Clicks Pattern Analysis for Online News Recommendation Systems

Jing Yuan[1], Andreas Lommatzsch[1], Benjamin Kille[1]

[1] DAI-Labor, Technische Universtät Berlin, Germany
{jing.yuan, andreas.lommatzsch, benjamin.kille}@dai-labor.de

Abstract. The NewsREEL challenge provides researchers with an opportunity to evaluate their news recommending algorithms live based on real users' feedback. Since 2014, participants evaluated a variety of approaches on the Open Recommendation Platform (ORP), yet popularity-based algorithms constitute the most successful ones. In this working note, we chronologically describe our participation in NewsREEL online task in the year 2016. With approaches including "most impressed", "newest", "most impressed by category", "content similar" and "most clicked", we reconfirm that content relevance is not a very good indicator for recommending news. Meanwhile, for the dominating portal *Sport1*, the extrapolation of the time series of impressions and clicks enables us to predict the items most likely to be clicked in the next hours. A sample analysis on one week data shows us that the duration of an item being popular is much longer than we expected. Thus, we propose that when designing recommenders in this contest, more attention should be paid on the time series patterns of clicks and impressions.

1 Introduction

News, as important media content, still keeps its role of guiding social opinion, even in modern world which is full of virtual social network and personal ideas. Many news providers employ recommender systems and similar personalization techniques to assist users in finding relevant news quickly and conveniently.

Different ways to incorporate recommendations in news publishers have been successfully launched in the current digital news content market. We exemplify three ways in which recommendations are pushed to news consumers. First, as an *e-Magazine Provider*, Flipboard aggregates news contents from different third party providers and then selects news which is relevant to a user's pre-defined topics forming their personalized news board. Second, some *Content Providers*, such as ByteDance, generate contents themselves. They recommend in a closed system based on internal users, news, and interaction in between both. Third, *Recommendation Providers*, e.g. PLISTA and OUTBRAIN, offer recommendation services for different kinds of websites, including news websites. Table 1 compares characteristics of the three main-stream news recommenders concerning aspects such as whether they generate content by themselves, the stability of users, and the stable range of news items, respectively. As a representative of

Recommendation Providers, PLISTA manifests its non-trivial condition in terms of variety of news portals and differences in users' expectations. Considering that NewsREEL competition receives data stream from PLISTA, participants have to cope with all these knotty conditions to win the contest[8].

Table 1: Characteristics of Main-stream News Recommender

News Recommender	**Generating Content**	**Stability of Users**	**Stable Range of News**
e-Magazine Provider (e.g. Flipboard)	✗	✓	✗
News Content Provider (e.g. Bytedance)	✓	✓	✓
Recommendation Provider (e.g. PLISTA)	✗	✗	✗

The NewsREEL challenge 2016 provides participants with the chance of evaluating recommender algorithms with online live user feedback [4, 3]. In the challenge, teams registered on Open Recommendation Platform (ORP) receive streamed messages describing published news articles, users' impressions and clicks on items, as well as recommendation requests from PLISTA. The challenging aspects of participating NewsREEL include: (1) recommendations must be provided in 100ms upon request; (2) participants need to deal with news portals from different domains; (3) user groups on specific portal alter; (4) number of messages varies largely among portals [8, 5].

In contrast to recommending movies or music, news items continuously emerge and become outdated constituting a dynamic environment. This makes the NewsREEL competition particularly challenging. Algorithms have to consider these dynamics in news articles and users' preferences. We focused on popularity and freshness to cope with the dynamics following the notion that users prefer important and recent news over insignificant and outdated articles. The success of the "most clicked" strategy in terms of CTR further supports this notion. Even though the method is rather simple, it captures crucial aspects. Visualizing clicks on items over time, we observe continued click activity stretching several hours for popular items. We compared contents of popular items and discovered that they overlap. Still, content-based algorithms have failed to benefit of these overlaps in previous editions of NewsREEL.

