Building Rich User Profiles for Personalized News Recommendation

Youssef Meguebli[1], Mouna Kacimi[2], Bich-liên Doan[1], and Fabrice Popineau[1]

[1] SUPELEC Systems Sciences (E3S), Gif sur Yvette, France,
{youssef.meguebli,bich-lien.doan,fabrice.popineau}@supelec.fr
[2] Free University of Bozen-Bolzano, Italy,
mouna.kacimi@unibz.it

Abstract. Nowadays, more and more people are using online news platforms as their main source of information about daily life events. Users of such platforms have access to an increasing amount of articles of different topics, stories, and view points. Thus, a news personalization service is needed to filter the flow of available information and satisfy users needs. To this end, it is crucial to understand and build accurate profiles for both users and news articles. In this paper, we propose a new approach that exploits users comments to recommend articles. We build the profile of each user based on (1) the set of entities he talked about it in his comments, (2) and the set of aspects related to those entities. The same information is extracted from the content of each news article to create its profile. These profiles are then matched for the purpose of recommendation. We have used a collection based on real users activities in four news web sites, namely The Independent, The Telegraph, CNN and AL-Jazeera. The first results show that our approach outperforms baseline approaches achieving high accuracy.

Keywords: User modeling, Personalization, News recommendation

1 Introduction

Media platforms, like CNN [1] and Al-Jazeera [2], deliver the latest breaking news on various topics about everyday events. The rich content of such platforms and their easy access make them a leading information source for Internet users. Typically, besides reading news articles, media platforms offer the possibility for users to write their comments, express their opinions, and engage in discussions with other users. However, before reacting to any content, users need first to find news articles of interest. This task can be challenging since, in many cases, a user may not even know what to look for. Consequently, there is a need for personalized news services that recommend articles based on user profile. The accuracy of personalized recommendation depends mainly on how well user profiles are defined. Naturally, users' comments represent a valuable

[1] http://www.cnn.com
[2] http://www.aljazeera.com/

information source since they reflect not only interesting entities for users but also more details about which entity aspects they are interested on. Therefore, several past studies have exploited, in different ways, users' comments for news recommendation [1–5, 7–9, 11]. Most of the approaches use tweets [2–4, 7, 11] and few others [1, 5, 9] exploit users' comments on news websites. hmueli et. al., [9] restrict user profile to a set of tags extracted from related comments. Abbar et. al., [1] build the profile of each user based on the set of entities he has commented on with their related sentiments. While the proposed approach is interesting, it does not exploit all available information in users' comments and thus it provides incomplete profiles. The reason is that a user can be interested to a specific entity when it is related to a given aspect and can be not interested on it when it concerns another aspect. For instance, we can have a user who is interested by the entity *Tunisia* when it is related to the aspect *Tourism* and be not interested when it is related to the aspect *Election*. In this paper, we propose a personalized news recommendation approach that pays particular attention to interesting aspects for each entity. To this end, we introduce a new approach that models the profile of users and articles based on a set of tuples representing entities and their aspects. The idea is to have a fine-grained description of users and articles regarding general topics together with more specific issues. The profile of a user is extracted from the set of comments he provides in the news platform, and the article profile is extracted from its content and described by a set of tuples $(entity, aspect)$. We define each profile by two main components: (1) *entities* which reflect well defined concepts such as persons, locations, organizations, objects, etc., and (2) their related *aspects* representing entity attributes or any abstract object. These profiles are then matched to recommend to each user the list of articles that match with user profile interests and the current article he is reading. We evaluate our approach using four real datasets including, The Independent [3], The Telegraph [4], CNN [5] and Al-Jazeera [6]. The experiments show that our approach outperforms baseline approaches with a large margin, in term of precision and NDCG.

2 Related Work

Exploiting user generated content in social networks to define users' interests have been extensively studied [2, 4, 7, 10, 11]. Stoyanovich et. al., [10] leverage the tagging behavior of users to derive implicit social ties which were shown to serve as good indicator of user's interests. Chen et. al., [3] exploits user Tweets to build a bag-of-words profile for each Twitter user. Abel et al., [2] build hashtag-based, entity-based, and topic-based user profiles from Tweets, and show that semantic enrichment improves the variety and the quality of profiles. Other approaches [4, 7] address the problem of extracting topics of interest in micro-

