Personalized News Recommendation using Classified Keywords to Capture User Preference

Kyo-Joong Oh*, Won-Jo Lee*, Chae-Gyun Lim**, Ho-Jin Choi*

* Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST),
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea

** Department of Computer Engineering, Kyung Hee University (KHU),
1-Seocheon-dong, Gyeonggi-do, Yongin-si 446-701, Republic of Korea

aomaru@kaist.ac.kr, mochagold@kaist.ac.kr, rayote@khu.ac.kr, hojinc@kaist.ac.kr

***Abstract*— Recommender systems are becoming an essential part of smart services. When building a news recommender system, we should consider special features different from other recommender systems. Hot news topics are changing every moment, thus it is important to recommend right news at the right time. This paper aims to propose a new model, based on deep neural network, to analyse user preference for news recommender system. The model extracts interest keywords to characterize the user preference from the set of news articles read by that particular user in the past. The model utilizes characterizing features for news recommendation, and applies those to the keyword classification for user preference. For the keyword classification, we use deep neural network for online preference analysis, because adaptive learning is necessary to track changes of hot topics sensitively. The usefulness of our model is validated through experiments. In addition, the accuracy and diversity of the recommendation results is also analysed.**

***Keywords*— Preference mining, keyword classification, deep belief network, user profile, news recommendation**

I. INTRODUCTION

Nowadays recommender systems are becoming an essential part of many mobile and web applications for smartphones and tablets. They generally aim to provide in-time, context-aware, personalized information services in order to increase product sales and user satisfaction. In most cases, a recommender system recommends unseen items (e.g., videos, movies, books, etc.) to users through the analysis of big data collected from big markets (e.g., YouTube, Netflix, Amazon, etc.). On the other hand, an enormous number of news articles are produced by various mass media and updated in every minute, motivating the need for personalized news recommender systems, which should be aware of the user's longer-term interest, shorter-term preference, business and situational context, and social relationships.

Due to their nature of short life-time usage, however, news should be treated differently from other items such as movies or products when building a recommender system. For example, collaborative filtering may not work because we do not have time to wait for collecting information on which news are popular among "similar" users, because new news become old news very soon, no good to recommend. Moreover, surveys report that people often select news articles based on the titles mainly, and read only three paragraphs in the body carefully [1, 2]. In addition, hot news topics change frequently, needing to consider the changes sensitively. User's long-term interests are also important for news recommendation. In the previous work, all of these aspects of news have not been considered seriously.

This paper proposes a neural network model to analyse user preference for news recommendation. The model extracts interest keywords to characterize the user preference from the set of news articles read by that particular user in the past. For the keyword classification, we use deep neural network for online preference analysis, because adaptive learning is needed to track changes of hot topics sensitively [3]. As a result, our approach would be able to recommend the right news at the right time.

The rest of this paper is organized as follow. Section II presents related work. Section III describes our proposed model of news recommendation based on deep neural network. Section IV shows experimental results. Finally, Section V concludes.

II. RELATED WORK

Currently, four techniques of document analysis are used for news analysis. Yang and et al. [4] propose a method to detect new events from the articles using term extraction technique. News categorization is studied using document clustering [5] and document indexing [6]. Document summarization [7] for news is adopted by Yahoo and Google for their news services. Recently, sentiment analysis [8] is often used for analysis of social networks.

This paper chooses to use the term extraction technique to capture user preference by identifying important keywords from the news articles read by the user. Keyword classification can be performed using traditional classification techniques such as Support Vector Machine [9], and neural network [10], or using traditional information retrieval technique such as Term frequency and Inversed Document Frequency (TF-IDF) as in the case of Lee and Kim [11]. In 2006, Blei and Lafferty introduced a dynamic topic model based on statistical approach to analyse topics of documents among in various documents [12].

For keyword classification in this paper, we will use a deep neural network, which consists of multi-layered perceptron, in line with the recent trend of deep learning such as Restricted Boltzmann Machine (RBM) and Deep Belief Network (DBN), known to improve the performance of learning and modelling in neural network [13].

III. Proposed Approach

This section introduces our news recommender system. Broadly speaking, our news recommender system consists of two parts: preference analysis and news recommendation, as shown in Figure 1.