The remainder of this paper is structured as follows. In Section 2, we briefly introduce the approaches we used in year 2016 and discuss other algorithms developed in previous years. Subsequently, we analyze characteristic user-item interaction patterns for different news portals in Section 3, and found that "most clicked items" has its own power of self-predicting. Finally, conclusion and an outlook to future work are given in Section 4.

2 Approach Used

In this section, we chronologically describe the approaches we have deployed in ORP, i.e. the online task of NewsREEL2016, and changes in our thoughts meanwhile. When the most simple approach "most clicked" finally shows its power to outperform other algorithms, it attracts our interest to dig deeper into clicks pattern from the perspective of time series analysis in the next section.

Most Impressed Inspired by the good performance of "baseline" in the past years (see [9]), which directly uses the most recently impressed items as recommendation candidates, we implemented a similar method by sorting the 2000 most recent impressions by their frequencies. Typically, this approach is called "most popular", but to distinguish it from "most clicked" which will be introduced later on, we refer to it as "most impressed" in this paper. The approach ran on ORP for two weeks (January 31 to February 13, 2016), and got the CTR 1.21% (ranked 3rd, team "artificial intelligence" got the first place with CTR 1.48%) and 1.35% (ranked 2nd, team "abc" got the first place with CTR 1.4%) in these two weeks separately.

Newest Considering that freshness represents a vital aspect of news, we also implemented an approach "newest" which provides the most recently created items from the same category as the currently visited item as recommendation. Given the good performance of "most impressed" mentioned above, we used it as an alternative solution when the request lacked an item_id, i.e. the category cannot be determined. In addition, for a recommendation request with 6 candidate slots, 3 positions are still filled by "most impressed" approach. Therefore, this approach can be seen as a simple ensemble of "most impressed" and "newest". With this solution, from 21–27 February, our team "news_ctr" got CTR 1.19% (ranked 5th, team "artificial intelligence" got the first place with CTR 1.45%) in the contest leader board.

Most Impressed by Category After witnessing how "newest" weakened the effect of "most impressed", we conducted another experiment which only considered the number of impressions, but separates the impression counts according to categories, thus for the recommendation request with item_id the recommending targets will only be the "most impressed" items in the relevant category. The approach ran on ORP for three weeks, from 6–12 March 2016 it got CTR of 0.82% (ranked 7th, team "is@uniol" got the first place with CTR 1.03%), from 13–19 March 2016 it got CTR of 0.97% (ranked 11th, i.e. the last one, team "xyz" got the first place with CTR 1.85%) and from 20–26 March 2016 it got CTR 1.24%(ranked 6th, team "xyz" got first place with CTR 2.16%).

Content Similar Having confirmed that considering categories in combination with popularity lead to worse performances, we implemented a pure content-based recommender using Apache Lucene to see how content relevance influence

recommending effect after all. We deployed this content-based recommender on ORP and noticed that from 27 March to 2 April it got a CTR of 0.77% (ranked 11th, team "xyz" got the first place with CTR 1.51%). This confirmed that in real-time news recommendation scenario as in this contest, pure content similarity is not sufficient for a successful recommending strategy. Said et al. [10] came to the same conclusion hypothesizing that content similarity fails to pick up on new stories but redirects users to similar contents.

Most Clicked While varying on different algorithms, we discovered an interesting phenomenon through the clicks message we received from ORP. Even though different contest teams used different algorithms, the clicked items for all of these recommendations tended to be similar. This consistent regularity reminded us to think whether characteristic patterns exist within clicks along the time axis. Hence we implemented the simplest approach "most clicked" which only serves the most frequently clicked items in the last hour to the recommendation requests. From 3–9 April 2016, this simple approach got a CTR of 1.14% winning the leader board ahead of "xyz" (0.96%). Figure 1 shows the result during this week.

