Users' Reading Habits in Online News Portals

Cagdas Esiyok
Technische Universität Berlin
DAI-Lab, Ernst-Reuter-Platz 7
10587 Berlin, Germany
cagdas.esiyok@tu-berlin.de

Benjamin Kille
Technische Universität Berlin
DAI-Lab, Ernst-Reuter-Platz 7
10587 Berlin, Germany
benjamin.kille@tu-berlin.de

Brijnesh-Johannes Jain
Technische Universität Berlin
DAI-Lab, Ernst-Reuter-Platz 7
10587 Berlin, Germany
jain@dai-lab.de

Frank Hopfgartner
Technische Universität Berlin
DAI-Lab, Ernst-Reuter-Platz 7
10587 Berlin, Germany
frank.hopfgartner@tu-berlin.de

Sahin Albayrak
Technische Universität Berlin
DAI-Lab, Ernst-Reuter-Platz 7
10587 Berlin, Germany
sahin.albayrak@tu-berlin.de

ABSTRACT

The aim of this study is to survey reading habits of users of an online news portal. The assumption motivating this study is that insight into the reading habits of users can be helpful to design better news recommendation systems. We estimated the transition probabilities that users who read an article of one news category will move to read an article of another (not necessarily distinct) news category. For this, we analyzed the users' click behavior within *plista* data set. Key findings are the popularity of category *local*, loyalty of readers to the same category, observing similar results when addressing enforced click streams, and the case that click behavior is highly influenced by the news category.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: selection process

Keywords

Click Behavior, News Category, User Modeling

1. INTRODUCTION

Newspapers have established digital news portals to provide the audience news contents. These portals attract more and more visitors. This might be due to the digital news portals' ability to provide breaking news amongst other factors. The volume of available news confronts visitors with a selection problem. Digital news portals have introduced news recommendation services to support users in such situations.

News recommendation exhibits some particularities compared to other domains. According to Billsus and Pazzani [2], these particularities include dynamic contents, required novelty, shifting user preferences, and brittleness. In addition, news recommender systems face highly sparse data. Most users interact with a small fraction of available news items. This scenario becomes especially severe when users visit the news portal for the first time. In such settings, the system has to infer user preference based on the initially visited article.

Due to the high sparsity, news recommenders typically incorporate various types of additional knowledge into their systems (see Section 2). We suggest to incorporate dynamic data into news recommender systems that take general reading habits into account. The reason is that reading news article is a sequential process. At each point the reader decides which article to read next. We consider this sequential decision process in a more coarser setting. Digital news portals – as their analog counterparts – have grown accustomed to group articles into categories such as, for example, politics, sports, and local. The sequential decision process we consider reduces to the level of news categories rather than to the article level. Considering all readers, the question at issue is, how likely it is that a random reader moves from one news category to the next.

We model this sequential process as a Markov process and estimate the transition probabilities between news categories. Then we analyze user behavior using the *plista* data set [10]. The survey shows that the transition probabilities are not uniformly distributed. The implication of this finding is that incorporating users' reading habits in terms of estimated transition probabilities between news categories can improve news recommender systems. The latter issue, however, is out of scope of this contribution.

This paper begins with Section 2 as a brief review of the related works. In Section 3, we outline the *plista* data set and the methods which we used. Results of our preliminary findings of on-going work are discussed in Section 4. Finally, we conclude our study and discuss our future work in Section 5.

2. RELATED WORKS

In this section, we present existing work on the use of transition matrices for recommender systems. Additionally,

IIiX'14, August 26 - 29 2014, Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637038

S \ C	Car	Education	Technology	Leisure Activity	Health	Culture	Politics	Local	Trip	Sports	Business	SUM
Car	14	5	63	12	50	45	78	442	15	93	62	879
Education	5	541	110	28	95	43	205	879	7	190	161	2264
Technology	81	213	2705	243	991	390	1703	6590	75	1868	954	15813
Leisure Activity	15	49	273	1942	456	150	272	2170	44	347	200	5918
Health	52	210	889	704	1545	358	1100	4831	74	1429	863	12055
Culture	25	90	635	116	364	404	601	2717	15	636	241	5844
Politics	90	306	2227	341	1440	759	9557	11983	60	4719	3259	34741
Local	377	971	6169	3035	5004	2497	13572	162038	411	17796	5859	217729
Trip	13	13	59	43	49	21	61	398	35	93	60	845
Sports	132	217	2171	445	1517	770	5217	16126	99	46338	1993	75025
Business	61	269	1251	286	1072	363	1869	5391	50	1674	2236	14522
												385635

