Online Media Communication Performance Statistics Based on Bayesian Algorithm

Xiaoxing Ma[1, a],

[1]Tianjin Hexi District Zhujiang Road 25# Tianjin University of Finance & Economics,

[a]xxingl@163.com,

Keywords: Bayesian statistics, Classical statistics, Prior distribution, Web crawler.

Abstract. Network media has become the main means of communication in modern life, this paper discusses how to do evaluation of the effect about network media dissemination, whether the media files reach the audience and the extent of the impact on the audience. To calculate the performance effect of network media, this paper designs evaluation method based on Bayesian weighted algorithm: it should not only consider the number of media files, but also consider the merits of the audience to accept the performance. We introduce a Bayesian statistical algorithm, considering various factors, to get the most objective evaluation results.

Introduction

The network has become a major part of people's life, it can not only be the source of material resources, and also be the sources of spiritual life. The online media file uploaded to Internet has various forms and ample content. Compared with traditional media, online media file have the advantage that spread fast, wide range of communication, big number of the audience, can get more detailed information about the behavior of the media audience: the number of visits, access time, frequency of access, residence time and other data. Communication performance is refers to the evaluation of whether the media files reach the audience and what degree to reach, influence to the audience and so on. The current Internet technology not only gives the audience the right to actively searching and release information, but also provides a new opportunity and platform for the audience's aggregation. Audiences conduct unreserved performance with different interests, hobbies, values, so we are more conducive the evaluation of communication performance.

We are difficult to achieve communication performances rely solely on artificial statistical analysis. We should make a more scientific and accurate evaluation method based on the study of network communication of evaluation and the audiences as the object. The evaluation apply the network search technology, audio and video analysis technology, Bayesian weighted statistical algorithm, can real time and automatically monitoring and evaluation situation of effective of the online media for reprint, click and view, to intuitively and quickly master and analysis of the performance of the communication of the online media file. We use natural language processing technology, content based audio and video analysis and retrieval technology provide a complete and accurate data for further evaluation to multi-modal collection of text, images, audio, video media file formats. For the accuracy of the statistical results, we started to apply the Bayesian statistical algorithms for effective monitoring and evaluation.

Rely on the support of a variety of traditional statistical software; we can get all the information of the media audience, including gender, age, income, education, occupation and other demographic data, as well as the number of visits, access time, access frequency, dwell time, comments and other behavioral data. Through in-depth research and analysis, we found that with the Internet data, the main index data associated with online media performance is the following:

Number of Media exposure: Refers to the number of the page released media is clicked, generally counted by the site counter. If the number of web pages released media exposure is higher, indicating that the media get the more attention.

Number of clicks: The number of times the media has been clicked by the user is called the number of clicks. The number of clicks can be more objective and accurate to reflect the performance of advertising. This is a statistical method that we usually use in the tradition.

Click rate: By Number of clicks divided by the media Number of Media exposure, you can get the click rate.

Comments: After watching the media, the audience will make comments for the content of the online media. For example, people rate the content, or to make a brief comment. Relative to the number of clicks, click rate more truly reflect the quality of online media communication performance. For example, an online media file is only 100 people have seen it and these 100 people gave it to comment, another file have been seen by 10000 people, but no one comment after watch, so that the communication performance is not the same.

In the performance of network media statistics, it should not only consider the number of media files, but also consider the merits of the audience to accept the performance, which is based on the Bayesian algorithm. We introduce the Bayesian weighted statistical algorithm to the communication performance and consider the various factors, in order to get the most objective evaluation results.

Bayesian algorithm

Among the statistics, the classical statistical and the Bayesian statistical are the two main schools of thought. Bayesian method is proposed by the British scholar Bayesian in the published paper “On the solution of the problem of opportunity ", and developed in the debate with the classical statistical school of thought, is also more and more being apply and extensive research by statistical workers.

Classical statistics

General information is the information that is included in the general distribution or in the population of the population, including the general knowledge, the range of parameters, the methods and features of the variables; Sample information is the information contained in the sample taken from the population.