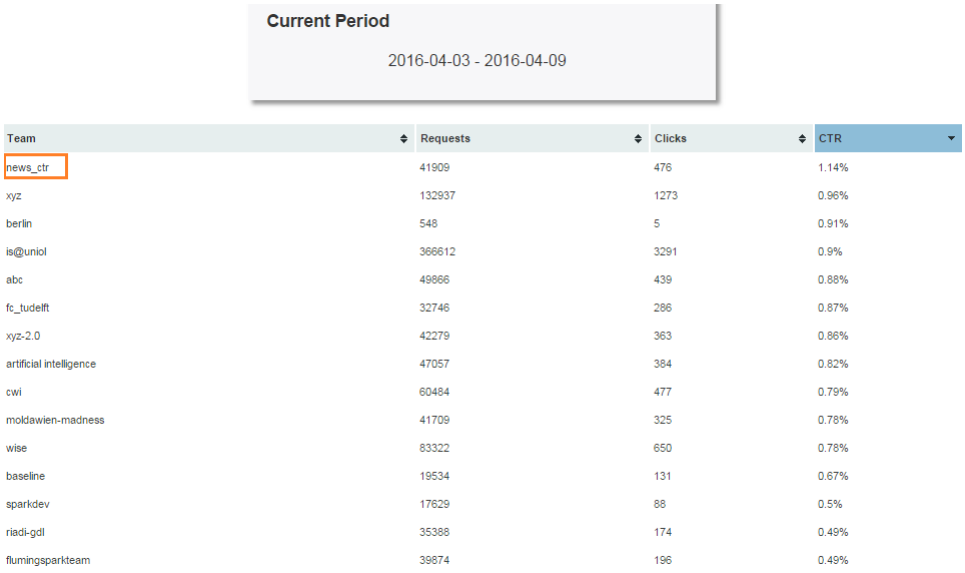

Fig. 1: *news_ctr* got first place in the period 3–9 April, 2016

Having observed this interesting phenomenon, we looked into previous work related to this contest. Kliegr and Kuchař [6, 7] implemented an approach based on association rules. They used contextual features (e.g. `ISP`, `OS`, `GEO`, `WEEKDAY`, `LANG`, `ZIP`, and `CLASS`) to train the rule engine. The results obtained in the online

evaluation indicate that association rules do not outperform other algorithms. Through the investigation in [2], Gebremeskel and de Vries found that there is no striking improvement through including geographic information on news recommendation, yet more randomness of the system should be taken into account when considering evaluation for recommenders. Doychev et al. introduced their 6 popularity-based and 6 similarity-based approaches in [1], but their algorithms seemed to perform poorly compared with "baseline" due to being influenced by content aggregating. As Said et al. concluded in [10] that news article readers might be reluctant to be confronted with similar topic all the way, but more pleased to be distracted by something breaking or interesting. In the following section, we are digging into how this breaking phenomenon is reflected in clicks behavior.

3 Clicks Pattern Analysis

As a further exploration of click patterns, we focus on clicks following recommendation requests in *Sport1* and *Tagesspiegel* on April 5, 2016. Clicks in *Sport1* follow more obvious and stable trends. We analyze this consistency for the "most clicked" recommender's suggestions in terms of the Jaccard similarity of temporally adjacent item groups.

3.1 Clicks Pattern for *Sport1*

First, we draw the histogram of clicks regarding recommendation requests on items in portal *Sport1*. Considering that PLISTA has only delivered part of all recommendation requests to ORP participants, we suppose that the click patterns might slightly differ amid contest teams scope and the whole PLISTA scope. ORP hides such scope information in its click_notification JSON "context.simple" object where key number '41' stands for "contest_team" and value number represents specific team_number. For instance, "news_ctr" is the contest_team with team_number 2465, while team_number -1 signals that the click happened outside contest team range. Thus, the figures are drawn separately by these two scales: contest teams scale excluding clicks outside the contest scope and whole PLISTA range without any restrictions on contest_team.