Chi-square (χ^2) = 214427.55 and P-value (α) < 0.005 where degree of freedom (df) = 100

Figure 1: Heat map illustration of the matrix which denotes the number of transitions.

we mention approaches suggested for news recommendation. Paparrizos et al. [11] investigate transition between job positions. Chen et al. [7] suggest to model the recommendation task as random walk. Hereby, transition probability matrix plays a central role. The authors evaluate their framework on movie ratings. Agarwal [1] investigates learning to rank methods applied to graphs. Providing ranked lists of entities, recommender systems can adopt learning to rank. The author mentions transition matrices incorporated to random walks as suited input to learning to rank procedures. Neither of these works target news recommendation.

Recommending news articles represents a challenge. Collaborative filtering techniques typically suffer from high sparsity which is apparent in the news domain. Thus, previous works suggest to augment the available data from other sources. These additional data sources include contents [3], semantic data repositories [4, 5, 6], location data [12], and micro-blogs data [8].

3. METHODS

This section describes the *plista* data set we use in our survey and formalizes the sequential decision process of a reader in terms of a Markov process.

3.1 Data Set

The *plista* data set has been released as a part of the ACM RecSys'13 Challenge on News Recommender Systems [13], in order for researchers to be able to develop novel recommendation algorithms due to this data set. The data set contains all interactions on 13 news portals corresponding to a time frame of one month ranging from June 1 - 30, 2013. The data ought to support researchers who are interested in cross-domain news recommendation, user modeling, and other related research topics. For further details about the evaluation scenario, the reader is referred to [9].

In order to start our investigation, we restricted our focus on an individual news domain[1] among 13 news domain. 385,635 transitions in total (see Figure 1) were generated from 4,258,277 impressions[2] which occurred in a time frame of one week ranging from June 1-7, 2013. All impressions of this individual news domain were classified into eleven main categories in order to be able to extract the users' click streams and set the transition matrix by means of these click streams. We have also drawn 162,192 items of click[3] collection in total – stored between 1st and 30th of June, 2013 – and then set a transition matrix (see Figure 2) based on click collection so as to compare it with the transition matrix based on impression collection.

3.2 Finite Markov Chains for News Categories

We are interested in how likely it is that a random reader decides to move from reading an article of one news category to reading an article of another news category. We model this process as a time discrete random process satisfying the Markov property.

The states S of the Markov process form a finite set consisting of the different news categories, such as, for example, politics, sports, and local. The states represent the relevant information we have about the reader.

The transition function of our Markov chain describes the probability that a random user who is reading an article of news category s_t at time t will move to read an article of news category s_{t+1} at time $t+1$. According to the Markov property, the transition probability takes the form

$$P(X_{t+1} = s | X_1 = s_1, \ldots, X_t = s_t) = P(X_{t+1} = s | X_t = s_t),$$

where the X_i are random variables at time i taking values from the finite set S of new categories. We call the current state at time t source news category and the next state at time $t+1$ clicked news category hereafter.

[1]According to statistics of *Alexa.com*, the domain is amongst the top 500 German web pages with respect to traffic.

[2]Whenever a user clicks on a news in news portal, an impression item is created in *plista* data set.

[3]Whenever a user clicks on a news in the recommended news list, a click item is created in *plista* data set.