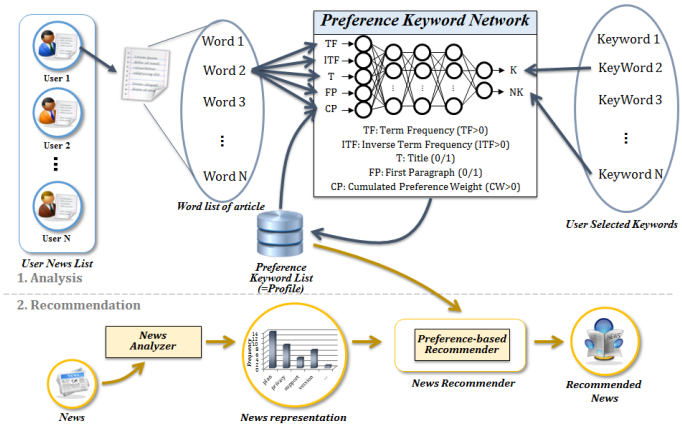

Figure 1. Overall process of the proposed approach

A. User Profiling Phase

The system analyses user interest keywords from read articles by user. From the analysis results, it recommends other personalized news. For the preference analysis, we used an analyser based on deep neural network to classify the interest keywords. Every word is classified into keywords or non-keywords by the proposed model. The concept of the proposed model is illustrated in Figure 2.

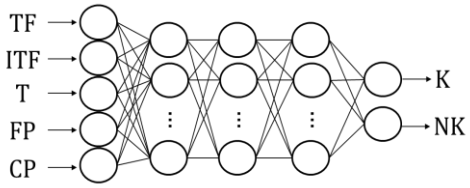

Figure 2. The proposed deep neural network model

There are five features consisting of the input layer. Term Frequency (TF) is the frequency of the word in the given document. Inverted Term Frequency (ITF) is an inverted value of the total frequency of the word in sample documents. This value helps to filter common words. Title (T) means existence of the word in the title of the given document, and First Sentence (FP) means existence of the word in the first sentence in the given document. Cumulated Preference Weight (CP) is long-term interest weight by words for keyword classification. Values of TF and ITF are greater than 0, value of CP is between 0 and 1, and values of T and FP are Boolean values.

Figure 3. Example of user profile (Oct. 2013)

The proposed model is consisted of 3-layer perceptron. Generally, RBM and DBN are consisted of 2~4-layer perceptron. We provide modification to the model from DBN for this study. We try to chase two hares, performance and precision of classification, using DBN.

In [3], they used a traditional one-layer neural network for keyword classification. In this case, the modelling was simple, and the result of classification had lower precision. Our approach has more layers of perceptron based on DBN for higher precision.

Especially, we don't use Inverted Document Frequency (IDF), because the IDF needs all the documents to compute the value, thus, it cannot calculate adaptively. Instead of using IDF, we compensate the IDF value using ITF. In addition, the CP involves the individual long-term interest of the keywords. The biggest advantage of the model is complexity of input layer. According to input features, it can be modified such as types and number of feature. Figure 3 shows some examples of users' profile.

B. News Ranking Phase

The user profiles are constructed by above model. We used the profiles for news recommendation in this step. For each user, every news article in the set is evaluated in its similarity to the user profile. First, when we crawled an upcoming news, we filtered noisy words and sentences to extract noun phases. Second, we compute the TF-IDF scores of the collected noun phases which are called by "news profile". Last, the similarity score between this "news profile" and the user profile is computed. In our implementation, we used cosine similarity for this purpose.

The Figure 4 illustrate the sequence how our recommender ranks the upcoming news. Finally, top-k news articles are recommended to the user.

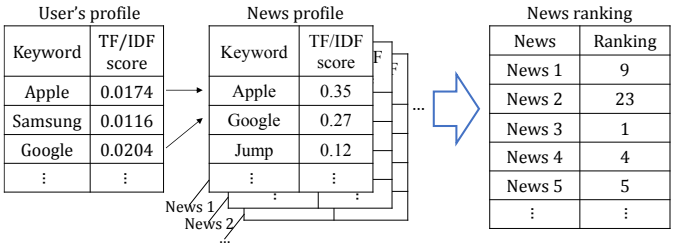

Figure 4. News ranking with user profiles

IV. EXPERIMENT

To validate the usefulness of the proposed method, we have implemented a prototype news recommendation service and performed a simple user study against a small group of users. Details of the experimental results are presented in this section.