[3] http://www.independent.co.uk/
[4] http://www.telegraph.co.uk/
[5] http://www.cnn.com
[6] http://www.aljazeera.com/

blogging environments. Hong et.al., [4] train a topic model on aggregated messages to improve the quality of topic detection in Tweets. Michelson et. al., [7] use a knowledge base to disambiguate and categorize the entities in user Tweets and then develop users profiles based on frequent entity categories. Our work does not fall in the previous classes since we exploit richer and longer comments than Tweets. Thus, we relate our work to the second class of approaches [1, 5, 9] which exploit users' comments on news websites to build user profiles. Li et. al., [5] enrich the content of each news article using users' comments and use the enhanced content to improve the accuracy of recommendation. However they do not build any user profile which results in a limited accuracy. Shmueli et. al., [9] restrict user profile to a set of tags extracted from related comments using a bag-of-words model. The closest work to ours is by Abbar et. al., [1] who build the profile of each user by extracting the set of entities he has commented on and their related sentiments. While the proposed approach is interesting, it does not exploit the different aspects of entities to have a more precise profile. In our work, we model user profile as set of interests reflected by the conjunction of *entities* and *aspects*. Another line of research related to this work is recommender systems [1–3, 5, 8, 9]. Two main strategies of recommender systems have been adopted and mostly combined in previous works. First, content filtering strategy creates a profile for each user or seed article and then recommends the best matching articles based on the user profile, the seed article, or both. Second, collaborative filtering strategy relies only on past user behavior without requiring the creation of explicit profiles. In our work, we adopt a content filtering strategy to recommend news articles to users based on their profile and potentially also on the article they are currently reading.

3 Personalized News Recommendation

3.1 Problem Definition

Our goal is to propose a personalized news recommendation model tailored to users' interests. Typically, interests represent the conjunction between entities and their related aspects. Entities reflect well defined concepts such as persons, location, organizations or objects, for example *"Aalborg"*, *"UMAP"*, and *"United Nation"*. While aspects reflect some specific issues related to the list of entities such as *"illegal immigration"*, *"recommender systems"*, or *"humanity acts"*. In our setting, we identify the interests of a given user based on the comments he has posted on the news platform. Using this information, the personalized news recommendation works as follows: Given a target user who is reading a seed article, we recommend a set of news articles that (1) are similar to the seed topic article for not deviating far away from user's interests and (2) match with specific issues that interest the user profile. The idea behind is to select, first, new articles that belong to the same topic than the seed article and then choose a subset that match with user interests. Formally, we define U as the set of users of a given news platform, and A as the set of articles provided by the news platform. Each user $u_i \in U$ provides a set of comments C_i about a set of

articles A' where $A' \subset A$. We assign to each user u_i a profile P_{u_i}, extracted from the set of his comments C_i, which reflects his specific issues about what he reads in the past. Similarly, we assign to each article a_j a profile P_{a_j} extracted from its content. When user u_i is reading article a_j , we proceed as follows. First, we compute the similarity between the article profile P_{a_j} and the profiles of the set of articles A_t where $A_t \subset A$ and A_t corresponds to all the articles that were published in time interval t. In this way, we can restrict our search space to any time period specified by the user. The time interval can range from a few days to months depending on user needs. The set of articles A_t is then sorted from the most similar article to a_i to the least similar one resulting in list L_1. Second, we compute the similarity between the user profile P_{u_i} and the profiles of the articles contained in the set A_t, thus, providing another sorted list L_2 from the most similar article to user profile P_{u_i} to the least similar one. As a last step, we aggregate the two lists L_1 and L_2 to obtain the final list of sorted articles from which we recommend the topk articles to user u_i.

3.2 Modeling User and Article Profiles

Due to the fact that both user comments and articles can express different types of information, including objective and subjective ones, we model both contents in the same way using the same structure for their profiles. To this end, we start by transforming the content of each comment (article) to a set of sentences S , using OpenNLP[7]. From each sentence, we extract two main components. First, a set of *entities*, where entities represent well defined concepts such as persons, locations, organizations, objects, etc. For example, given the sentence *"Obama is wrong to give work permits to young illegal immigrants"* we extract the entity *"Obama"*. Second, we extract the set of aspects, where aspects can be entity attributes or some abstract objects. In the previous example, the set of aspects are: *"Work permit* and *"illegal immigration"*. Note that for extracting entities we have used OpenCalais[8] and for aspects we have used Zemanta [9] to process a huge corpus containing $1,101,094$ Wikipedia articles.

