

PADAC²: Real-Time News Recommendation System with Heterogeneous Social Footprints

Dong-Yup Kang^{*1}, Dong-Kyun Han^{*1}, Gyumin Sim^{^2}, JongHyuk Jung^{#2}, HyunKi-Jeon^{^2}, Soobin Lee^{*}, Joonyoung Park^{*}, Seunghyeon Moon^{*}

^{*}KAIST Institute, [^]KAIST Computer Science, [#]KAIST Industrial Design, Daejeon, Republic of Korea

^{*}{dykang, handk, soobinlee, jypark, shmoon}@itc.kaist.ac.kr, [^]{gyumin92, heretor92, jeonhk}@kaist.ac.kr
(The authors¹ equally contributed on the paper, and other authors² interned at KAIST Institute)

Abstract—Everyday in Korea, more than hundred thousands of News articles and postings are generated by either writers or users. Many people read News articles and write their opinions on the articles through major News portal systems such as Naver or Daum. However, they are sometimes time-consuming, biased, and distracted by unnecessary information. We propose a real-time News recommendation system called PADAC², that is more passive process for users to browse their interests from massive News media. We propose a recommendation algorithm called HeteRoCommender based on heterogeneous source of social footprints given.

Keywords— News Recommendation, Social Evidence, Media Aggregation, Collaborative Filtering

I. INTRODUCTION

In South Korea, there are more than hundreds of News press and magazines (e.g., Joogang ilbo and Donga ilbo), dozens of famous internet communities (e.g., Ilbe, Wootda) and social media services (e.g., Twitter, Facebook). Daily, more than hundred thousands of articles and postings are generated by either writers or users. Indeed, we are living in a deluge of information.

Many people in Korea read News articles and write their opinions on the articles through major News portal systems such as Naver [1] or Daum [2] (Figure 1). The sequence of actions is called *News Browsing activity* in our paper. When they have spare time or needs for browsing News or other media, they visit the portal sites, read some headlines and click the original article when they like to read more carefully. While this browsing activity is more active process triggered by users' interests, it is sometimes (1) time-consuming or tiresome until they got what they want, (2) biased by the information provider who arranges the headlines even though a bunch of articles are not covered, and (3) distracted by other advertisements or unnecessary information. Such problems may make our News browsing activities unpleasant and annoying.

We propose a real-time News recommendation system called PADAC², that is more passive process for users to browse their interests from massive News media. Our system recommend real-time trending/breaking and personalized News articles to each user. The meaning of the name, PADAC², is the sound flapping wings by fishes or birds

pronounced in Korean. To express the freshness of our recommended News articles metaphorically, we named our system. In this paper, we focus on proposing a novel recommendation algorithm. In addition, we also propose a general architecture of our system and Web and Mobile interfaces that are actually developed to release our service publicly.



Figure 1 Snapshot of a major portal site: some headlines, search ranks, ads, and categories.

The most famous algorithm for item recommendation task is collaborative filtering (CF) [4]. The basic idea of CF is to recommend items that similar groups of users are interested in. Even though CF shows increasing success on not only academic researches but also practical recommendation system such as Amazon or Netflix, it assumes that the user-item matrix data should be collected first. That is, the recommendation system at initial stage should be not applicable for CF. Also, CF is very vulnerable to real-time items such as News articles and its dynamics.

We propose a recommendation algorithm called **HeteRoCommender** based on heterogeneous source of social footprints given. People keep a record of their browsing activities with different forms: The social footprints mean other meta-sources of information collected by many users and their voting activities such as Like in Facebook, ReTweet in Twitter, comments in articles, and the rank of search queries collected by major portal sites such as Naver and Daum. Then, we combine our algorithm with existing rating prediction method [6].

The recommended articles would be retrieved by REST API calls and the results are displayed in our mobile and web version of News browsing system. We also developed

Flipboard like interface for users to read and explore their interests on magazine style interface

Finally, we evaluated real-timeness and browsing effectiveness of our system compared to existing portal based browsing system. The results show that our system is much more effective than commercial portal system in terms of browsing and fast access of real-time news.

Our major contributions of this paper are:

- 1) We designed and developed a real-time News recommendation system called PADAC² to release actual news recommendation service in South Korea.
- 2) We proposed a novel recommendation algorithm called HeteRoCommender with socially gathered footprints such as search ranks, Tweets, Facebook postings, and postings from many domestic communities, and combined the heterogeneous information with existing collaborative filtering based algorithm.
- 3) We evaluated real-timeness and browsing effectiveness of our system with mobile and web users. Moreover, all the data that we have been collecting would be publicly released for research purpose.