A. Data Collection and Pre-processing

First of all, we collected history of read news. For the collection, we implemented two platforms: Google Chrome Extension "Daum News Tracker," and Android application "KECI News." Both implementations can download and install from each application market: Chrome web store and Google play. Basically, the data of collection is consists of timestamps, IP address, twitter ID, and URL of articles. Our service is supporting to identify the user via twitter open authentication (OAuth). Also, it is possible to use the service whoever doesn't have twitter ID via IP address. We have collected the data for two month for training and three months for testing. Table 1 shows the test data of 8 identified users. The whole dataset are shown at http://kecidev.kaist.ac.kr/.

TABLE 1. NUMBER OF READ ARTICLES

User	Month1 (Sep 2013)	Month2 (Oct 2013)	Month3 (Nov 2013)	Sum	Average
u1	127	79	168	**374**	**124.7**
u2	51	241	162	**454**	**151.3**
u3	51	103	29	**183**	**61.0**
u4	109	109	70	**288**	**96.0**
u5	846	513	607	**1966**	**655.3**
u6	34	145	49	**228**	**76.0**
u7	52	132	45	**229**	**76.3**
u8	100	191	91	**382**	**127.3**
Average	**171.3**	**189.1**	**152.6**	**513.0**	**171.0**

From the history data, we distributed the log by user or IP address. We extracted news IDs from the logs. We analysed the latest 50 articles to extract user profile. We parsed body of the articles from the HTML document using "Jericho HTML Parser 3.3." We analysed the Korean articles, so we needed morphology analysis to extract nouns using "JHanNanum."

B. Keyword Classification and User Profiling

We generated personalized profile for each user using above data. We used about 1500 articles data from about 70 users for the training; those are obtained from March to April in 2013. The labels in output layer were obtained from user directly for training and test evaluation. Our dataset contains anonymous users' data. Therefore, the user was anonymous user then, we used top 50% results of TF-IDF as keyword labels. Using the data, the model could calculate parameters of edges in the neural network using back-propagation algorithm adaptively. Based on the trained network, we tested other read articles set. The test data set was approximately 2000 articles of 50 users for about a month in September, 2013. Based on the result, we made user profile consisted of interest keywords and preference weights of the keywords.

TABLE 2. COMPARISON OF TRAINING TIME AND ACCURACY BETWEEN ADAPTIVE TF-IDF AND PROPOSED MODEL (5 USERS)

IDs	Adaptive TF-IDF		Proposed Model	
	Time (s)	*Accuracy (%)*	*Time (s)*	*Accuracy (%)*
ND****	113.4	56	182.6	75
Moc****	18.3	63	29.8	79
Gfbah****	28.8	68	46.1	80
Dohyoj****	18.4	64	29.4	75
Syl****	13.2	48	21.0	67
Average	**38.4**	**59.8**	**61.8**	**75.2**

Table 2 shows the comparisons of results: training time of test dataset and accuracy of interest keywords. We compared the results with Adaptive TF-IDF by [4], because our proposed model could be trained the dataset adaptively. The Adaptive TF-IDF predicts the IDF value by TF and ITF value of the words and their equations.

From the result, the learning time spent average 0.22 sec per article in Adaptive TF-IDF and average 0.32 sec per article in our model. It took about 45% more for training, because of multi-layered structure of deep neural network model. However, the accuracy of classification results was increased to a considerable degree, about more than 10% for every user. In our model, the TF and ITF values were considered as inputs and considered to classify the interest keywords similar with Adaptive TF-IDF. In addition, the position of the words and cumulated (long-term) preference of words also considered synthetically.

