Dean&Francis

ISSN 2959-6157

Application of Markov Chain in Information Retrieval

Kecan Zhu[1, *]

[1]The Affiliated International School of Shenzhen University, Shenzhen, 518000, China

*Corresponding author: Keshuai.Zhu_30P@tsinglan.org

Abstract:

This paper introduces the fundamental role of Markov chain algorithm in recommendation system, especially its "memory-free" property, that is the future state only depends on the current state and has nothing to do with historical data. This feature is useful for simulating user behavior on the platform. The researchers defined the concept of "status" and selected key metrics based on how users interact with search results (like likes, favorites, time spent watching videos, etc.). These indicators are crucial for building predictive models. The paper also discusses the application of Markov chains in recommendation systems, including how transfer probabilities are used to guide information retrieval behavior and population retrieval history is used to weigh these probabilities. In addition, the potential of combining Markov chains with other algorithms and models is also explored. Finally, the abstract emphasizes the importance of considering user data and privacy ethics when developing software and concludes that Markov chain algorithms are an important foundation of recommendation systems, capable of predicting user preferences and enhancing user experience. The paper will show how to use Markov chains to make more accurate predictions.

Keywords: Markov chain; information retrieval; personalized recommendation.

1. Introduction

In short video software or search engines such as TikTok and Google, that people usually use, or even in people's usual typing, people can often feel that this software seems to anticipate what they want to see and recommend it very accurately. For example, when a person types, the software already has the whole word when the person just enters half of the words. Most of the time, it is very accurate. In some papers, experiments have been done, and the results show that most people are satisfied with the results and will use what is recommended [1]. This algorithm - Markov chain - has also become the core competitive strategy of short video software and Internet companies [2].

This paper concentrates on the workings of the Markov chain algorithm, which is fundamental to the recommendation systems used by this software. The Markov chain's basic characteristic is its "memory-

less" property, which means that the future state is solely dependent on the present state, independent of historical data [3]. This feature is important in modeling user behavior as they transition between different items, content, or state.

In order to use the Markov chains in recommendation systems, first of all, technical personnel must establish a clear definition of what is a "state". In the user's engagement, a state can be defined by the duration of user interaction with an application or specific content, such as the time spent viewing a video, the act of liking or favoriting it, or the browsing of web pages or products [4]. By defining these states and capturing user preferences, a model of user behavior can be constructed. These models can forecast the user's next action, and this system is now often used in the software.

The application of Markov chains in recommendation systems often is the subject of extensive research and experimentation. For example, Cao et al. proposed a Markov model - based information retrieval (IR) model that uses transition probabilities to guide information retrieval behavior and uses population information retrieval history to weigh these probabilities [5]. Yang et al. described the performance of local search algorithms in discrete optimization problems [6]. In addition, Green created a recommendation system based on the Markov chains and the type grouping (RSMCG), which uses the Markov chains to predict the user's next actions and takes into account the user's recent actions [7].

Furthermore, the integration of the Markov chains with other algorithms and models has been discussed by Yang and Rannala, who emphasizing the importance of mixing many kinds of computing methods and a range of software together [8]. A variety of data is available to determine users' likes and preferences and train this model. The sequential recommendation mentioned by Roberts can not only predict the content what the user is likely to be interested in, but also predict the sequence of the user's behavior at a specific time or situation [9]. The hybrid recommendation system mentioned in this paper is a very important system. By combining similarity models (such as collaborative filtering) and the Markov chains, the accuracy and efficiency of the recommendation system are improved. Similarity models can help systems understand the relationship between users and items, while the Markov chains can capture sequential dependencies in user behavior, which means the user's next action depends on the current action or previous action [10].

Experiments also take into account the ethics of user data and privacy as people compute and explore deeper algorithms, more in-depth research, complexity, and personalized recommendations. This affects the user experience and the user's trust in the product. This is also very important when developing software.

In conclusion, the Markov chain algorithm is a very important foundation of a recommendation system, providing a powerful system for predicting user preferences and enhancing the overall user experience. This article will introduce how to use Markov chains to make predictions.

