Consolidating Metrics and Developing Best Practices: Bridging the Gap in Recommender Systems Evaluation

Position Paper

Patrick Turgeon
Principal, Figurs*
Toronto, Ontario, Canada
patrick@figurs.ca

ABSTRACT

In industry, analytics and business challenges arise when attempting to upgrade a recommender system (RS). The drivers for this divergence which occurs between the analytics and the business groups are explored. Moreover, some of the metrics and evaluation criterion used by each groups are reviewed. It is proposed that the current number of metrics used by analysts to evaluate recommendation systems needs to be consolidated and inclusive of business metrics. An interactive dashboard or scorecard as an evaluation tool is proposed. It's also suggested that ACM RecSys should consider developing a best practices guide for evaluating RS and promote it widely to industry.

Categories and Subject Descriptors

General and Reference: evaluation, metrics, and measurement. **Information systems**: recommendation systems, personalization, collaborative filtering, content filtering.

General Terms

Management, Measurement, Documentation, Performance, Design, Human Factors, Standardization, and Verification.

Keywords

Recommender systems, best practices, metrics.

1. INTRODUCTION

Recommender system (RS) are often born out of a behavioral segmentation which show that a given website is losing a set (segment) of users who are not able to find what they came for. In web analytics, this is reflected by looking at the search versus product view ratio over a period of time.

Otherwise, the conception of a RS is initiated at the launch of a site, with the prime objective of increasing "stickiness"; ensuring full user experience and to increase the monetization of content.

Regardless of the driver, at the beginning of the lifecycle of a recommender system, both analysts and business sponsors' objectives are aligned since most of the evaluation criteria centers around product development and delivering something new for the organization and its users. Hence, the goals are narrowly defined. The assumption tends to be linear in nature across the enterprise: better user experience will lead to more sales.

As the lifecycle of the RS matures, the overall goal remains the same, but the process and criteria for the RS evaluation are often not well articulated and defined. Hence, a divergence between the analytics and business groups occurs.

2. DIVERGENCE

The issue is that the term "performance" takes on different meanings and definition for both groups. Anderson (2006) illustrated the reasons for potential divergence for both group by showing that as we move down the long tail, the range of quality and satisfaction of the recommendation increases, along with the necessary amount of filtering power required [1].

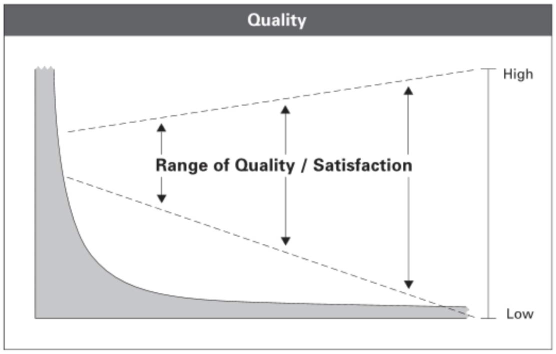


Figure 1. Range of Quality/Satisfaction [1].

Moreover in industry, impressions for recommendations can be impacted by factors such as creative and/or inventory space availability, as well as, business rules (driven by detailed use cases). These factors layered over the algorithm(s) of a RS can further complicate the issue of performance evaluation.

Consequently, a RS may present a given product more often (by default or on purpose – ie. recommendation persistence [2]) and not meet the business criteria for better user experience, namely diversity [3] and coverage [4].

In addition, it's not unusual for the business to point out instances, where the RS may not "perform well" by considering different evaluation metrics (clicks), and ranking criteria (i.e. margins, sales, or conversion) amongst other factors that may not have been scoped by analytics.

Hence, what tends to happen over time is that the initial linear justification for the RS becomes much more sophisticated and

Copyright is held by the author. Workshop on 'Recommender Systems Evaluation: Dimensions and Design' (REDD 2014), held in conjunction with RecSys 2014. October 10, 2014, Silicon Valley, United States.

complex; business becomes concerned with user experience and other metrics, while analytics stays focused on initial scoped definition of "performance" within the context of scalability.

Therefore it's not uncommon that at the time of a release upgrade, the evaluation criteria are not well articulated in the scope of the project. These often need to be "re-casted" as requirement or side project which require further simulation. This in turn, tends to delay deployment and potentially affect the company's bottom line by postponing the release of RS to production

3. EVALUATION CRITERIA

"The literature on recommender system evaluation offers a large variety of evaluation metrics but provides little guidance on how to choose among them." [5]

From an analytics perspective in both industry and research, the focus has been to evaluate RS by using an abundance of metrics. Although there has been some outstanding research on the subject, many of metrics identified are not fit for business consumption. For instance, some are not easily communicated, others are not relevant to user experience and/or the bottom line.