C. Personalized News Recommendation

Our news recommendation approach is contents-based recommendation. Already, we made the user profile using above proposed model. We tried to find recommendable news for each user based on the user profile. First of all, we collected the latest news set via implemented news crawler day to day. We also did same pre-processing for the each article in the latest news set such as parsing, noun extraction, and TF-IDF. Finally, we could get bag-of-words of articles and the importance weights of each words similar user profile. After these pre-processing, the recommender calculated cosine similarity every bag-of-words of the articles with user profiles. Then, every article could get the similarity score. Based on the score, the recommendable articles are ordered by rank. The recommendation results were delivered to each user.

TABLE 3. HIT RATIO (%) – RECOMMENDED USING PROFILES

User	Day 1	Day 2	Day 3	Day 4	Day 5	Average
User 1	40	40	60	70	50	**52.0**
User 2	50	50	30	50	40	**44.0**
User 3	30	50	20	50	80	**46.0**
User 4	60	80	60	80	70	**70.0**
User 5	50	60	60	80	90	**68.0**
User 6	50	80	60	50	70	**62.0**
User 7	50	20	40	40	20	**34.0**
User 8	60	60	40	60	50	**54.0**
Average	**48.8**	**55.0**	**46.3**	**60.0**	**58.8**	**53.8**

Table 3 shows hit ratio of the recommendation results for five days of qualified 8 users. We measured the reading rates from the top-10 ranked news. Then, average accuracy was achieved 54% for five days. The value is higher than recommendation of news randomly (37%).

TABLE 4. DIFFERENCES BETWEEN USER PROFILES (%)

User Profile	P (u1)	P (u2)	P (u3)	P (u4)	P (u5)	P (u6)	P (u7)	P (u8)
P(u1)	0	67	89	74	51	75	66	64
P(u2)	67	0	89	80	51	78	71	58
P(u3)	89	89	0	79	82	90	94	89
P(u4)	74	80	79	0	70	87	87	77
P(u5)	51	51	82	70	0	70	53	54
P(u6)	75	78	90	87	70	0	88	70
P(u7)	66	71	94	87	53	88	0	76
P(u8)	64	58	89	77	54	70	76	0
Average	**69.4**	**70.6**	**87.4**	**79.1**	**61.6**	**79.7**	**76.4**	**69.7**

Table 4 shows a matrix how different the interest keywords with each other. The values were average values for five days. At least 51% keywords in user profile were different from others. Based on the user profile, we could get personalized news.

TABLE 5. AVERAGED DIFFERENCES BETWEEN USER PROFILES – FOR FIVE DAYS (%)

Days	Day 1	Day 2	Day 3	Day 4	Day 5	Average
Differences	61	75	72	66	72	**69**

Table 5 shows the average values from above matrix each day. Average 69% of the interest keywords were different in user profile for five days such as Table 4.

TABLE 6. AVERAGED DIFFERNECES BETWEEN RECOMMENDED NEWS – FOR FIVE DAYS (%)

Number of Topic cluster	Day 1	Day 2	Day 3	Day 4	Day 5	Average
10	58	78	79	91	90	**79**
20	59	77	88	99	93	**83**
30	59	80	85	95	91	**82**

Table 6 shows diversity of the recommended news lists. We should consider articles which were similar topic, we analysed topic distribution of each article using LDA. We changed the cluster number and observed the differences. There are no quite big differences among the number of topics.

Table 3, 4, and 5 were indices how the recommended articles are personalized by our approach.

V. CONCLUSION

In this paper, we have proposed a model to classify interest keywords more suitable for extracting user preference of news topic using a deep neural network model based on DBN. For the contents-based news recommendation, we have gotten the personalized user profiles adaptively. To achieve the purpose we have considered 5 features of news. The features are affected positively to classify the words as interest keywords. The proposed model has supplemented chronic disadvantages of neural network model in learning and modelling, while having more accurate classification results. In addition, we also have evaluated accuracy and diversity of our news recommender. The results were meaningful for personalized news recommendation. In the future, we will consider the more factors and features to mining user preferences such as time, location and other contexts. Also, we will try to calibrate the long-term interest through other sources such as SNS usage data.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics Co. Ltd.

REFERENCES

[1] M. Amy, C. Leah, and R. Tom, "The tablet revolution and what it means for the future of news," *The State of the News Media 2011: An Annual Report on American Journalism*, Pew Research Center, Washington DC, 2011.