Figure 2 shows the click conditions among contest teams. In order to track the top clicked items in a time sequence, we draw the figure for each hour for April 5, 2016, i.e. 24 subfigures covering the whole day. In each subfigure, news items are located on the x-axis as points sorted by the click frequency in descending order. A red vertical line separates the six most frequently clicked items—most recommendation requests ask for six suggestions. For a majority of intervals, we observe a power law distribution. Few highly popular items occupy a majority of clicks. The percentage of clicks occupied by the top six items is shown in the red boxes.

We analyze how popularity transitions into the future. Therefore, we highlight the item_ids of the six most popular items in the top right corner of each

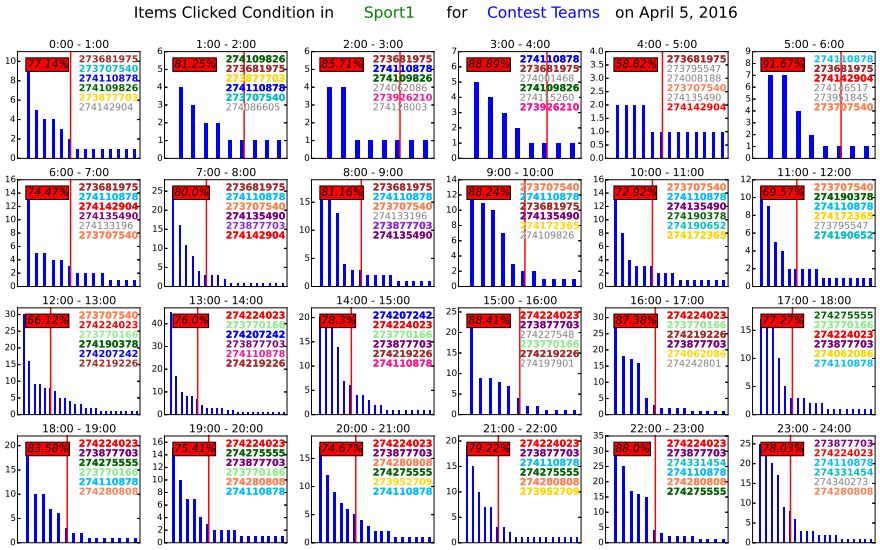

Fig. 2: Top clicks condition of *contest teams* on *Sport1* on April 5, 2016

subfigure. We color item_ids to facilitate tracking individual items throughout the plot. Items tinted in gray only appear in a single one hour interval. From Figure 2, we see that on April 5, 2016, items ranked in the top three manifest more continuity, i.e. they are more likely to re-appear in the next hour's top clicked 6 items group.

Aside from the scope of contest teams in ORP, we are also interested in the power law distribution of clicked items in the whole PLISTA range. In Figure 3, we find that along with the increasing number of distinct clicked items in the "whole" range, the power law distribution of clicked items is even more significant. In all one hour time windows, more than 87% clicks are contributed by the top 6 items. The more complete data may cause the increased steepness of the histograms. The distribution can be described by Zipf's law. The significant advantage of the six most frequently clicked items reminded us that we should pay more attention to the short head with higher business value. Thereby, we can keep a relatively high CTR. Still, more sophisticated methods are required to leverage the potential of the long tail.

When focusing on the most frequently clicked items within this one day, we find some clues for the future work. First, Table 2 illustrates the four items occurring most frequently in the *top 6* group. It lists their item_id, the duration contained in the *top 6*, their ranking trends, and the date they had been created. All four items remained in the *top 6* for at least eleven hours. We had expected considerably less time as news continuously emerge. We notice fluctuating rankings of these four items. Recognizing patterns in shifting rankings will be subject to future work. The dates of creation subvert our previous expectations. We as-