Bayesian statistics

Classical statistics gradually exposed some problems, many scholars of the two statistical school of thought found that during comparative study, compared with classical statistical methods, Bayesian statistical method has advantages in many aspects, such as intuition, accuracy, and so on. Two basic concepts of Bayesian statistical method is the prior distribution and posterior distribution, prior distribution is a probability distribution parameter. The Classical statistics is based on the general information and the sample information to infer. Bayesian statistics using the prior distribution, the prior distribution is mainly based on the experience and historical data, which is based on the Classical statistical method. The fundamental point of view in Bayesian is that any statistical inference about the distribution parameter, in addition to the use of the information provided by the sample, must provide a prior distribution, it is an indispensable factor in the statistical inference, they consider the prior distribution do not need to have an objective basis, can be partially or completely represent subjective beliefs, the posterior distribution, according to the prior distribution of samples and unknown parameters, with method for the conditional probability distribution used in probability theory, under the condition of known sample, compute shows the distribution conditions of unknown parameter. In the Classical statistics, the sample is considered as a population with a certain probability distribution, and the parameters in the population are the ordinary unknown variables; In contrast, Bayesian statistics are considered as random variables in any of the unknown parameters, according to the no determinacy using a probability distribution to describe the unknown parameters. The key method of Bayesian statistics is to deduce any inference must and only according to the posterior distribution, without involving the distribution of samples. In the statistical inference, only using the data that already appeared, namely the sample information, this is the "conditional view" in Bayesian statistics. Based on the difference in the use of the sample, the Bayesian statistics does not

recognize the criteria "bias" in the classical statistics, because they consider that the sample must be extracted from the real data, rather than the estimated amount of all possible sample space.

Bayesian statistics analysis method has been widely used in many fields, such as astronomy, meteorology, medical diagnosis, economic management, communication engineering, control technology of quality and reliability etc. With the continuous development and improvement of Bayesian statistics, many of the latest knowledge and methods have been integrated into the theory of Bayesian statistics, and Bayesian theory plays a more and more important role in statistics.

Let us give an example to illustrate: Fig.1 the chart 1 and chart 2 can be seen, the number of clicks of A file and B file on the statistics are the same in the unit period, however, the number of comments for the A file only in March, reached the sum of the 6 months of B file. If we simply look at the total value, number of clicks and comments of these two files are the same, the communication performance should be the same. But for the communication performance, the performance of communication is much better for the file has person to click and comment in the unit period. Even if the file's has more access and more comments, it can not illustrate the communication performance is very wide, because the statistics of the communication performance is a comprehensive evaluation value.