RS Evaluation Criteria from Business Perspective	
Not easily Understood	*Understood*
F1	Coverage
Precision	Rank
Recall	Average Precision
DCG, iDCG, nDCG	

Figure 2. RS Evaluation Criteria from Business Perspective.

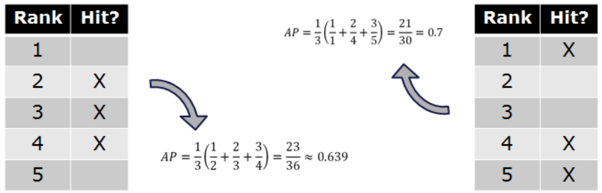
Case in point, f1; precision and recall are identified as key metrics in the analytics community but are often not well understood by business [4, 5]. Precision and recall are both dependent on the number of recommendation or threshold selected. As a result, they tend to be inversely proportional, rise in one leads to a decrease in the other and is not necessarily connected to sales.

Similarly discounted cumulative gain (DCG), idealized (iDCG), and normalized (nDCG) also suffer from the same fate [4, 5].

On the other hand, coverage, rank and average precision are essential for getting an overview of performance and refining business rules for use cases [4, 5]. These are easily understood by business and impact performance. Below are some definitions:

Item Space Coverage
Most commonly, the term coverage refers to the proportion of items that the recommendation system can recommend. This is often referred to as catalog coverage.

Average Precision (AP)
Is a ranked precision metric that places emphasis on highly ranked correct predictions (hits)



Rank	Hit?
1	
2	X
3	X
4	X
5	

$$AP = \frac{1}{3}\left(\frac{1}{2} + \frac{2}{3} + \frac{3}{4}\right) = \frac{23}{36} \approx 0.639$$

Rank	Hit?
1	X
2	
3	
4	X
5	X

$$AP = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5}\right) = \frac{21}{30} = 0.7$$

Figure 3. Item Space Coverage and Average Precision (AP).

4. RECOMMENDATION

In closing, it is proposed that as the RS matures, best practices for evaluation criteria should be developed to reflect a balance between both the objectives of business and the analytics group.

The issue is that term "performance" with RS evaluation is relative to the number of recommendations or threshold selected. And ultimately, the analytics group exists to serve the business.

As a result, the evaluation of RS should contain selective metrics (i.e. coverage, rank, average precision). Furthermore, these should be inclusive of units that are relevant to the business (i.e. clicks, item rank margins, and conversion).

By consolidating evaluation metrics, the gap between business and analytics perspectives can be bridged. One way to address the challenge would be to move to an interactive dashboard or scorecard as a tool for evaluation. This way, multiple metrics can be visualized at once by both groups, and back-end micro-simulations can inform how change in recommendation threshold affects overall business metrics.

It is recognized that this solution is not simplistic and brings considerable additional challenges to a project (i.e. data collection, aggregation, and processing time) but the merits should be debated in industry and research.

ACM RecSys can also facilitate the process by developing a "best practice guide" to evaluate RS which can be used in industry. This will solidify the reputation of the organization in the marketplace, but also demystify and standardize the RS evaluation process.

Interestingly, Konstan and Adomavicious (2013) have identified a similar issue with algorithmic research [6].

REFERENCES

[1] C. Anderson 2006 The Long Tail: Why the future of business is selling more of less. Hyperion Books, NY, NY, USA.

[2] J. Beel, S. Langer, M. Genzmehr, A. Nürnberger (2013). Persistence in .Recommender Systems: Giving the Same Recommendations to the Same Users Multiple Times. In T. Aalberg and M. Dobreva and C. Papatheodorou and G. Tsakonas and C. Farrugia. Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013). Lecture Notes of Computer Science (LNCS) 8092. Springer. pp. 390–394.

[3] C. N. Ziegler, S. M. McNee, J.A. Konstan, & G. Lausen (2005). Improving recommendation lists through topic diversification. Proceedings of the 14th international conference on World Wide Web. pp. 22–32.

[4] G. Shani, A. Gunawardana 2011 Evaluating Recommender Systems. (pp 257-297). In – F. Ricci, L. Rokachi, B. Shapira, P. Kanto 2011 Recommender Systems Handbook, Springer, NY, NY, USA. pp 257-297

[5] Schröder, G., Thiele, M., & Lehner, W. Setting Goals and Choosing Metrics for Recommender System Evaluations.

[6] J. A. Konstan, G Adomavicius 2013 Toward the adoption of Best Practices in Algorithmic Recommender Systems Research. Proceedings of Rec Sys 2013, Hong Kong, China.