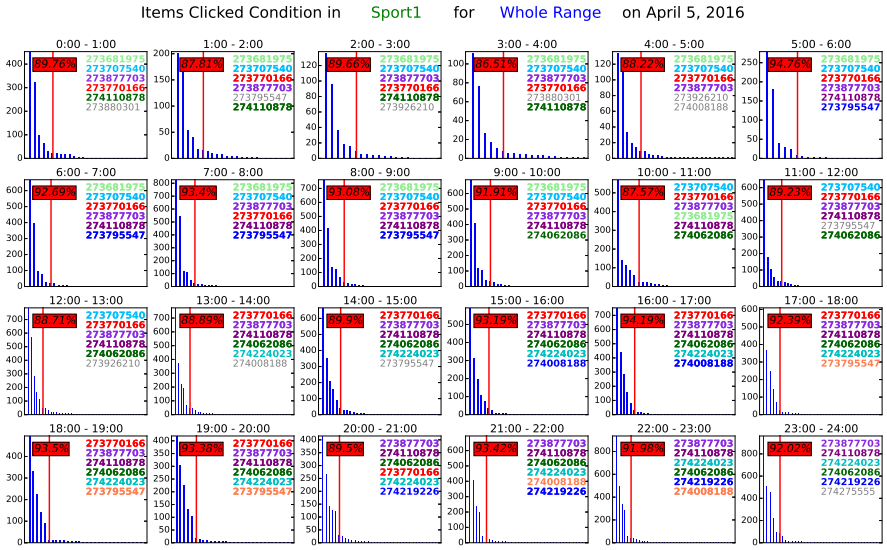

Fig. 3: Top clicks condition of *whole* range on *Sport1* on April 5, 2016

sumed that news would remain relevant for a very limited time. In contrast, news articles created on April 2, 2016, dominated the *top 6* news three days later. This indicates a noticeably longer life-cycle of news than we anticipated.

Table 2: Stable top clicked items condition on April 5, 2016

Item Id	**Being in Top6**	**Ranking**	**Created At**
273681975	0:00–11:00	Most of the time 1st	2nd, Apr. 2016
273707540	0:00–13:00	2nd → 1st	2nd, Apr. 2016
273877703	0:00–24:00	3rd/4th → 2nd → 1st	3rd, Apr. 2016
273770166	0:00–21:00	2nd → 3rd/4th → 1st	2nd, Apr. 2016

Next, we analyze the categories, titles, and descriptions of these four items in order to get a better understanding of the contents. Table 3 highlights co-occurring terms in these four items. Among those, we see that the breaking news that coach *Pep Guadiola* will be leaving *Bayern* attracted many users' interest on relevant articles. We hypothesize that content similarity affects recommenders, but only for popular items. Still, a majority of users only pays attention to the most popular articles which is why pure content-based recommenders frequently suggest articles with minor click chances.

Table 3: Stable top clicked items description on April 5, 2016

Item Id	**273681975**	**273707540**	**273877703**	**273770166**
category	fussball	intenational-fussball	fussball	fussball
title	**Guardiola** macht **Götze** froh	**Van Gaal** watscht di Maria ab	Robben:"**Van Gaal** ist wie **Guardiola**"	Mittelfeldbestie **Götze**: Wechsel nur im Notfall.
text	Der Coach des **FC Bayern** lässt Youngster Felix **Götze** erstmals mit den Profis trainieren. Javi Martinez und Manuel Neuer stehen derweil vor dem Comeback.	Als Rekordtransfer geholt, nach nur einer Saison wieder vom Hof gejagt. Jetzt geht die Geschichte zwischen United-Trainer Louis **van Gaal** und Angel di Maria in die Verlängerung.	Arjen Robben vergleicht den ehemaligen **Bayern**-Coach Louis **van Gaal** mit dem aktuellen Trainer Pep **Guardiola**. Mit dem **FC Bayern** will Robben noch viel erreichen.	Mario **Götzes** beherzter Auftritt gegen Frankfurt zeigt: Er will sich unbedingt beim **FC Bayern** durchsetzen. Der Verein mauert noch beim Thema Transfer.