3.3 Profile Similarity Measure

We have adopted cosine similarity to compute the similarity between profiles. This measure has been shown to be very effective in measuring similarity and detecting novelty between news articles [6]. In a standard search problem, a news article or user profile is represented by a vector of n dimensions where a term is assigned to each dimension and the value of the dimension represents the frequency of the term in the profile. In our setting we are interested in computing similarity between profiles described by a set of tuples, for this end we modify the vector representation as follows: each profile is represented by one vector

[7] http://opennlp.apache.org/
[8] http://www.opencalais.com
[9] http://www.zemanta.com/

representing the set of tuples and the value of each dimension represents the frequency of the tuple on news article or user profile.

4 Experiments

4.1 Setup

We have crawled a dataset based on the activities of 164 users from The Independent news site. The choice of this site was based on the fact that it has a large number of active users that continuously post comments on articles of various topics. Additionally and more importantly, users of The Independent follow also other news websites including The Telegraph, CNN and Al-Jazeera, so they have access to different types of articles covering different aspects for the same entity. For each of those users, we have crawled his comments in the four news sites mentioned earlier. Additionally, we have collected all the articles commented by each user from May 2010 to December 2013. Statistics about the number of comments and articles from each news web site are shown in Table 1. To evaluate our approach, we have randomly selected 23 users. For each user we performed recommendation at different time points $t_1, t_2, ..t_n$. The reason behind time dependent evaluation is two fold: (1) to take into account profile updates since users continuously post comments bringing new information about their interests, and (2) to use data before time point t_i for recommendation and data starting from time point t_i for assessment, as described later. The time points $t_1, t_2, ..t_n$ are chosen in such a way that between t_{i-1} and t_i, there is at least m comments posted by the user. In our experiments, we have set $m = 100$ to have enough evidence that the user profile needs to be updated. This setting resulted in 189 rounds of recommendation. We have simulated the recommendation system in the following way. For each user and at each time point t_i, we build the user profile based on his comments posted before t_i. Then, we choose as a seed article the first article that the user commented after time point t_i. We choose an article commented by the user to make sure that it matches user's interests. Based on the seed article and the user profile we return a set of articles that are similar to the seed article and at the same time have similar interests as the expressed in the user profile. Figure 1 shows the distribution of articles by topic. We can see

#Comments	482, 073
#Independent articles	26, 096
#Telegraph articles	23, 154
#CNN articles	535
#Aljazeera articles	303

Table 1: Datasets Statistics

that most articles and comments concerns politics. Note that the list of the seed articles we have selected follow a very similar distribution to the overall set of

articles. To assess the effectiveness of our approach we have used an automatic evaluation to avoid the subjectivity of manual assessments. We have considered the action of commenting on an article to be an indicator that the article fits the interests of the user. Based on this assumption, we check the list of recommended articles. The one that user has commented on are considered relevant. Note that it is probable that we systematically underestimate the interest of the user . A person might well be interested in an article even though he does not comment on it.

4.2 Results

We use two baselines strategies to assess our approach. The first one is based on aspect-centric profiles for both users and articles. The aspects were generated from users' comments and news articles content using Zemanta Api as we described earlier. The second strategy is based on entity-centric profiles for both users and articles. This strategy has been proposed in [1] and it represents our second baseline. We compare the two above strategies to our contribution where we define a global profile for both users and articles. To compare the results of the different strategies, we use Precision and NDCG at k ($P@k$ and $NDCG@k$). The $P@k$ is the fraction of recommended articles that are relevant to the user considering only the top-k results. It is given by:

