Recommending News Articles using Rule-based Classifier

Christián Golian and Jaroslav Kuchař

Web Intelligence Research Group, Faculty of Information Technology,
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic

{goliachr, jaroslav.kuchar}@fit.cvut.cz

Abstract. In this paper we summarize our experiments with a rule-based classifier as a recommender within CLEF NewsREEL 2017 challenge. Systems that recommend news articles are suitable to solve information overflow in digital editions of newspapers, when users have problems choosing what they want to read. They face challenges unknown to the systems recommending books or movies such as a frequency of producing the new content. This paper deals with an approach based on association rules acting as a classifier. In our approach we experimented with settings that allow reducing the amount of rules used for the classification and increasing the performance that is crucial for real recommendations.

Keywords: news recommender, association rules, CLEF NewsREEL.

1 Introduction

The enormous number of available news causes information overflow resulting in users having problems choosing what they want to read. News recommendation systems should solve this problem and offer them an article or a collection of articles, which they could find worth a read.

In this paper we summarize our experiments with a rule-based classifier as a recommender within CLEF NewsREEL 2017 challenge[1] [1]. The challenge enables to compare and evaluate news recommendation systems both offline and online. This challenge allows to participate in two tasks: 1) NewsREEL Live [4] which uses real-time information about interactions between users and items. It is realized by redirecting a part of internet traffic to a recommender system of a participant in the challenge and 2) NewsREEL Replay [3] which uses historical data. The provided benchmarking system simulates the real stream of events by replaying of recorded historical data. Those tasks allow benchmarking of algorithms in terms of quality of recommendations (measured by the Click-Through Rate) and technical aspects (measured by reliability and response time)[5].

[1] www.clef-newsreel.org

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 51-55.

Over the years, many different approaches to news recommendation have been developed. Given that the rule-based approach yielded promising results [6] we decided to do further work in this area. The advantage of the rule-based approach is the possibility to easily explain the recommendations, since rules are considered as on of the most understandable representation of models. In our algorithm we focused on the rule-based classifier *CBA* [6] that is also focused on reducing amount of rules using a pruning of available rules. The proposed solution applies several existing algorithms and approaches that together build a competitive recommender system.

2 Approach

In our approach we decided to mainly use the contextual information about the reader or the article being read. During offline evaluation, we made several experiments to explore, which available features would perform best. Based on these experiments, we selected settings and a subset of twelve attributes for the online recommendation task.

The reasoning behind the use of rules was following: If a certain number of user interactions with an item often include values of attributes repeating themselves, they may be interesting either for users sharing these attributes or for users reading articles sharing these attributes. In order to provide an example: if an article was read frequently during evening hours, it may be interesting to someone else reading news in the evening or late at night.

To get list of relevant rules we use standard existing algorithms (e.g. *Apriori*, *FPGrowth*). Each rule is composed from a left-hand side, right-hand side and it is described by its quality measures: support and confidence. To prefer certain rules to other, we sort rules in the same way as in *CBA* according to the confidence, support and length of the rule. Since the amount of rules returned by the standard implementation can be huge, we use the rule pruning: it removes rules that can be never used for subsequent classification, usually due to their redundancy, lower significance etc. Our algorithm works as follows: an article (article id on the right hand side of the rule) is recommended only when values of attributes contained on the left hand side (contextual information) of a rule are equal to values of attributes in the recommendation request. If there are more matching rules, we use all unique article identifiers as a list of recommended articles. If no recommendation was made using a matching rule, implementation of baseline algorithm provided by organizers of CLEF News-REEL was used. Examples of rules from the classifier (values of features are anonymized):

```
{browser:40052, geo_user_zip:61958} => {itemId:341743113}
    supp = 0.02 , conf = 1.0
{isp:6, category:420949} => {itemId:367259468}
    supp = 0.01 , conf = 1.0
{browser:16064801, device:504182} => {itemId:315791779}
    supp = 0.1 , conf = 0.7
{} => {itemId:322334534}
    supp = 0.01 , conf = 0.01
```

In every domain there was a rule with an empty left hand side called the default rule. This means that every recommendation request from this domain matched it, and so at least one recommendation was always made using association rules.

The technical solution is built on top of the provided *Java SDK*[2]. Our implementation uses only the latest *n* interactions for rule mining. It enables to reflect outdating of news items and dealing with performance issues as well. Main implementations of rule mining algorithms and corresponding operations are available for the programming language R; we thus use a binary server *Rserve*[3] to provide communication between Java and R. To create association rules from interactions between users and news articles *arules* [2] library for R was used. The *rCBA* [7], a classification based on association classifier for R was used to prune and remove redundant association rules.

3 Evaluation

For the offline evaluation we selected subsets from the very large dataset provided by organizers. We compared results of our approach with the provided baseline implementation that recommends the most recent articles. The comparison of both approaches for selected news portals is on the Table 1. The rule-based approach can provide slightly better results. It takes into account the contextual features and temporal aspects at the same time.