3.2 Jaccard Similarity Between Clicks in Neighbor Hours

In this subsection, we quantify the continuity of most frequently clicked items and analyze this continuity behavior concerning contextual factors such as time of day and day of week. Jaccard Similarity, as defined in Equation 1, is a metric to measure the similarity of two sets A and B. The value of this metric equates to the cardinality of the intersection divided by the size of union of these two sets. In our scenario, A and B refer to the sets of the six most frequently clicked items of two neighboring one hour time slots. The higher the Jaccard similarity, the more items users constantly are concerned with across neighbor hours.

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

We expand our view from a single day to the week 3–9 April, 2016. Thereby, we obtain $24 \times 7 = 168$ one hour time windows. Thereof, we derive 167 pairs of subsequent time windows to compute the average Jaccard similarity. Figure 4 illustrates our findings overall, for specific times of day, and for each weekday. We distinguish the contest scope and the whole PLISTA scope by cornflowerblue and violet colors. Throughout the three subfigures, we noted that the PLISTA scope's Jaccard metric exceeds the contest scope. The gap is most obvious in the night (0:00–8:00). Still, we have to consider the fact that the night has relatively few

interactions compared with the day time. Independent of context, we observe Jaccard scores in the range of 40–60%. These signal that more than half of the most popular items re-occur in the next hour's *top 6* group. Thus, recommending popular items guarantees a good chance to perform well. This explains the good performance of the "baseline" in previous editions of NewsREEL.

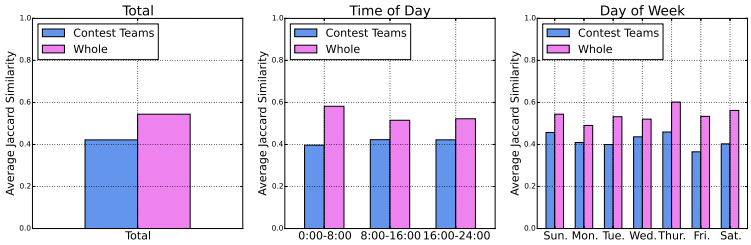

Fig. 4: Jaccard similarity of top clicked items between continuous hours for *Sport1* in the week 3–9 April, 2106

3.3 Predicting Ability of Impressions and Clicks

Empirically speaking, "Most Impressed" approach always performs well on CTR, thus we compare the predicting ability of impressions and clicks regarding clicks in the next hour. As the six most frequently clicked items receive more than 80% of all hourly clicks, we define predicting ability here by the Jaccard similarity between the set of recommended items and the set of the six most frequently clicked items in a specific hour. The six items most frequently viewed in the last hour form the recommending set "Most Impressed". On the other hand, the six items most frequently clicked after having been suggested in the last hour characterize the recommending set "Most Clicked". Figure 5 shows both methods' performances over time. The cyan curve refers to "Most Clicked" while the magenta line refers to "Most Impressed". The upper subfigure shows the comparison of "Most Impressed" and "Most Clicked" in the range "contest teams", and the bottom subfigure presents the same comparison in range "whole PLISTA". "Most Clicked" outperforms "Most Impressed" in both scenarios. This indicates that at least on 5 April 2016, users' reactions to recommendations let the system better predict future clicks than what they read.

3.4 Clicks Pattern for *Tagesspiegel*

Hitherto, we focused on *Sport1*. We repeated our experiments for the second largest publisher—*Tagesspiegel*. Figure 6 shows a considerably lesser number of clicks compared with *Sport1*. Some one hour intervals have less than six clicks in total. Even considering the whole PLISTA range of clicks, Figure 7 shows a