S \ C	Car	Education	Technology	Leisure Activity	Health	Culture	Politics	Local	Trip	Sports	Business
Car	0.010	0.015	0.083	0.000	0.010	0.015	0.083	0.683	0.000	0.024	0.078
Education	0.000	0.081	0.068	0.007	0.035	0.004	0.170	0.515	0.000	0.035	0.086
Technology	0.001	0.000	0.272	0.007	0.016	0.018	0.109	0.461	0.013	0.053	0.050
Leisure Activity	0.000	0.001	0.116	0.286	0.036	0.018	0.046	0.450	0.000	0.026	0.021
Health	0.000	0.001	0.092	0.078	0.177	0.008	0.085	0.392	0.028	0.039	0.099
Culture	0.000	0.001	0.148	0.009	0.021	0.123	0.087	0.529	0.001	0.042	0.039
Politics	0.001	0.000	0.112	0.003	0.010	0.012	0.287	0.420	0.002	0.117	0.035
Local	0.000	0.000	0.065	0.007	0.010	0.010	0.064	0.781	0.001	0.039	0.022
Trip	0.000	0.001	0.098	0.020	0.099	0.005	0.082	0.527	0.045	0.054	0.068
Sports	0.000	0.000	0.057	0.002	0.007	0.006	0.061	0.139	0.001	0.717	0.010
Business	0.001	0.002	0.123	0.006	0.054	0.012	0.192	0.425	0.003	0.065	0.118

Figure 2: Heat map illustration of transition matrix based on click collection of *plista* data set.

S \ C	Car	Education	Technology	Leisure Activity	Health	Culture	Politics	Local	Trip	Sports	Business
Car	0.016	0.006	0.072	0.014	0.057	0.051	0.089	0.503	0.017	0.106	0.071
Education	0.002	0.239	0.049	0.012	0.042	0.019	0.091	0.388	0.003	0.084	0.071
Technology	0.005	0.013	0.171	0.015	0.063	0.025	0.108	0.417	0.005	0.118	0.060
Leisure Activity	0.003	0.008	0.046	0.328	0.077	0.025	0.046	0.367	0.007	0.059	0.034
Health	0.004	0.017	0.074	0.058	0.128	0.030	0.091	0.401	0.006	0.119	0.072
Culture	0.004	0.015	0.109	0.020	0.062	0.069	0.103	0.465	0.003	0.109	0.041
Politics	0.003	0.009	0.064	0.010	0.041	0.022	0.275	0.345	0.002	0.136	0.094
Local	0.002	0.004	0.028	0.014	0.023	0.011	0.062	0.744	0.002	0.082	0.027
Trip	0.015	0.015	0.070	0.051	0.058	0.025	0.072	0.471	0.041	0.110	0.071
Sports	0.002	0.003	0.029	0.006	0.020	0.010	0.070	0.215	0.001	0.618	0.027
Business	0.004	0.019	0.086	0.020	0.074	0.025	0.129	0.371	0.003	0.115	0.154

Figure 3: Heat map illustration of transition matrix based on impression collection of *plista* data set.

4. RESULTS & DISCUSSIONS

4.1 Chi-squared Test of Independence

Figure 3 represents the transition matrix, and shows the estimated transition probabilities. As can be easily seen from Figure 3, there is not a uniform distribution. So as to determine whether users' click behavior is influenced by the category of source news in a click stream of user's reading list (for example, in click streams, some users mostly read articles from category politics at first, and then articles from category sports.), we applied chi-squared test of independence. We deal with the matrix shown in Figure 1 as if it is a 11x11 contingency table. According to chi-squared test of independence, chi-squared test statistic is 214,427.55, while critical value for chi-squared distribution equals 140.169 where significance level is 0.005 and degree of freedom is 100. We therefore reject the null hypothesis that users' click behavior is independence and assume that the next news category depends on the current news category; since 214,427.55 is greater than the critical value of 140.169, and P-value is less than significance level of 0.005.

4.2 Popularity of Category Local

As can be seen from the transition matrices, for each category except sports, a great majority of audience clicks on a news which belongs to category local after reading a news.

4.3 Loyalty to the Same Category

Figure 1 shows the remarkable high value of the total number of the transitions (i.e., 227,355 transitions among 385,635) where source news category and clicked news category are the same. It presents that the source news and the clicked news are in the same category, with a percentage of 58%. We can observe in Figure 3, audience of sports and local categories are more loyal to their category than the other categories (that is, they insistently read the news in the same category as sports and local, respectively); on

the other hand, audience of some categories, such as culture, could be very open to new categories.