Chart 1

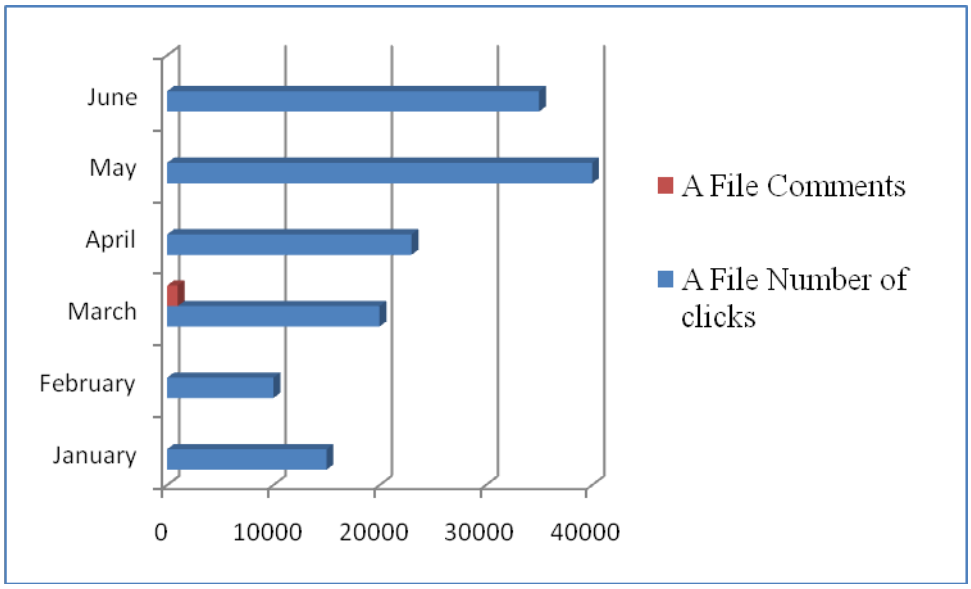

Chart 2

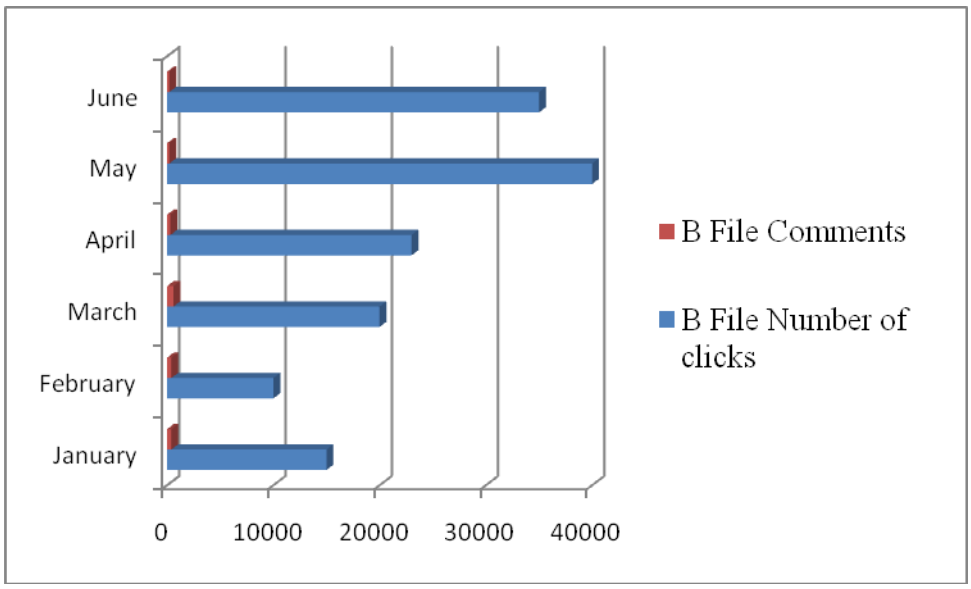

Fig.1 Example diagram

With the continuous development and improvement of Bayesian theory, many of the latest knowledge and methods have been integrated into the theory of Bayesian's analysis, and Bayesian's theory plays a more and more important role in statistics. A characteristic of the Bayesian approach is very obvious, approaches is in a lot of uncertainty in the incident, by learning and inductive methods to calculate the probability of occurrence of a particular event. Bayesian theory has a wide application in data mining, such as causal reasoning and uncertain knowledge representation, user classification, regression analysis and clustering pattern discovery. Bayesian statistics provides a posteriori assumptions and a computational method.

$$\mathbf{P}(\mathbf{A}\,|\,\mathbf{X}) = \frac{P\ (X\,|\,A\ )P\ (A)}{P\ (X\ )} \tag{1}$$

Wherein, $P\ (X\ )$ represents the prior probability of the data space X, that is, when the $\mathbf{A}$ is not set up, the probability of X is established; $P\ (A)$ represents there is no given space X, the initial probability of assuming space A establishment; $P\ (X\,|\,A\ )$ said the hypothesis $\mathbf{A}$ holds in the case, the probability of training space X; $\mathbf{P}(\mathbf{A}\,|\,\mathbf{X})$ that we need to calculate the results, represents in the training space X, established probability of hypothesis $\mathbf{A}$, which is a posteriori probability, react the confidence level of $\mathbf{A}$ when X was established. In general, X can be divided into two opposing events, that is, $\mathbf{A}$ and $\sim A$ (the opposite of the event $\mathbf{A}$), the $P\ (X\ )$ can be expressed as:

$$P(X) = P(X\,|\,A)P(A) + P(X\,|\sim A)P(\sim A) \tag{2}$$

The general form of Bayesian formula can be expressed as:

$$\mathbf{P}(\mathbf{A}\,|\,\mathbf{X}) = \frac{P\ (X\,|\,A\ )P\ (A)}{P\ (X\ )} \tag{3}$$

Model and results analysis Based on Bayesian weighted statistical optimization

The process of comprehensive evaluation of the online media files is divided into three parts: The searching of the online media files, the online media data collection, and finally the results calculated to get the value. At present, multimedia document search engine mainly uses the related text in the web page to extract the key words of the multimedia information to carry on the multimedia information retrieval, text in the web page to extract the key words of the multimedia information to carry on the multimedia information retrieval. But because of the content of multimedia information is extremely rich, sometimes it is difficult to use a simple description of several key words, and sometimes the automatic extraction of keywords and multimedia content does not match, resulting in multimedia retrieval results based on keywords is unsatisfactory. The search results often contain too much and are inconsistent to the theme of the content search, users often need to browse to choose which consumes much time, reduce the efficiency of the use of cyber source.

Content based multimedia retrieval technology developed, from the content of multimedia retrieval to solve the hierarchy problem, improve the retrieval accuracy, avoid subjective and incomplete caused by text description. The so-called content based multimedia retrieval is automatic analysis and extraction of the content multimedia for low-level audio-visual features (such as image color, texture, shape, video of the scene, the camera, frame, sound strength, tone, etc.). To obtain the middle-level and high-level features of the theme layer, we used pattern recognition technology, aiming at the features of multimedia retrieval.