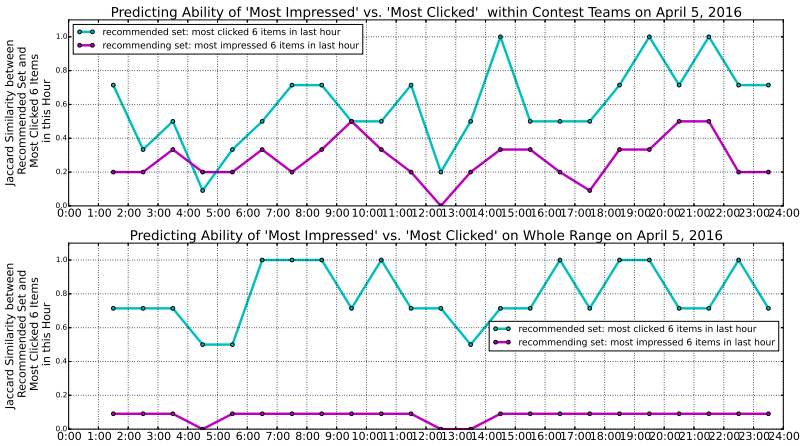

Fig. 5: "Most Impressed" vs. "Most Clicked" on predicting next Hour "Most Clicked" in *Sport1* on April 5, 2016

flatter power law distribution. The top items only account for 20–35% of clicks. In addition, we observe more variation in the most frequently clicked items such that many items appear only for a single hour in the *top 6* group. We hypothesize that the increased variation is caused by the higher diversity in topics. *Sport1* exclusively provides sport-related news. Contrarily, *Tagesspiegel* covers a wide range of topics including politics, economy, sports, and local news.

4 Conclusion and Future Work

In this working note, we describe our experience with the real-time news recommendation contest NewsREEL online task in 2016. Through evaluating approaches such as "Most Impressed", "Newest", "Most Impressed by Category", "Content Similar", and "Most Clicked", we found out that a small subset of news items attracted most clicks. This holds true beyond the scope of individual algorithms. Hence we started analyzing the patterns of clicked items on the dominating portals *Sport1* and *Tagesspiegel*. In particular for *Sport1*, item popularity followed a power law distribution and items continued to be popular for hours. This phenomenon was less pronounced on *Tagesspiegel*. Monitoring which articles users clicked provided better information to predict future clicks than tracking which articles users read. These observations inspire us to change the perspective of implementing recommender from analyzing features and contextual factors to investigating clicked items' time series patterns. Thus, as long as *Sport1* continues to be the dominant news source in the contest, we can focus on the following points as future work: (1) analyzing the duration regularity of an item staying in the most clicked items group; (2) the ranking prediction of

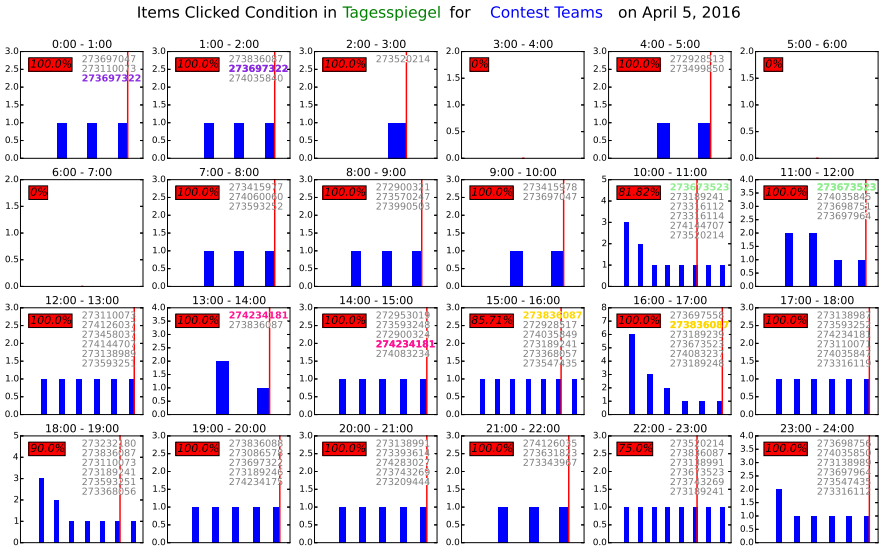

Fig. 6: Top clicks condition of *contest teams* range on *Tagesspiegel* on April 5, 2016

an item of being popular; (3) making use of long tail to find relevant features and contexts.