4.4 Similar Results with Enforced Streams

In addition to transition matrix based on the impression collection of the *plista* data set, we have also generated the transition matrix (see Figure 2) which depends on the click collection of *plista* data set. This is because we wanted to compare the transition matrices in order to analyze the differences arising from the fact that we get a click stream which is enforced by the recommender system indeed, when we address the click collection of *plista* data set instead of impression collection. As a result of this comparison, we have noticed that transition matrices are so similar; which means that although a recommender system forces the users for clicking recommended news, users' click behaviors seems not to be influenced by the system, i.e., they keep on reading the news in accordance with their interests.

5. CONCLUSION & FUTURE WORKS

This preliminary study of our ongoing work aims to investigate the users' news reading habits and the relations between the category of source news and the category of clicked news in *plista* data set. Within this study, we presented that the categories of the news have a strong influence on the users' click behavior. That is to say, news read by users follow certain patterns; for example, some users first read news from category politics, and then news from category sports.

As a part of future work, by making use of the transition matrix based on impressions, we are going to develop a model which represents "the role of news categories on users' click behavior" in order to mitigate the effects of cold-start problem due to short click histories of new users. This model will be used to suggest a recommendation list based on the transition matrix until a system gets enough past data about new users who have rated a few items yet. The most important issue for future work is going to be the construction and evaluation of a recommender system that uses our findings.

6. ACKNOWLEDGEMENTS

The first author has been funded by the Republic of Turkey Ministry of National Education. The work leading to these results has received funding (or partial funding) from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 610594.

7. REFERENCES

[1] S. Agarwal. Learning to rank on graphs. *Machine Learning*, 81(3):333–357, 2010.

[2] D. Billsus and M. Pazzani. Adaptive news access. In *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 550–570. Springer Berlin Heidelberg, 2007.

[3] T. Bogers and A. van den Bosch. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 141–144, New York, NY, USA, 2007. ACM.

[4] I. Cantador, A. Bellogín, and P. Castells. News@hand: A semantic web approach to recommending news. In *Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, AH '08, pages 279–283, Berlin, Heidelberg, 2008. Springer-Verlag.

[5] I. Cantador, A. Bellogín, and P. Castells. Ontology-based personalised and context-aware recommendations of news items. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '08, pages 562–565, Washington, DC, USA, 2008. IEEE Computer Society.

[6] M. Capelle, F. Hogenboom, A. Hogenboom, and F. Frasincar. Semantic news recommendation using wordnet and bing similarities. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 296–302, New York, NY, USA, 2013. ACM.

[7] Y.-C. Chen, Y.-S. Lin, Y.-C. Shen, and S.-D. Lin. A modified random walk framework for handling negative ratings and generating explanations. *ACM Trans. Intell. Syst. Technol.*, 4(1):12:1–12:21, Feb. 2013.

[8] G. De Francisci Morales, A. Gionis, and C. Lucchese. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 153–162, New York, NY, USA, 2012. ACM.

[9] F. Hopfgartner, B. Kille, A. Lommatzsch, T. Plumbaum, T. Brodt, and T. Heintz. Benchmarking news recommendations in a living lab. In *CLEF'14: Proceedings of the Fifth International Conference of the CLEF Initiative*. Springer Verlag, 09 2014. to appear.

[10] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz. The plista dataset. In *NRS*, pages 16–23. ACM, 2013.

[11] I. Paparrizos, B. B. Cambazoglu, and A. Gionis. Machine learned job recommendation. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 325–328, New York, NY, USA, 2011. ACM.

[12] J.-W. Son, A.-Y. Kim, and S.-B. Park. A location-based news article recommendation with explicit localized semantic analysis. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 293–302, New York, NY, USA, 2013. ACM.

[13] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, F. Hopfgartner, B. Kille, T. Plumbaum, A. Lommatzsch, T. Brodt, A. Bucko, and T. Heintz. Workshop and challenge on news recommender systems. In *RecSys*, RecSys '13, pages 481–482, New York, NY, USA, 2013. ACM.