the left edge of the figure represent three input sub-queries. The three sets of three rectangles to the right of the sub-queries represent the amount of those matching sub-queries in three videos.

The search starts by iteratively comparing the set of input sub-queries to each video's sub-query match. The comparison takes place with the help of the measure f - where if

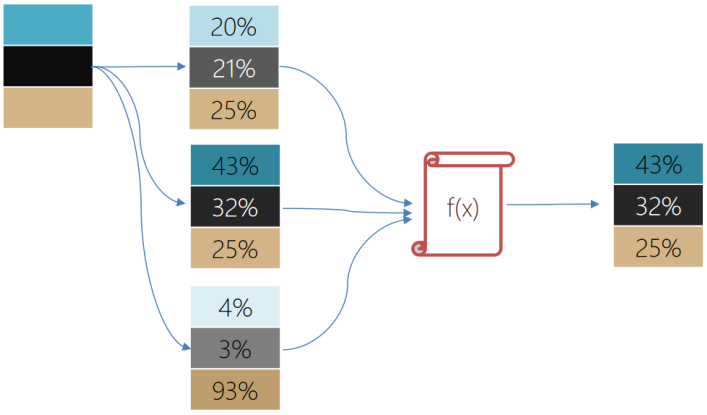

Figure 1: An example of the search methodology.

the value of f for the current video is less than the preceding video, then the current video is a better match. The first video has a 20% match for the corresponding first sub-query, a 21% match for the second sub-query, and a 25% match for the third sub-query; and so on for the remaining two videos. In this example, the values of f for {21%, 20%, 25%}, {43%, 32%, 25%} and {4%, 3%, 93%} are 0.051, 0.038 and 0.076 respectively. The video for which the value of f would be the lowest is the second video, and that would be ranked most favourably. As can be seen in this example, this dissimilarity measure ranks videos favourably when the spread of the sub-query matches is sufficiently low, and the mean of the sub-query matches is sufficiently high, with the sufficiency defined as per Equation 1.

3.6. Measuring Dissimilarity between Video Content and Titles

The search methodology described in subsections above, aid in quantitatively measuring the dissimilarity between the input query and the video content in order to rank and retrieve videos. This concept can be naturally extrapolated to measure the dissimilarity between a video's audio-visual content and its title. In fact, it can be viewed as a direct application of the dissimilarity measure f and Algorithm 1.

Instead of having the input x_i in Equation 1 be defined as an intersection of the tokenized input query q with the k^{th} video content (O_k and C_k), the caption content (S_k), or the title (T_k) each separately; it can be defined as an intersection of the title with the video content and the caption content as stated in Equation 2.

$$x_i = \begin{cases} x & | \; x \in C_k \; \forall i \in T_k \cap O_k \\ tf(i, S_k) & \forall i \in T_k \cap S_k \end{cases} \tag{2}$$

Table 1
Dataset Statistics.

Dataset Parameter	Min	Mean	Max
Video Length	0.86 minutes	5.05 minutes	14.99 minutes
Title Length	3 words	7 words	11 words
Caption Length	137 words	884 words	2796 words

4. Empirical Analysis

4.1. Dataset

The evaluation is done on a dataset of 304 videos that were collected from the large scale instructional video database HowTo100M compiled by Miech et al. [4] by selecting the top 5 results for each search query that was in the set of all 80 common everyday objects that ImageAI [17] can detect. Repeated results, results in languages other than English, and results longer than 900 seconds are excluded from the sample. The evaluation is done on a set of 100 queries. The sampling rate for video processing was $5fps$. Table 1 shows various statistics for the data.

4.2. Evaluation Metrics

To evaluate video retrieval, standard precision at k defined by the proportion of top k documents that are relevant; and recall at k defined by the proportion of relevant documents that are in the top k, are used as is ubiquitous in the field of information retrieval [11]. To evaluate clickbait classification, precision, recall, accuracy, F1-score and Matthews correlation coefficient (MCC) [21, 22] metrics are used. MCC is a more reliable metric than F1-score as it takes into account all four confusion matrix classes [21].

The dissimilarity measure in Equations 1 and 2 quantifies the dissimilarity between video content and titles and is applied to classify clickbait videos into two classes, namely clickbait and not clickbait based on a threshold value $f = 100$. The selection of this threshold value for f is arbitrary within reason - videos with $f \geq 100$ or where effectively $\leq 1\%$ of the audio-visual content match the title are classified as clickbait. Both of these can be considered as equivalent.