5 Acknowledgement

The work of the first author has been continuously funded by China Scholarship Council (CSC). The research leading to these results is partially supported by the CrowdRec project, which has received funding from the European Union Seventh Framework Program FP7/2007–2013 under grant agreement No. 610594.

References

1. D. Doychev, R. Rafter, A. Lawlor, and B. Smyth. News recommenders: Real-time, real-life experiences. In *Proceedings of UMAP 2015*, pages 337–342, 2015.
2. G. Gebremeskel and A. P. de Vries. The degree of randomness in a live recommender systems evaluation. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*. CEUR, 2015.
3. F. Hopfgartner, T. Brodt, J. Seiler, B. Kille, A. Lommatzsch, M. Larson, R. Turrin, and A. Serény. Benchmarking news recommendations: The clef newsreel use case. *SIGIR Forum*, 49(2):129–136, Jan. 2016.
4. B. Kille, A. Lommatzsch, G. Gebremeskel, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Serény, T. Brodt, and A. de Vries. Overview of newsreel'16: Multi-dimensional evaluation of real-time stream-recommendation algorithms. In

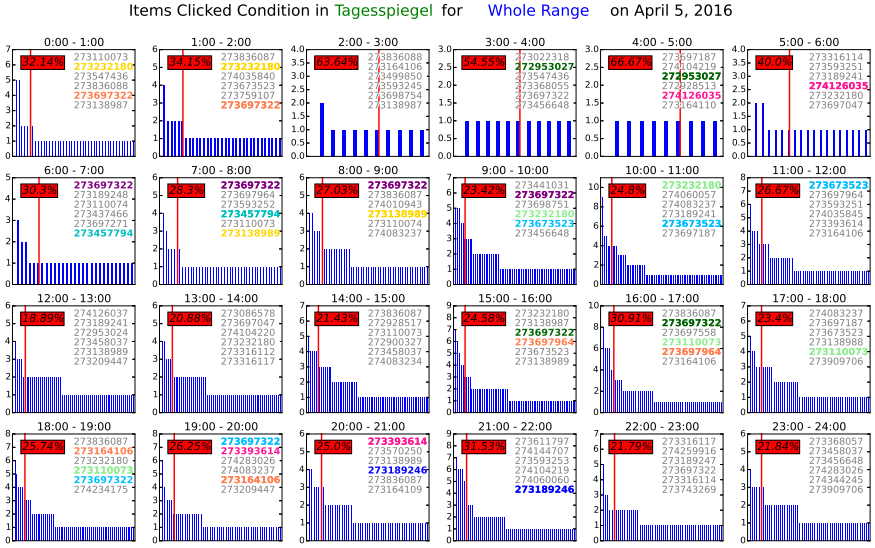

Fig. 7: Top clicks condition of *whole* range on *Tagesspiegel* on April 5, 2016

N. Fuhr, P. Quaresma, B. Larsen, T. Goncalves, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016.* Springer, 2016.

5. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Stream-based recommendations: Online and offline evaluation as a service. In *Proceedings of the Sixth International Conference of the CLEF Association, CLEF'15*, pages 497–517, 2015.
6. T. Kliegr and J. Kuchar. Benchmark of rule-based classifiers in the news recommendation task. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, and N. Ferro, editors, *CLEF*, volume 9283 of *Lecture Notes in Computer Science*, pages 130–141. Springer, 2015.
7. J. Kuchar and T. Kliegr. InBeat: Recommender System as a Service. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, pages 837–844, 2014.
8. A. Lommatzsch. Real-time news recommendation using context-aware ensembles. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 51–62, 2014.
9. A. Lommatzsch. Real-time recommendations for user-item streams. In *Proc. of the 30th Symposium On Applied Computing, SAC 2015*, SAC '15, pages 1039–1046, New York, NY, USA, 2015. ACM.
10. A. Said, A. Bellogín, J. Lin, and A. P. de Vries. Do recommendations matter?: news recommendation in real life. In *Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, February 15-19, 2014, Companion Volume*, pages 237–240, 2014.