$$P@k = \frac{|Relevant\_Articles \cap topk\_Articles\_Results|}{k}$$

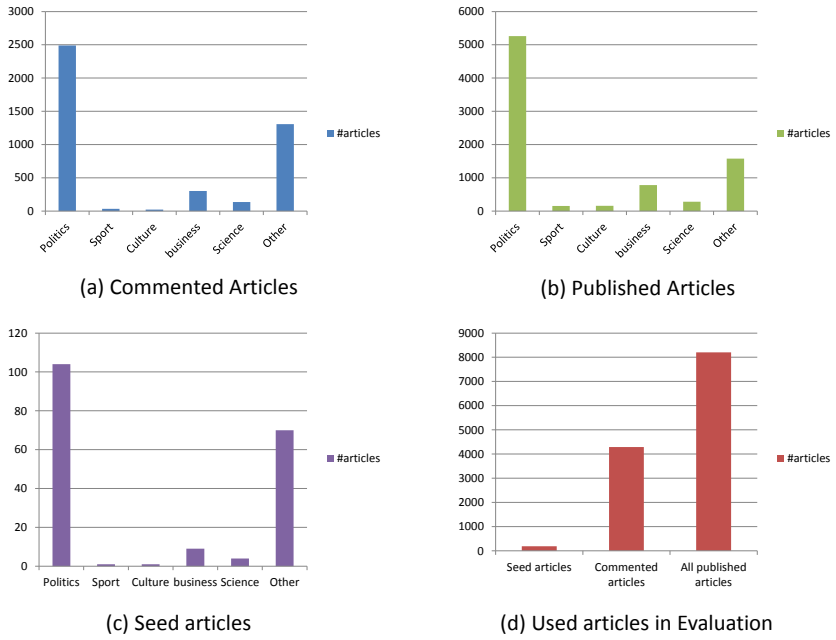

Fig. 1: Statistics about categories of used articles in Evaluation

Additionally, we compute $NDCG$ to measure the usefulness (gain) of recommended articles based on their (geometrically weighted) positions in the result list. It is computed as follows:

$$NDCG(E,k) = \frac{1}{|E|}\sum_{j=1}^{|E|} Z_{kj} \sum_{i=1}^{k} \frac{2^{rel(j,i)}-1}{log_2(1+i)}$$

where Z_{kj} is a normalization factor calculated to make $NDCG$ at k equal to 1 in case of perfect ranking, and $rel(j,i)$ is the relevance score of a news article at rank i. In our setting, relevance scores $rel(j,i)$ have two different values: 1(relevant) if the news article was commented by the user u, and 0(not relevant) if the news article was not commented by the user u. The precision and $NDCG$ results for the three strategies are shown in Table 2. We can clearly see that

	P@5	**P@10**	**NDCG @5**	**NDCG @10**
Aspect-centric Profile	0.396	0.392	0.734	0.689
Entity-centric Profile [1]	0.412	0.409	0.806	0.768
Global Profile	**0.52**	**0.507**	**0.855**	**0.797**

Table 2: Precision and NDCG values for all users

our approach of using global profile outperforms the baseline approach with a gain of 10% in terms of precision and 5% in term of ranking at NDCG@5. We also observe that using only aspects to build user and article profiles performs worst. The reason is that most of the news articles do not address certain aspects without relating them to some entities. Thus, disregarding entities leads to poor results. Moreover, when viewpoints are expressed about entities, they usually refer to certain aspects of those entities. Thus, using only entities to build profiles penalizes the performance. Consequently the combination of both entities and aspects give the best results. Note that real precision values must be higher than the one presented here. The reason is that comments can tell us if a user is interested in an article or not but their absence does not mean the opposite.

5 Conclusion and Future Works

In this paper, we have proposed a personalized news recommendation approach that takes into account fined-grained users interests. Existing approaches used only tags and entities to model interests which does not contain complete information. Thus, we have proposed a new model for user and article profiles based on entities and their related aspects. We have performed experiments based on four news websites, namely The Independent, The Telegraph, CNN and Al-Jjazeera. The results show that using both entities and aspects in the profile outperforms both entity-centric and aspect-centric approach with a minimum

precision gain of 10% and 5% in term of ranking at $NDCG@5$. This work represent a first attempt for a personalized news recommendation based on user and article viewpoints. As future works, we plan to test our model with larger set of users. It is also very promising to explore diversification techniques to improve our model by recommending articles outside of the current scope of the user profile.

References

1. S. Abbar, S. Amer-Yahia, P. Indyk, and S. Mahabadi. Real-time recommendation of diverse related articles. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1–12, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, UMAP'11, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.
3. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.
4. L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.
5. Q. Li, J. Wang, Y. P. Chen, and Z. Lin. User comments for news recommendation in forum-based social media. *Inf. Sci.*, 180(24):4929–4939, Dec. 2010.
6. Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang. Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 57–66, New York, NY, USA, 2011. ACM.
7. M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.
8. O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.
9. E. Shmueli, A. Kagian, Y. Koren, and R. Lempel. Care to comment?: Recommendations for commenting on news stories. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 429–438, New York, NY, USA, 2012. ACM.
10. J. Stoyanovich, S. Amer-yahia, C. Marlow, and C. Yu. Leveraging tagging to model user interests in del.icio.us. In *In AAAI SIP*, 2008.
11. J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.