[2] M. R. Tom and C. Leah, "Mobile devices and news consumption: some good signs for journalism," *The State of the News Media 2012: An Annual Report on American Journalism*, Pew Research Center, Washington DC, 2012.

[3] Widyantoro, D. H., T. R. Ioerger, and J. Yen, "An Adaptive Algorithm for Learning Changes in User Interests," *Proceedings of the 8th ACM International Conference on Information and Knowledge Management (CIKM 1999)*, pp. 405-412, Kansas City, USA, Nov. 2-6, 1999.

[4] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 32-43, Jul.-Aug. 1999.

[5] V. Hatzivassiloglou, L. Gravano, and A. Maganti. "An investigation of linguistic features and clustering algorithms for topical document clustering," *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2000)*, pp. 224-231, Athens, Greece, July 24-28, 2000.

[6] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol.16, no.10, pp. 1279-1296, Oct. 2004.

[7] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, and S. Sigelman, "Tracking and summarizing news on a daily basis with Columbia's Newsblaster," *Proceedings of the Second International Conference on Human Language Technology Research*,, pp. 280-285, San Francisco, CA, USA, 2002.

[8] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment Analysis in the News," *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 2216-2220, Valletta, Malta, May 2010.

[9] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," *Advances in Web-Age Information Management*, Lecture Notes in Computer Science, pp. 85-96, Springer Berlin Heidelberg, 2006.

[10] T. Jo, M. Lee, and Gatton, T. M, "Keyword extraction from documents using a neural network model," *Proceedings of the International Conference on Convergence and Hybrid Information Technology (ICHIT 2006)*, vol. 2, pp. 194-197, Cheju, South Korea, Nov. 2006.

[11] S. Lee and H. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF," *The Journal of Society for e-Business Studies*, vol. 14, no.1, pp. 59-73, Nov. 2009.

[12] D. M. Blei and J. D. Lafferty, "Dynamic topic models," *Proceedings of the 23rd ACM International Conference on Machine Learning*, pp. 113-120, New York, USA, 2006.

G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Journal of Neural Computation*, vol. 18 no. 7, pp. 1527-1554, 2006.

Mr. Kyo-Joong Oh
He received a bachelor's degree of computer science in 2011 from Korea Advanced Institute of Science and Technology (KAIST). He is currently a MS/Ph.D. candidate student in the department of computer science at Korea Advanced Institute of Science and Technology (KAIST). His current research interests include artificial intelligence, data mining, software engineering, and recommendation system.

Mr. Won-Jo Lee
He received a bachelor's degree of computer science in 2012 from Hanyang University. He is currently a MS candidate student in the department of computer science at Korea Advanced Institute of Science and Technology (KAIST). His current research interests include artificial intelligence, data mining, and big-graph mining.

Mr. Chae-Gyun Lim
He received a bachelor's degree of medical computer science in 2011 from Eulji University. He is currently a MS candidate student in the department of computer engineering at Kyung Hee University (KHU). His current research interests include artificial intelligence, data mining, and biomedical informatics.

Prof. Ho-Jin Choi
He is currently an associate professor in the Dept. of Computer Science at KAIST. In 1982 he received a BS in Computer Engineering from Seoul National University, Korea. In 1985 he got an MSc in Computing Software and Systems Design from Newcastle University, UK. And in 1995, he got a PhD in Artificial Intelligence from Imperial College, London, UK. From 1982 to 1989, he worked for DACOM, Korea, and between 1995 and 1996 worked as a post-doctoral researcher at Imperial College. From 1997 to 2002, he served as a faculty member at Korea Aerospace University, Korea. He moved to Information and Communications University (ICU), Korea, from 2002 to 2009. And since 2009 he has been with the Dept. of Computer Science at KAIST. Between 2002 and 2003 he visited Carnegie Mellon University, Pittsburgh, USA, and served as an adjunct professor the Master of Software Engineering (MSE) program. Between 2006 and 2008, he served as the Director of the Institute for IT Gifted Youth at ICU. Since 2010, he has been participating in the Systems Biomedical Informatics National Core Research Center at the Medical School of Seoul National University. Currently, he serves as a member of the board of directors for the Software Engineering Society of Korea, for the Computational Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics.