머신 러닝을 사용한 개인화된 뉴스 추천 시스템

펭소니 [1], 양예선 [1], 박두순 [1*], 이혜정 [2]
[1] 순천향대학교 소프트웨어융합학과
[2] 순천향대학교 AI · SW 교육원

peng.sony61@gmail.com, parkds@sch.ac.kr

Personalized News Recommendation System using Machine Learning

Sony Peng[1], Yixuan Yang[1], Doo-Soon Park[1*], HyeJung Lee[2]
[1]Dept. of Software Convergence, Soonchunhyang University, South Korea
[2]Institute for Artificial Intelligence and Software, Soonchunhyang University, South Korea

peng.sony61@gmail.com, parkds@sch.ac.kr

요 약

With the tremendous rise in popularity of the Internet and technological advancements, many news keeps generating every day from multiple sources. As a result, the information (News) on the network has been highly increasing. The critical problem is that the volume of articles or news content can be overloaded for the readers. Therefore, the people interested in reading news might find it difficult to decide which content they should choose. Recommendation systems have been known as filtering systems that assist people and give a list of suggestions based on their preferences. This paper studies a personalized news recommendation system to help users find the right, relevant content and suggest news that readers might be interested in. The proposed system aims to build a hybrid system that combines collaborative filtering with content-based filtering to make a system more effective and solve a cold-start problem. Twitter social media data will analyze and build a user's profile. Based on users' tweets, we can know users' interests and recommend personalized news articles that users would share on Twitter.

1. Introduction

Digital technology, notably the emergence of smartphones, has dramatically increased the massive of data that can be retrieved online. In addition, social media sites such as Facebook, YouTube, Instagram, and Twitter are key data generators and the source of data information [1]. Users can access all information with just one click through their mobile phone or computer. However, even if they access the information or news quickly, it does not mean the user can get the right content of news they prefer or are interested in. Due to a large amount of news, the user might find it hard to get the news they want to read. With this type of overloading problem, they find a solution by applying Recommendation Systems. Recommendation systems have been widely used in many areas such as e-Commerce (Amazon, Google, AliExpress, eBay…), Entertainment (Netflix, YouTube, Spotify…), and Healthcare systems [2]. Recommendations have been used to assist the user in reducing their finding time and getting suitable items for their need.

In our work, we mainly focus on hybrid recommendation systems by combining collaborative filtering with content-based filtering, the proposed technique intends to improve system effectiveness and alleviate the cold-start problem. A user's profile is built using Twitter social media data. Then, depending on users' tweets, we may learn their interests and propose tailored news items. The news data must be the retrieve from the latest site such as Times, New York Times, BBC News, CBS NEWS, etc. The retrieval attributes will be scraping from these popular sites in the same format (CSV). Then convert it into data frame using panda’s library which available in python language. With the greatest help of Natural language processing.

2. Techniques in Recommendation System

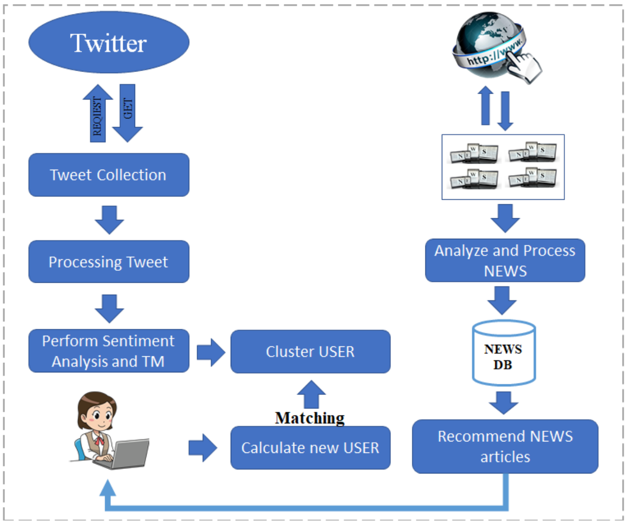
Recommendation systems are filtering systems that adjust information presented to a user depending on his/her interests, item similarity, frequently item sets and so on [3]. Moreover, recommendation systems are commonly utilized to suggest suitable movies, articles, news, restaurants, locations, products purchase, etc. Due to the real-world issues, researchers are offered various approaches related to recommendation systems and their own definition to solve the overloading problems which start from data mining methods to deep learning according to their defining problem [4]. For traditional recommendation systems, are categorized into 2 vital parts: Content-based filtering and Collaborative filtering. One is Content-based filtering refer to a system that recommends based on user choices that they decided in the past (experience) [5]. For instance, if readers read the news content and give some comments, it might recommend recent news (related content) based on their previous content to the target reader. The specific content (features) can be text, images, sound, etc. Through this analysis, a similarity between entities can be established for recommending articles/news similar to those that a user has purchased, visited, heard, seen, and rated highly or positively. Second, Collaborative filtering is a technique that allows users to give ratings about a set of elements (movies, songs, products, news…). It is used to identify relationships between users (user behaviors). For example, users give a rating to the items and product purchase history. Moreover, the system does not require many product features to work. Among them. Collaborative filtering is a widely used and famous technique based on users' similarities [6]. There are two types of collaborative filtering, including memory-based and model-based. The memory-based approach is a method that computes the similarity between users or items using the user's previous data based on ranking. Another model-based approach is used to predict unrated products by previous rating experience of the users. These algorithms include Matrix factorization, deep learning methods, a clustering approach, and more. Indeed, these two techniques have many benefits, but they also produce another disadvantage. In Collaborative filtering faced some problem including cold start (item or user), scalability, data sparsity and more. Thus, to solve this such issues, researchers are proposed other techniques call Hybrid approach to enhance the result and performance of recommendation systems in various ways. In ref [7] proposed personalized real-time movie recommendation system from practical prototype until evaluation period using user's demographic to divide into various clusters and the results could reduce time complexity and to improve recommendation performance. Since the era of big data, and research trending of deep learning and graph learning theory, ref. [8] applied graph-based features to recommendation systems. Moreover, they used autoencoder in feature extraction to extract new features based on attribute combination in movie fields.