II. RELATED WORKS

There are some existing News recommendation services: major News applications (e.g., NYT, CNN, YTN), News summarization service (e.g., Summly), and News curation service (e.g., FlipBoard, Pulse, GeniNews). We have five criteria to compare our system with existing services: personalization, information source, service type, environment, and business model. In summary, PADAC² supports personalized alerts from multiple source of information on mobile and web interfaces. Also, by subscription to our services, people who want some analytic functions are provided data analysed information by our data mining modules.

Table 1. Comparison of Existing News Recommendation Service

	Major News	Curation Service	Summarization Service	PADAC ²
Personalization	-	Setting by users	Keywords by users	Personalized alerts
Information Source	single	multi	multi	multi
Service Type	Press	Manual filtering	Summarization algorithm	Analytic mining
Environment	Web/App	App	App	Web/App
Business Model	Ads	Ads	-	Subscription

III. DATA CRAWLING AND SYSTEM ARCHITECTURE

Since there are various media sources to crawl, we focus on designing flexible and extendible crawler. Mainly, our crawler is implemented in Java and leverage commonly used frameworks such as Apache Maven, Spring Framework, and Hibernate. The crawler is consisted of three main components: model, parser, and repository (see Figure 2).

There are various media services and each service has different data model. It is hard to handle heterogenetic data models in an application. Therefore, we design abstract, but contains all necessary information to recommend news, unified data model. We focus on unifying not only data model, but also name of similar elements. For example, we use a term 'favorCount' which represent users' favour. 'Like' in Facebook and 'retweet' in Twitter can be represented as 'favorCount' in our model.

Since news services have various layout, it is labour-intensive to implement parser for each service. Fortunately, a Korean portal site provides articles from various presses in the same layout. We implemented a parser for Naver news service to crawl news articles. Since community services use table structure web page, we implemented a parser which is specialized to parse table structure web pages. For each community we simply change the name of table elements. To crawl Twitter and Facebook, we use open APIs of each service to retrieve data.

Repository module manages database transactions. We implemented it as a single-tone pattern class by leveraging Spring-framework to loose coupling and cohesion between classes. After crawler parse data, repository saves data to the database. We re-use the repository module in HeteRoCommender as well due to its flexible design. Therefore, we ensure consistent database transaction APIs across PADAC² service. In the performance perspective, we guarantee quick responsibility by leveraging 2nd level caching in Hibernate framework.

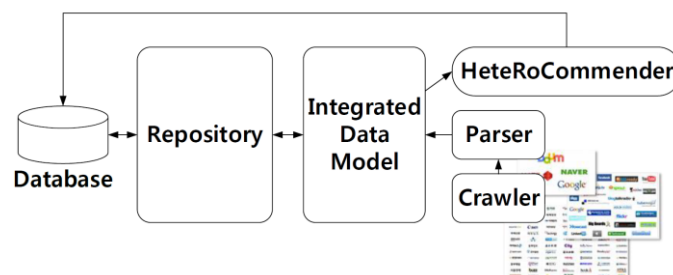


Figure 2 Architecture of PADAC²

The Table 1 shows the number of articles and postings that our crawler has been crawling from June 2013 until now. We have crawled more than twenty millions of News articles from hundreds of major and minor presses, two millions of Korean tweets and six hundreds of Facebook posts from major News pages. Also, more than hundreds of thousands of postings from major communities have been collected.

Table 2. Number of crawled items until September 14, 2013. (Crawled since around June, 2013)

News	Social Media		Ranks		Community		
	Twitter	Facebook	Daum	Naver	OU	Humor	Ilbe
20,151,079	2,002,868	619,537	163,865	156,159	52,942	170,301	161,282

IV. HETEROCOMENDER

A. Heterogeneous Social Footprint

Our system collects heterogeneous source of data from many media channels such as News, social media, search ranks, and communities. In this section, we record the social footprints from the heterogeneous sources.

If a user wants to discuss about an issue with other users or express their own feeling, she can submit a posting to media service (except news services which is managed by press). Posting is consisted of content, time, location tag, and hash tag. Location tag denotes where the user make posting. Recent services automatically trace the place where user submit a posting. Users can intentionally use hash tag when they want to discuss about common topic with other people. Therefore, hash tag is usually topic word. In Twitter, for example, hash tag is appeared with a character ‘#’ in posting content.

Postings are appeared in not only the writer’s timeline but also the writer’s neighbours’ timeline. The timeline is collection of postings that user is interested in. Users can discuss or talk about the posting through comment, a short message which appears with a posting. User can use location tag and hash tag in comment as well. Some media services such as community services support reply which is a posting as an answer for original posting.

If user interested in a posting, she can mark it as her favourite and share it with her near friends. For example, Facebook have ‘Like’ which denote a user is interested in the posting. Also she can share it with other users. Similarly, many Korean communities have ‘chucheon’, a Korean word means recommendation in English, to express a user’s interests.