2. Methods

2.1 Data Source and Description

The data sources of this study mainly include two parts: one is the experimental data in existing literature, and the other is the data collected through the self-designed experiment of this study. The former belongs to secondary data and may have some bias, but it provides basic background and preliminary analysis for the study. The latter is first-hand data, which has higher accuracy and authenticity through direct experimental collection and provides direct evidence for research.

2.2 Indicator Selection and Description

In this paper, the selection of key indicators mainly focuses on the interactive behavior of users to search results, such as the likes, favorites, viewing time of video content, and further interaction behavior of similar content. In addition, the e-commerce platform also includes the user's purchase behavior of the product and the purchase tendency of similar products. These metrics are crucial for building predictive models and are relatively easy to collect through experiments or log analysis to obtain more accurate first-hand data.

Table 1. Search engine Data in China (2024/08/30~2024/09/28)

Search engine	Page view	Number of visitors	Bounce rate	Average access duration
Baidu	11,721,209	7,322,119	83.45%	00:01:54
What browser	1,825,819	970,709	75.99%	00:01:01
360 research	2,045,281	999,545	70.95%	00:01:36
Sogou	824,491	470,026	73.02%	00:02:11

ISSN 2959-6157

Google	72,053	19,308	47.06%	00:05:45
Quark	10,839	3,866	63.94%	00:01:15
Headline search	1,852	860	75.07%	00:01:40
Bing	158	73	68.75%	00:04:14

Table 1 shows that in China, most people will choose Baidu to search for things, so this paper will focus on Baidu.

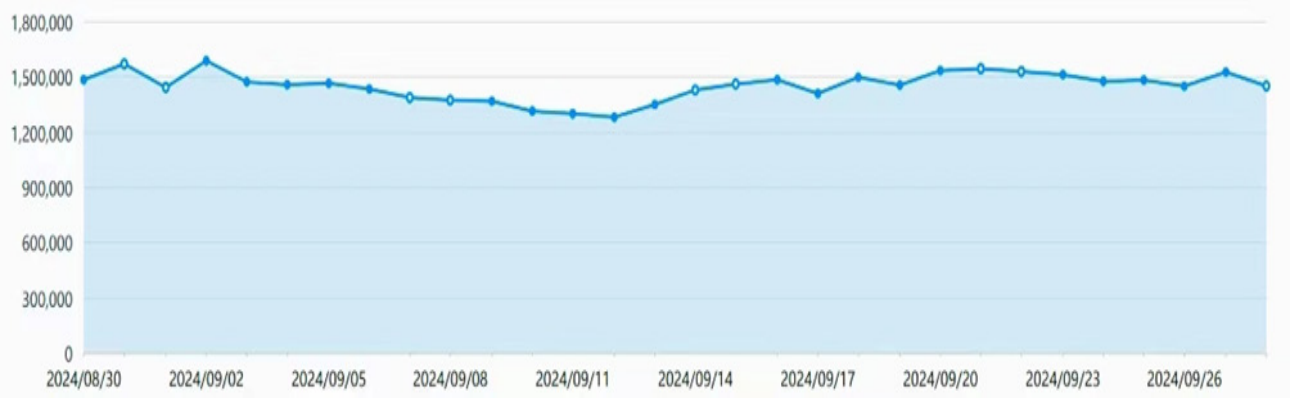

Fig. 1 Page views of Baidu (2024/08/30~2024/09/28)

Figure 1 is a line chart, which shows that the page views of Baidu are almost the same every day, and the ups and downs will not be large, indicating that the predicted quantity gap will not be large.

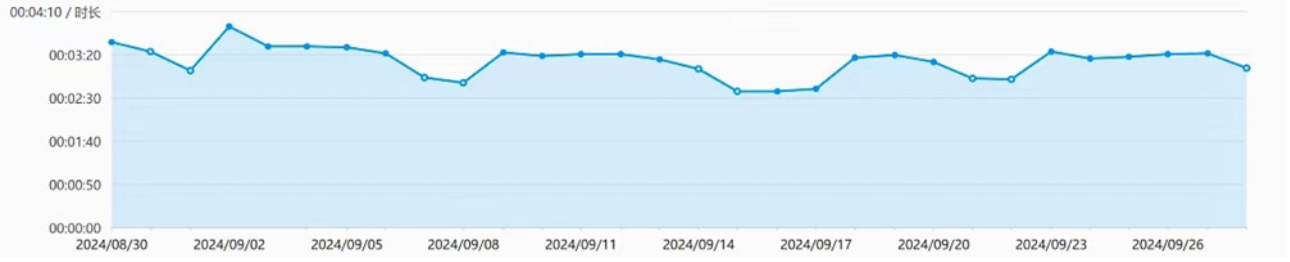

Fig. 2 Average visit duration of Baidu (2024/08/30~2024/09/28)

Figure 2 is also a line graph, and the conclusion is the same as that of Figure 1, indicating that the Average visit duration of Baidu does not change very much every day, so it is easier to predict the subsequent user behavior.

Table 2. Data of new and old customers of Baidu Data (2024/08/30~2024/09/28)

Classify	Page view	Percentage of page views	Visit times	Number of visitors	Bounce rate	Average access duration
New Visitors	19,756,537	45.44%	12,817,680	10,979,325	75%	00:02:08
Returning Visitors	23,717,634	54.56%	12,020,061	8,624,297	67.49%	00:04:18

Table 3. The source of people who browse Baidu (2024/08/30~2024/09/28)

Source	Page view	Number of visitors	Visit times	Bounce rate	Average access duration
Direct access	21,945,556	7,792,535	9,682,877	58.84%	00:05:08
Other ways	20,599,422	12,081,478	14,628,243	41.16%	00:03:12

Table 2 and 3 are the main charts, which contains the initial data of two states, one is the number of people who

visit Baidu directly, and the other is the number of people who use other ways to visit Baidu.

2.3 Method Introduction

In this study, the Markov chain model is used to analyze user behavior, and a state transition matrix is constructed. First, all possible states of the Markov chain are clearly defined to ensure that these states are independent of each other and cover all possible uses interaction scenarios. Then, based on the experimental data, the transition probability of each state is calculated, and the corresponding state transition matrix is constructed. In the construction process, each row of the matrix is verified to ensure that the sum of the state transition probabilities is 1, so as to meet the basic requirements of probability theory. The model has been applied in a number of short video platforms and shopping software and can effectively predict user behavior and provide personalized recommendations for users, thereby improving users' purchase intention and satisfaction.

3. Results and Discussion

3.1 Calculate

First, this paper defines two states, one is "direct access" (search the URL directly or use a bookmark to access), and the other is "access by other means" (jump from another hyperlink). Since the table is the sum of the previous 30 days, people need to find the average number of days first.

$$Averagenumberofdirectvisitorsperday = \frac{7,792,535}{30} \approx 259,751 \tag{1}$$

$$Averagenumberofvisitorsperdaybyotherways = \frac{12,081,478}{30} \approx 402,716 \tag{2}$$

Determine the initial state vector (which is the known data in the table). This initial state vector V contains the number of daily visitors for both access methods. This vector represents the starting point of the prediction, which is the current state:

$$V_0 = \begin{bmatrix} 259,751 \\ 402,716 \end{bmatrix} \tag{3}$$

In this paper, the transition probability is given the bounce rate and the average access time as an approximation of the transition probability because there is no direct transition data. It is assumed that visitors who visit directly have a 41.16% probability of visiting through another website on their next visit. In addition, visitors who visit other sites have a 58.84 percent chance of going directly to them on their next visit. The transition matrix P is:

$$P = \begin{bmatrix} 0.5888 & 0.4112 \\ 0.4116 & 0.5884 \end{bmatrix} \tag{4}$$

Finally, P can be used in the transition matrix to predict the number of visitors tomorrow.

$$V_1 = V_0 \times P \tag{5}$$

$$V_1 = \begin{bmatrix} 259,751 \\ 402,716 \end{bmatrix} \times \begin{bmatrix} 0.5888 & 0.4112 \\ 0.4116 & 0.5884 \end{bmatrix} = \begin{bmatrix} 318,699.2944 \\ 343,752.9024 \end{bmatrix} \approx \begin{bmatrix} 319,000 \\ 344,000 \end{bmatrix} \tag{6}$$

The result means that about 319,000 people visited Baidu through direct search on September 29, while about 344,000 people visited through other means.

3.2 Analyse and Advises

As the data shows, forecasts are based on historical data, so that researchers can continue to improve the model. By analyzing past user behavior, researchers can predict future user behavior. This approach assumes that past patterns of behavior will continue in the future. The analysis points out that most users who use Baidu access Baidu through other websites, which means that Baidu's traffic is largely dependent on recommendations from external websites. This is an important finding because it points to a key channel through which Baidu gets traffic. Based on this discovery, Baidu can adopt the following strategies to optimize promotional ways and increase user access.

Partnerships or Advertising on other platforms, such as websites, blogs, news portals, and so on, could be established with more websites or social media companies to place Baidu search boxes or create hyperlinks of Baidu to direct users to search on Baidu, which makes it easier for users to access Baidu. Baidu can use social medium broad user base to attract more visitors.

There is also a range of other ways. Baidu can increase partners, encourage bloggers and other content creators to actively recommend Baidu, and increase the enthusiasm for promotion through commission or reward programs. In addition, Baidu can optimize its search engine ranking to ensure that it can also get a high ranking in other search engines, which can attract users to click. Thirdly, Baidu needs to ensure the quality of its search results and website loading speed, so that users are willing to visit and use it for a long time through external links, which can increase the number of return visitors.

4. Conclusion

By continuously analyzing user visit data and understand-

ISSN 2959-6157

ing which external websites users visit Baidu through, Baidu can focus on optimizing these channels. Through these strategies, Baidu can increase the number of users who visit external websites, thereby increasing the overall traffic and market share. Predictive analytics reveal the future needs of users. Baidu can develop new features and services based on these predictions to ensure that they meet the expected needs of users. Ongoing predictive analytics provide insight into market trends and user behavior. Baidu can use this information to adjust its marketing strategy to adapt to market changes. If the forecast shows that traffic may decline. Baidu can take measures in advance, such as optimizing search algorithms or increasing advertising, to prevent user loss.

References

[1] Afeng Wang, Yiming Zhao, Yijin Chen Information Search Trail Recommendation Based on Markov Chain Model and Case-based Reasoning. Pages 228-241. In 2020 ASIS&T Asia-Pacific Regional Conference (Virtual Conference), 2021, 12-13.

[2] Gerald Benoît. Application of Markov chains in an interactive information retrieval system. Information Processing and Management, 2005, 843-857.

[3] Alexander G. Nikolaev, Sheldon H. Jacobson. Using Markov chains to analyze the effectiveness of local search algorithms. Journal of Heuristics, 2011, 160-173.

[4] Fatma Mlika, Wafa Karoui. Proposed Model to Intelligent Recommendation System based on Markov Chains and Grouping of Genres. Procedia Computer Science, 2020, 868-877.

[5] Guihong Cao, Jian-Yun Nie, and Jing Bai. Using Markov Chains to Exploit Word Relationships in Information Retrieval. University of Montreal, 2005.

[6] Yeongwook Yang; Hong-Jun Jang; Byoungwook Kim. A Hybrid Recommender System for Sequential Recommendation: Combining Similarity Models With Markov Chains. IEEE Access (Volume: 8), 2020, 2169-3536.

[7] Green P J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 1995, 82(4): 711-732.

[8] Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology & Evolution, 1997, 7: 717-724.

[9] Roberts A F M S O. Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society, 1993, 55(1): 3-23.

[10] Neal Radford M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational & Graphical Statistics, 2000.