The relevance assessment for retrieval per query was manually established. The ground truth for clickbait classification of all the videos was manually established as well by annotating all the videos into two classes, clickbait and not clickbait.

4.3. Results

All results stated hereon are averaged across all queries. Tables 2 and 3 show the results of video retrieval compared with state-of-the-art separable 3D CNN (S3D) [23] implementation of Multiple Instance Learning Noise Contrastive Estimation (MIL-NCE) [13, 4] that uses the cosine similarity. These results show that the dissimilarity and

Table 2
Video Retrieval/Ranking Statistics for Precision at k.

Method	P@1	P@2	P@3	P@4	P@5
MIL-NCE [13, 4]	40.00%	60.00%	45.66%	57.50%	51.40%
Method described in this work (visual data)	75.00%	73.50%	76.67%	67.50%	70.00%
Method described in this work (caption data)	35.00%	59.50%	47.30%	57.52%	52.00%
Method described in this work (title data)	79.00%	75.00%	76.67%	77.50%	75.20%

Table 3
Video Retrieval/Ranking Statistics for Recall at k.

Method	R@1	R@2	R@3	R@4	R@5
MIL-NCE [13, 4]	5.71%	17.14%	19.57%	32.86%	36.71%
Method described in this work (visual data)	10.71%	21.00%	32.85%	38.57%	50.00%
Method described in this work (caption data)	5.00%	17.00%	20.27%	32.87%	37.14%
Method described in this work (title data)	11.28%	21.43%	32.86%	44.28%	53.71%

Table 4
Clickbait Classification Statistics.

Recall	Precision	Accuracy	F1-Score	MCC
0.77	0.85	0.97	0.81	0.79

method described in this work significantly outperforms MIL-NCE [13, 4] for small data sets.

Table 4 shows the statistics for clickbait classification. MCC is a contingency matrix method of calculating the Pearson correlation coefficient [22], and hence can be interpreted in a similar way [24]. An MCC of 0.79 implies a significant agreement between predictions and observations and that the classification works reasonably well for small data sets. These results could in part be attributed to the fact that the dissimilarity measure gives equal importance to the titles, captions, and the audio-visual content of videos together. The video retrieval and the clickbait classification both show promising preliminary results for small data sets. An implementation of this method is available at: https://github.com/kalpaghada/savs.

5. Conclusion and Future Work

The objective of this work was to attempt to formulate an alternative mathematical statistical method of ranking results for multi-modal video retrieval through audio, visual, and titular data by introducing a novel dissimilarity measure. This objective naturally extrapolated to measure the dissimilarity between a video's audio-visual content and titles that can aid in detecting clickbait videos. Results show that this method outperforms a state-of-the-art video retrieval and ranking approach and performs reasonably well in

classifying clickbait videos for small data sets, serving as a viable alternative to both applications.

For future work, the size of the data set sample this system is tested on can be increased. The video retrieval method can be used in conjunction with methods like MIL-NCE [13, 4] and tested alongside methods suited to large scale data sets [25]. The three modalities in this work, namely, the audio, the video, and the title, although being considered together for clickbait classification, are considered separately for video retrieval; an attempt can be made to combine these three modalities for video retrieval. The clickbait classification method can be combined with methods like naive Bayes [26], logistic regression [27] or random forests [28, 14] and the threshold value can be treated as a learnable parameter and optimized. Clickbait content on online platforms tend to attract viewers by triggering an emotional response [29]; in the video format, it could be an initial emotional response through misleading thumbnails, descriptions and other meta-information; integrating this aspect with this approach could be looked into. A large-scale video library can be compiled as well, for further research on multi-modal video clickbait classification.

References

[1] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, Y. Kompatsiaris, Near-duplicate video retrieval by aggregating intermediate cnn layers, in: L. Amsaleg, G. þ. Guðmundsson, C. Gurrin, B. þ. Jónsson, S. Satoh (Eds.), MultiMedia Modeling, Springer International Publishing, Cham, 2017, pp. 251–263.

[2] X. Wu, W.-L. Zhao, C.-W. Ngo, Near-duplicate keyframe retrieval with visual keywords and semantic context, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 162169. URL: https://doi.org/10.1145/1282280.1282309. doi:10.1145/1282280.1282309.

[3] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, X. Wang, Dual encoding for zero-example video retrieval, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9338–9347. doi:10.1109/CVPR.2019.00957.