Tab. 1. Offline evaluations – Click-Through Rate

News Portal (id)	Baseline algorithm	Rule-based algorithm
418	0.07%	0.14%
1677	0.40%	0.45%
35774	1.30%	1.62%
All	1.04%	1.29%

Table 2 presents the experiments with setting of parameters that influence the building of the rule-based classifier. It allows getting brief insights to the influence of specific settings. Those experiments helped us to find appropriate settings for further experiments. Please note, that only subset of experiments is presented.

[2] https://github.com/plista/orp-sdk-java
[3] https://rforge.net/Rserve/

Tab. 2. Offline evaluations – parameters setting

Parameter/Portal	418	1677	35774	All
conf: 1%, sup: 0.5%	0.14%	0.45%	1.62%	1.29%
conf: 1%, sup: 1%	0.16%	0.43%	1.59%	1.27%
conf: 5%, sup: 2%	0.13%	0.46%	1.67%	1.34%
max. clicks 5000	0.14%	0.41%	1.65%	1.31%
max. clicks 100000	0.14%	0.40%	1.65%	1.31%
max. rule length 2	0.16%	0.42%	1.66%	1.32%
max. rule length 10	0.13%	0.42%	1.60%	1.30%
max. rule length 12	0.13%	0.43%	1.62%	1.29%
pruning disabled	0.18%	0.53%	1.75%	1.40%

Several conclusions were drawn from these experiments. The prediction accuracy increased together with increasing of maximum length of association rule only up to a certain length. Increasing the maximum number of clicks (size of data that is used to mine rules) did not significantly increase the prediction accuracy. Disabling pruning of rules can bring better results, but at a cost of higher number of rules leading to higher number of erroneous responses and longer computing time. It can lead to the decreased reliability and increased response times.

Tab. 3. Online evaluations – Selected teams from the official results [5]

Team	Clicks	Impressions	CTR
BL2Beat	726	193014	0.376%
B	879	244334	0.360%
WIRG	600	154419	0.389%
I	764	236332	0.323%
N	1268	255663	0.496%
O	896	130221	0.688%
Q	12	1380	0.870%

In the online evaluation, our algorithm took 13th place of 21 contestants based on the CTR that relates to impressions - how often the recommendations of a system have been shown to readers (Algorithm called WIRG in the Table 3). The table summarizes results from the final evaluation period. The setting of our algorithm we selected according to the best results from the offline evaluation and two online testing periods allowing tuning of participated algorithms. We were able to beat the baseline (BL2Beat) with our approach. Since we do not know any details about other participating algorithms, it is not possible to state any conclusions yet. The important fact is that our algorithm was able to handle incoming messages, process data and provide recommendations (up to 100 messages per second with responses within 100ms). The recommendations are at the same time easily explainable.

4 Conclusion

In this paper our news recommender system based on association rules was examined. The main idea of the algorithm is to build a rule based classifier using contextual features and study influence of several settings. The advantage is that the recommendations are explainable and the recommender is technically designed as a scalable solution allowing real-time recommendations. Our algorithm took part in both tasks of the CLEF NewsREEL 2017 challenge. The algorithm managed to beat baseline. In our future work we would like to address the detected issues and limitations of our solution. One aspect that we would like to try to overcome is the need for repetitive computation of models on the background using stream-based version of the rule-mining algorithm.

References

1. Golian Ch., Kuchar J.: News Recommender System based on Association Rules @ CLEF NewsREEL 2017. In: Working Notes of CLEF 2017 – Conference and Labs of the Evaluation forum, Dublin, Ireland, 11-14, 2017.
2. Hahsler M., Gruen B.,Hornik K.: arules – A computational environment for mining association rules and frequent item sets. Journal of Statistical Software, 14(15):1–25, 2005.
3. Hopfgartner F., Brodt T., Seiler S., Kille B., Lommatzsch A., Larson M., Turrin R., Sereny A.: Benchmarking news recommendations: The CLEF newsreel use case. SIGIR Forum, 49(2):129–136, 2015.
4. Hopfgartner F., Kille B., Lommatzsch A., Plumbaum T., Brodt T., Heintz T.: Benchmarking news recommendations in a living lab. In: CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative, LNCS, pages 250–267.Springer Verlag, 09 2014.
5. Kille, B., Lommatzsch, A., Hopfgartner, F., Larson, M. and Brodt, T.: CLEF 2017 NewsREEL Overview: Offline and Online Evaluation of Stream-based News Recommender Systems. In: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14, 2017.
6. Kliegr T., Kuchar J.: Benchmark of rule-based classifiers in the news recommendation task. In: Proceedings of the Sixth International Conference of the CLEF Association, CLEF'15, pages 130–141, 2015.
7. Vojir S., Zeman V., Kuchar J., Kliegr T.: Easyminer/r preview: Towards a web interface for association rule learning and classification in r. In: Proceedings of the RuleML 2015 Challenge, Berlin, Germany, August 2-5, 2015., 2015.

Acknowledgements: This research was supported by Faculty of Informatics, Czech Technical University in Prague.