File collection. The acquisition and processing of data is filtered based on the vitality of media data information through interface. In the media file collection, we will combine keyword search mechanism and search mechanism based on the content of the media file to compare the similarity, complement each other, to establish the association, to improve the retrieval precision. First, using context information to automatic extract media files, according to media file name, the page theme,

file URL form feature set to extract text keyword. We use keywords search media files in the Web, and then use the combined content-based search method. Analysis of the content and structure, extract the features about visual and object, combined with the keyword extracted from related text classification to recognition and increase the depth of indexing, to establish feature index with the extract feature of file. Establish the media resource index database, finally, by using retrieval condition combined natural language fuzzy query mode, matching the file according to the similarity of keywords and feature index, export the retrieval results sorted by the similarity.

We use the technology of web crawler to collect the media file for analysis, web crawler is a procedure that grabs program automatic crawling along the Internet, to extract each web page, at the same time we extract hyperlinks, as a clue to further crawl. The content of the video is more abundant than the image, and the data is very large in the media file. According to the description of the video content features, we use strict matching method for video file format, size, category, video length, including shot type. The collect of media file will sorted by the similarity , we want to evaluation for the performance of media files such as: the amount of each page access, number of access for IP, number of comments, forwarding capacity and other detailed data.

Using Bayesian weighted algorithm. Uses Bayesian weighted algorithm to analyze these data to determine the performance of communication. The following treatment for the web page contained media file:

Step1 Get the statistics information of the current media file for file types, reviews, the number of clicks on, the total number of clicks -- $ClickNumi$

Step2 Calculate the averages number of review in unit time -- Ri;

Step3 Get the least number of requires hits in the database of current file type -- mi;

Step4 Get the number of average review for each media types in the unit time in the database -- Ci;

Step5 According to the formula, get the weighted score;

$$PageEffecti = (ClickNumi \div (ClickNumi + mi)) \times Ri + (mi \div (ClickNumi + mi)) \times Ci \quad (4)$$

$\sum PageVisiti$ For the number of access times of media for all the web page contained the media. The weight of each page the media is the comprehensive communication performance for $Pi = PageVisiti / \sum PageVisiti$. This method not only considers the number of the audience to the multimedia file, but also takes into account the communication performance.

Summary

The biggest difference between network media and traditional media is that the nonlinear structure, the interaction function of the network, and the extensive existence of hypermedia links, make it in a large state of disorder for a long time. It will be a very difficult task to carry out the order statistics of the complex content dissemination performance. The significance of this work is very important, and its results can solve the current online media development and dissemination of various problems in the process, Bayesian method has played an important role in the analysis of communication performances. Bayesian statistical method for the use of prior distribution data, become a reasonable supplement to the lack of posterior sample to compute the communication performances, comprehensive consider of many factors, to calculate communication performance of the specific each page, and communication performance of the media file. The calculation process is objective and comprehensive, so as to get the scientific and objective evaluation results.

This article is on the basis of the existing statistical construction method of communication performance statistics, according to the statistics method and computer subject. It puts forward a case in computer science and technology disciplines background; describe the application of this statistical method and preliminary implementation. This paper emphasizes the feasibility. In the next step, we will focus on how to use and maintain the statistical database of the communication performance.

References

[1] Fan Long, Content analysis method applied in the study of network communication, Information science.05 (2010) 67-70.

[2] Bai Bing, Digital statistics of the network communication effect, Journalism of University, 08 (2001);

[3] Hu Chunling, Based on interesting frequent patterns of Bayesian network computation and pruning, Journal of software, 12 (2011) 43-47;

[4] Yin Yuhong, The similarities and differences between online advertising and traditional media advertising, Academic review, 07 (2011) 32-37;

[5] Yuan Zhigang, Mining massive scientific data based on Bayesian theory, Computer Software andTheory, 02 (2013)22-25;

[6] Jin Chanming, Research on web crawler module of search engine, Modern computer, 03(2010)66-68;

[7] Rajiv N. Rimal, Adrienne H. Chung, Nimesh Dhungana, Media as Educator, Media as Disruptor: Conceptualizing the Role of Social Context in Media Effects, J Commun, 2015, Vol.65 (5);

[8] Xiao-Lin Wu, Daniel Gianola, Guilherme J. M. Rosa, Bayesian model averaging for evaluation of candidate gene effects, Genetica, Genetica, 2010, Vol.138 (3), pp.395-407;

[9] Hideaki Ishigami, Relative age and birthplace effect in Japanese professional sports: a quantitative evaluation using a Bayesian hierarchical Poisson model, Journal of Sports Sciences, 2016, Vol.34 (2), pp.143-154;

