History-related Content Recommendation in Social Networks based on User's Interest Estimation

Kouki Yoshida
Dept. of Computer Science,
Takushoku University
Japan
r88483@st.takushoku-u.ac.jp

Adam Jatowt
Dept. of Computer Science,
University of Innsbruck
Austria
adam.jatowt@uibk.ac.at

Yasunobu Sumikawa
Dept. of Computer Science,
Takushoku University
Japan
ysumikaw@cs.takushoku-u.ac.jp

ABSTRACT

Knowing and analyzing the events of the past is important as it allows us to understand the formation process of modern society. Twitter users are able to acquire and share information of their interests including also content on history. We propose an algorithm to recommend history-related content according to the interests of Twitter users. Our approach analyzes users' past tweets to determine their interests by modeling the forgetting curve of memory. It determines the users' current interests by considering the time passage and the repetition of posted content on particular topics.

KEYWORDS

social media analysis; public history; collective memory; Twitter

ACM Reference Format:
Kouki Yoshida, Adam Jatowt, and Yasunobu Sumikawa. 2023. History-related Content Recommendation in Social Networks based on User's Interest Estimation. In *JCDL '23: The 23th ACM/IEEE-CS Joint Conference on Digital Libraries.* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

It is generally believed that history plays an important role in our society. The reason for this is that by knowing history, we can understand the processes that shape the present, and give meaning and direction regarding the past. To let everyone learn about history, history education starts early on from elementary schools. Besides the structured educational curricula, many initiatives encourage history learning by communicating historical knowledge to the public, or by sharing historical findings by experts. In the recent years, social media has been used as one of the ways to effectively share history-related content with large numbers of users. Prior studies have analyzed this kind of activities taking place on Twitter in the context of commemorating WWI [2] or, from a more general angle, based on classifying history-related tweets into six basic categories of commemoration [7]. Other related study has also investigated and visualized how Twitter users refer to the past [5].

In this paper, we propose a novel chatbot system[1] operating on Twitter that recommends history-related content in a customized way according to the estimated interests of users. It analyzes the tweets that a given user posted in the past to determine what kind of topics and issues the user is likely to be currently interested in. Our approach to user interest modeling assumes that the scope of user's interests changes as time passes, and it prioritizes topical areas that are similar to the content tweeted more often by the user in the present than in the past. Furthermore, we assume that content that has been posted repeatedly for a long time is of stronger interest than the most recent one. We represent the above assumptions by modelling the memory forgetting curve [3] that describes the way in which content is retained through repeated learning.

Chatbots used for educations started also to appear recently. The common application is the study of foreign languages (e.g., [6]) under the assumption that learning languages need to be done in an interactive manner. The reader is referred to a survey paper for more details [1]. To the best of our knowledge, we are the first to propose a Twitter chatbot system for sharing history-related content.

2 HISTORICAL KNOWLEDGE BASE

To automatically extract past event descriptions we have decided to use Wikipedia. One form in which Wikipedia records important events are year and month-day articles which briefly describe as a list the key events that occurred in the corresponding time frames. We have used year pages for 1AD to 2019AD and day pages covering articles devoted to days from January 1 to December 31. In total, the past events' knowledge base we have prepared stores the descriptions of 71,374 events. The collected events cover wide range of topics, time periods and geographical areas, as well as they were collectively chosen by the Wikipedia editors as important events in each particular year or month-day. Note that despite its obvious problems, Wikipedia is considered as relatively accurate, and has been increasingly used in the history and memory science.

3 USER PREFERENCE ESTIMATION

If the user posts a reply text or direct message to our chatbot, the chatbot executes the proposed algorithm to share some past event's description. Our chatbot accepts a simple message pattern signaling user's wish to receive recommendation like "please tell me about history" as a reply text. Once the request has been received, the chatbot performs two procedures: collecting words from the user's past tweets and estimating interested topics from the words extracted from those tweets.

We assume that *when a Twitter user posts a certain tweet, the user is basically interested in the content of the tweet. The interest level however fades with time*. It is also assumed that *the content of repeated tweets forms continuous interest of the user*. Based on

[1] https://mobile.twitter.com/user_pref_hist

JCDL '23, June 26–30, 2023,

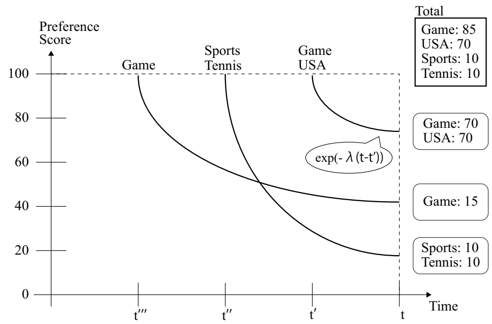

Figure 1: Example of term weighting in the user preference modeling.

these assumptions, we estimate what is currently of the most likely interest of the user through interest modeling by applying the concept of a forgetting curve of memory.