[4] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in: ICCV, 2019.

[5] L. Shang, D. Y. Zhang, M. Wang, S. Lai, D. Wang, Towards reliable online clickbait video detection: A content-agnostic approach, Knowledge-Based Systems 182 (2019) 104851. URL: https://www.sciencedirect.com/science/article/pii/S0950705119303260. doi:https://doi.org/10.1016/j.knosys.2019.07.022.

[6] T. Xie, T. Le, D. Lee, Checker: Detecting clickbait thumbnails with weak supervision and co-teaching, in: Y. Dong, N. Kourtellis, B. Hammer, J. A. Lozano (Eds.), Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, Springer International Publishing, Cham, 2021, pp. 415–430.

[7] S. Zannettou, S. Chatzis, K. Papadamou, M. Sirivianos, The good, the bad and the

bait: Detecting and characterizing clickbait on youtube, in: 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 63–69. doi:10.1109/SPW.2018.00018.

[8] B. Gamage, A. Labib, A. Joomun, C. H. Lim, K. Wong, Baitradar: A multi-model clickbait detection algorithm using deep learning, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2665–2669. doi:10.1109/ICASSP39728.2021.9414424.

[9] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, Similarity-aware deep attentive model for clickbait detection, in: Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, S.-J. Huang (Eds.), Advances in Knowledge Discovery and Data Mining, Springer International Publishing, Cham, 2019, pp. 56–69.

[10] T. Korenius, J. Laurikkala, M. Juhola, On principal component analysis, cosine and euclidean measures in information retrieval, Information Sciences 177 (2007) 4893–4905. URL: https://www.sciencedirect.com/science/article/pii/S0020025507002630. doi:https://doi.org/10.1016/j.ins.2007.05.027.

[11] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. doi:10.1017/CBO9780511809071.

[12] S. Chambon, A. Crouzil, Similarity measures for image matching despite occlusions in stereo vision, Pattern Recogn. 44 (2011) 20632075. URL: https://doi.org/10.1016/j.patcog.2011.02.001. doi:10.1016/j.patcog.2011.02.001.

[13] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman, End-to-end learning of visual representations from uncurated instructional videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[14] M. Potthast, S. Köpsel, B. Stein, M. Hagen, Clickbait detection, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2016, pp. 810–817.

[15] P. Bourgonje, J. Moreno Schneider, G. Rehm, From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles, in: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 84–89. URL: https://aclanthology.org/W17-4215. doi:10.18653/v1/W17-4215.

[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] Moses, J. Olafenwa, Imageai, an open source python library built to empower developers to build applications and systems with self-contained computer vision capabilities, 2018–. URL: https://github.com/OlafenwaMoses/ImageAI.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[19] pocketsphinx, CMUSphinx, Accessed: 2022-05-03. URL: https://cmusphinx.github.io/.

[20] D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, A. Rudnicky,

Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, volume 1, 2006, pp. I–I. doi:10.1109/ICASSP.2006.1659988.

[21] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6. URL: https://doi.org/10.1186/s12864-019-6413-7. doi:10.1186/s12864-019-6413-7.

[22] D. M. W. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, CoRR abs/2010.16061 (2020). URL: https://arxiv.org/abs/2010.16061. arXiv:2010.16061.

[23] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[24] P. Schober, C. Boer, L. A. Schwarte, Correlation coefficients: Appropriate use and interpretation, Anesthesia & Analgesia 126 (2018). URL: https://journals.lww.com/anesthesia-analgesia/Fulltext/2018/05000/Correlation_Coefficients__Appropriate_Use_and.50.aspx.

[25] J. Lokoč, G. Kovalčík, T. Souček, J. Moravec, P. Čech, Viret: A video retrieval tool for interactive known-item search, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 177181. URL: https://doi.org/10.1145/3323873.3325034. doi:10.1145/3323873.3325034.

[26] G. H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, p. 338345.

[27] S. L. Cessie, J. C. V. Houwelingen, Ridge estimators in logistic regression, Journal of the Royal Statistical Society. Series C (Applied Statistics) 41 (1992) 191–201. URL: http://www.jstor.org/stable/2347628.

[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[29] B. Ghanem, P. Rosso, F. Rangel, An emotional analysis of false information in social media and news articles, ACM Trans. Internet Technol. 20 (2020). URL: https://doi.org/10.1145/3381750. doi:10.1145/3381750.