3. Proposed News Recommendation System

In the session, we explain our news recommendation systems.

3.1. System Architecture

Figure 1 shows the design flow of our proposed work. Two main components are related to each other. The first one is Twitter Collection, and the second is NEWS DB. Next, we will briefly detail our work as follows.



(Figure 1) System Architecture.

3.2. Twitter Data

- Collect Twitter user's data (based on hashtags) and save it in the tweet Collection
- Analyze user's tweet

a. Data Preprocessing

The tweet contains userName, non-English words, punctuations, URLs, and emoji. Based on the data, tweets might be composed of multiple languages, emojis, or some links. So, there are several steps to remove these:

- Clean Tweets: Remove useless contents (URLs, numbers, emojis …)
- Remove Stop words: Remove common words in English (e.g., I, you, we, there).
- Tokenize tweet: Breaking stream of textual content into words.
- Stemming and Lemmatization: Apply text normalization to reduce derivational and inflectional forms of a word to a common base form.
  - Stemming: Chops off words without any context (e.g., eating: eat, cried: cry)
  - Lemmatization: Find the lemma of a word with the use of a vocabulary (e.g., are: be, better: good)

b. Cluster Users based on their similar interests

According to their similar interests in news based on their tweeting behaviors, we can cluster users who share similar interests from their timelines or tweets. Moreover, it requires vectorized representation of tweets. To perform it, we use the sklearn library, which calls TfidfVectorizer. TF-IDF refers to the Term Frequency-Inverse Document Frequency. TF-IDF refers to a weight that ranks the importance of a term in its contextual document corpus. The main idea of TF is to give the frequency of words in each user's tweets. On the other hand, IDF is to compute the weight of rare words across users' tweets. Lastly, perform K-mean to cluster users based on tf-idf matrix.

c. Perform Sentiment Analysis and Topic Modeling

With the help of TextBlob, the sentiment property returns a namedtuple of the form Sentiment (polarity, subjectivity). The polarity score refers to a float within the value [-1.0, 1.0] range. The subjectivity is about a float within the value [0.0, 1.0] range, where 0.0 is objective and 1.0 is subjective.

For Topic Modeling, LDA is performed. Tuning of some topics for each cluster was accomplished using the coherence measure from **gensim.models.coherencemodel**.

3.3. NEWS DB

We use a python library called Newspapers3k to retrieve news collection to scrape news websites. The various topic can be collected/Analyzed and then saved to the news DB. F

3.4. Making Recommendations

Finally, we recommend users based on their similarities and identify their interests in the various topic. New news or articles with the most similar topic to the group user cluster will be recommended.

4. Conclusion

This paper proposes a personalized news recommendation system to help users find the right and relevant content and suggest news that readers might be interested in. The proposed mechanism aims to build a hybrid system that combines collaborative filtering with content-based filtering to make a system more effective and solve a cold-start problem. First, Twitter social media data will analyze and build a user's profile. Then, based on users' tweets, we can know users' interests and recommend personalized news articles that users would share on Twitter. Based on this proposed mechanism, we will enhance recommendation performance and to solve cold-start problem.

참고문헌

[1] Cooley, D., & Parks-Yancy, R. “The effect of social media on perceived information credibility and decision making”. Journal of Internet Commerce, 2019, 18(3), 249-269.

[2] Qin, L., Xu, X., & Li, J. “A real-time professional content recommendation system for healthcare providers’ knowledge acquisition”. In International Conference on Big Data, Springer, 2018, 367-371.

[3] Mohamed, M. H., Khafagy, M. H., & Ibrahim, M. H. “Recommender systems challenges and solutions survey”. In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), IEEE, 2019, 149-155.

[4] Da’u, A., & Salim, N. “Recommendation system based on deep learning methods: a systematic review and new directions”. Artificial Intelligence Review, 2020, 53(4), 2709-2748.

[5] Reddy, S. R. S., Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. “Content-based movie recommendation system using genre correlation”. In Smart Intelligent Computing and Applications, Springer, 2019, 391-397.

[6] Nassar, N., Jafar, A., & Rahhal, Y. “A novel deep multi-criteria collaborative filtering model for recommendation system”. Knowledge-Based Systems, 2020, 187, 104811.

[7] Zhang, J., Wang, Y., Yuan, Z., & Jin, Q. “Personalized real-time movie recommendation system: Practical prototype and evaluation”. Tsinghua Science and Technology, 2019, 25(2), 180-191.

[8] Darban, Z. Z., & Valipour, M. H. “GHRS: Graph-based hybrid recommendation system with application to movie recommendation”. Expert Systems with Applications, 2022, 200, 116850.