B. Collaborative Filtering based on Heterogeneous Social Footprints

For real-time News recommendation, two important factors are considered: real-timeness and personalization.

First, we only cache recent articles and posting within 30 minutes into our memory. Then, if title or body of the articles or postings contain following keywords, we choose the articles or postings as global candidates for recommendation: "속보", "충격", "아찔", "섹시", "노출", "자극", "환상", "살인", "사고", "자극". The second step is to expand those seed keywords based on real-time search keywords collected by our crawler. For example, "한국시리즈 5 차전 123", "권상우 커피차", "길 먹살", "런닝맨 엑소", "로빈 기븐스" shows top five keywords ranked by a major portal service. We used those keywords to extract additional global candidates for recommendation. These two simple keyword based models help to solve the cold start problem of our recommendation system.

Finally, when certain number of users’ logs are collected, we have used Hidden Factors as Topics (HFT) models [6] to recommend personalized News to appropriate users. HFT model optimize the objective function of Mean Square Error from both textual information and rating information. Previously, LDA [7] based models discovers hidden dimensions called topics from text while latent factor recommendation systems [8] discovers hidden ratings from

users and items. The HFT model combines two approaches by restricting the latent factor model with corpus likelihood as a regularizer obtained from LDA. For brevity, please refer the paper if you are interested in the model. The learned topics are used to combine with basic collaborative filtering method that uses user-article matrix that our system has collected in the logs.

V. INTERFACE

In this section, we briefly introduce the web and mobile interface of our system. The figure 3 shows magazine style of our web interface. The layout is automatically aligned according to size of images or articles.

In the first page, we show title, image if exists, and summarized sentences. By doing so, we avoid the copyright violation. If you click an article, you can see large view of the clicked article (see Figure 4).



Figure 3 Snapshot of Web Interface for PADAC²: multi article view



Figure 4 Snapshot of Web Interface for PADAC²: single article view

VI. EXPERIMENT

We evaluated real-timeness and browsing effectiveness of our system compared to existing portal based browsing system. Real-timeness means how fast a system alerts breaking News or events, and browsing effectiveness means how well the information is provided.

For real-timeness, we collect top 50 keywords from Korean search ranks, and calculate the difference between time of the search keyword firstly appears in the rank and time of our system alerts the keyword. On average, we could get 19.45

seconds until users obtain alerts from our system. This time is very fast especially for the breaking News keywords.

For browsing effectiveness, we tested user study from ten non-technology related students. The figure 5 shows the results for three metrics: usability, distraction, and information quality. According to the result, our system has very comfortable interface for browsing. Also, our interface and architecture help avoiding distraction from Ads and other additional unnecessary information. Finally, the quality of provided information is very high.

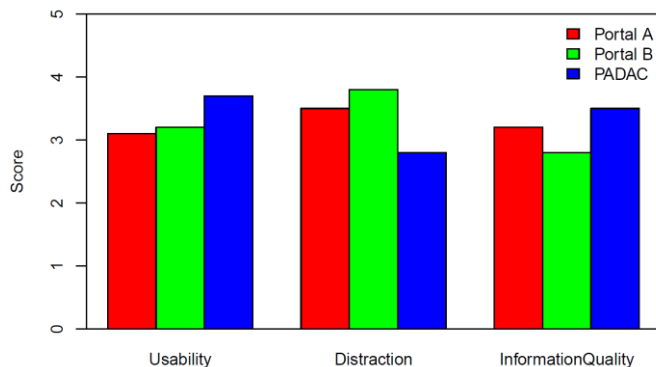


Figure 5 User Study of PADAC²

VII. CONCLUSIONS

In this paper, we propose a real-time News recommendation system from multiple sources of information such as social media, News, search ranks, and communities. We propose our crawling strategy and efficient architecture. Also, News articles are recommended by our proposed algorithm, HeteRoCommender that merges multiple heterogeneous sources of information with existing collaborative filtering methods. Finally, we propose web and mobile version of our interface. The user studies show that our system provide high-quality articles without distraction compared to existing portal systems.

ACKNOWLEDGMENT

This research was funded by the MOTIE(Ministry of Trade, Industry & Energy), Korea in the ICT R&D Program 2013 under Grant No. 10035166

REFERENCES

- [1] <http://www.naver.com/>
- [2] <http://www.daum.net/>
- [3] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
- [4] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295). ACM.
- [5] Flipboard.com/
- [6] J. J. McAuley, J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text", RecSys 2013.
- [7] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. JMLR, 2003.
- [8] Y. Koren and R. Bell. Advances in collaborative filtering. In Recommender Systems Handbook. Springer, 2011.