Weight Assignment for Collected Words from Past Tweets. First, we collect words extracted from past tweets excluding stop words and calculate their weights to predict the current interests of the user.

Fig. 1 provides an example showing how our model predicts the current interests of a Twitter user from the user's posted tweets. In this figure, the horizontal axis represents time. This example uses t''', t'', and t' as the times when a Twitter user tweeted. The vertical axis represents scores calculated by our model to estimate the user's interest based on the words contained in the tweet text at the present time (t). In our example, a user posted tweets containing the word "Game" twice, once in t''' and once in t'. When calculating the level of interest in "Game", instead of calculating the same weight for either of the two words, we reduce the weight as time passes. Considering the passage of time compared to t, we recalculate the weight (15) to be smaller than the value (70) of t' since the posts of t''' are spread farther apart in time.

Here, the text in the tweet posted at t'' includes "Sports" and "Tennis", which in this figure are smaller than the value of the "Game" weight of t''' tweeted before. This method takes into account the fact that the content of interest for each user tends to be different for each topic. For example, if we analyze user's past tweets and find that most of them are actually about games, then the game-denoting words in a given target tweet text would have an increased weight. On the other hand, if the rarely posted tweets are from sports, the weight of the words in those tweets should be reduced.

For calculating weights models the forgetting curve of memory, we use the exponential function exp. The larger the difference between the current time t and the time t' of the previously posted tweet, the smaller the value of interest; thus, this difference is given as the argument of exp. The following equation shows the proposed method to find the interest for a single tweet: $f(t, t', \omega, \lambda) = \omega \times exp(-\lambda(t - t'))$. λ is used as an argument of the function exp to adjust the weight of the difference between the current time and the time when each tweet was posted. If the total period of past tweets that could be collected is short, it is difficult to draw an appropriate forgetting curve; thus, the weights should be reduced. The λ is the value obtained by dividing 1 by the entire period of past tweets over which the posts could be collected. The ω is a parameter that indicates the size of the interest depending on the topic. We set the ω score as a similarity between the latest tweet and t.

To find the current most interesting content from past tweets T, we compute the sum of the above equation as follows:

$$Interest(t, w, \omega, \lambda) = \sum_{t' \in T(t)} f(t, t', \omega, \lambda) \times \delta(w, tweet(t'))$$

where w is a word and $tweet$ is a function to retrieve the body of a tweet posted at the time given in the argument. δ is a function that returns 1 if w is used in $tweet(t')$; otherwise, it returns 0. $T(t)$ is a function that collects timestamps of tweets posted before the tweet t given in the argument.

Next, we calculate weights for the collected words as the sum of all tweets. The equation is as follows:

$$W(w, lt, \lambda) = \sum_{t \in T(lt)} Interest(t, w, TopicSim(lt, t), \lambda)$$

where lt is the reply text to the chatbot. The function *TopicSim* calculates the ω score by applying any similarity measurement techniques such as LSA and cosine similarity.

Estimating Interested Topics. As our chatbot outputs past events that are similar to what collected words represent, the next step is to find the topics that the collected words represent. At this point, instead of simply performing a topic analysis on the collected words, we additionally use the types of words obtained from Wikipedia by applying TAGME [4]. This makes it possible to distinguish whether a particular name is a person, such as a writer or chemist, a song, or other entities. TAGME identifies named entities from text by linking words to their corresponding Wikipedia articles. We collect Wikipedia categories from the Wikipedia articles. We regard the collected words and their Wikipedia categories as a document, and then create the document's feature vector. Finally, we calculate cosine similarity between the feature vectors equipped with the LSA created for past events and the collected words, and we then output the most similar ones.

4 CONCLUSIONS & FUTURE WORK

We proposed a chatbot to present a history-related content in line with user interests by analyzing the interests of Twitter users based on their posted tweets and considering the passage of time.

In future we will analyze what type of history-related content attracts the interest of users. In particular, we plan to establish what kind of history effectively attracts interest depending on the time period in which the history is presented.

REFERENCES

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. An Overview of Chatbot Technology. In *Artificial Intelligence Applications and Innovations*. Springer International Publishing, Cham, 373–383.

[2] F. Clavert, B. Majerus, and N. Beaupré. [n.d.]. #ww1. Twitter, the Centenary of the First World War and the Historian. Twitter for Research 2015.

[3] Hermann Ebbinghaus. 1987. *Memory : a contribution to experimental psychology* (reprint ed.). Dover Publications.

[4] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities). CIKM, 1625–1628.

[5] Adam Jatowt, Émilien Antoine, Yukiko Kawai, and Toyokazu Akiyama. 2015. Mapping Temporal Horizons: Analysis of Collective Future and Past Related Attention in Twitter. WWW, 11 pages.

[6] Jiyou Jia. 2009. CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems* 22, 4 (2009), 249–255.

[7] Yasunobu Sumikawa and Adam Jatowt. 2021. Analyzing history-related posts in twitter. *Int. J. Digit. Libr.* 22, 1 (2021), 105–134.