Recommendations with a Purpose

Dietmar Jannach
TU Dortmund, Germany
dietmar.jannach@tu-dortmund.de

Gediminas Adomavicius
University of Minnesota, USA
gedas@umn.edu

ABSTRACT

The purpose of recommenders is often summarized as "help the users find relevant items", and the predominant operationalization of this goal has been to focus on the ability to numerically estimate the users' preferences for unseen items or to provide users with item lists ranked in accordance to the estimated preferences. This dominant, albeit narrow, view of the recommendation problem has been tremendously helpful in advancing research in different ways, e.g., through the establishment of standardized evaluation procedures and metrics. In reality, recommender systems can serve a variety of purposes from the point of view of both consumers and providers. Most of the purposes, however, are significantly underexplored, even though many of them are arguably more aligned with the real-world expectations for recommenders than our current predominant paradigm. Therefore, it is important to revisit our conceptualizations of the potential goals of recommenders and their operationalization as research problems. In this paper, we discuss a framework of recommendation goals and purposes and highlight possible future directions and challenges related to the operationalization of such alternative problem formulations.

Keywords

Foundations of Recommender Systems; Recommendation Goals and Purposes

1. INTRODUCTION

Automated recommendations have become a pervasive part of the daily user experience on the web. Today, many major e-commerce websites, media streaming platforms, and social networks use a part of the user interface to display recommendations to their users [3]. The overarching goal for such recommendation services is to create some sort of *utility* (or benefit), e.g., provide users with relevant information, improve customer retention, increase revenue. A number of studies show that recommenders can measurably influence the behavior of online consumers [2, 5].

RecSys '16, September 15 - 19, 2016, Boston , MA, USA

ISBN 978-1-4503-4035-9/16/09…$15.00

DOI: http://dx.doi.org/10.1145/2959100.2959186

Academic research on recommender systems has matured substantially over the past two decades. Because the interpretation of "utility" highly depends on the specific application context, the research community has developed general domain-independent frameworks to operationalize this concept. In particular, the canonical tasks of a recommendation system are often described as *find good items* or *predict an item's relevance to a user* [1, 4]. Operationally, assessing the system's performance on such tasks often translates to predicting held-out user preference ratings, i.e., to a rating matrix completion problem, or to the creation of item lists that are ranked according to the estimated user preferences.

This abstract, canonical view of the recommendation problem has been tremendously helpful to the recommender systems community in establishing a common terminology for the field, providing a clear set of problems to focus on, facilitating the creation of a number of general-purpose recommendation algorithms (i.e., abstracting away from domain information), and supporting a certain level of reproducibility and continuous improvement through standardized metrics and experimentation procedures.

Despite the advantages of having an established, canonical problem formulation, it is important to note that there exist numerous other potential interpretations of the purpose of a recommender system than "help users find relevant items", from the point of view of both consumers and providers. Consider, for example, that a provider aims to use a recommender to guide online customers to other parts of the product catalog, e.g., to increase sales of long-tail or high-margin items. The predicted relevance of these items might still be important for the recommended items to be selected by the user, but has to be balanced with the estimated (e.g., economic) utility for the provider.

Such types of recommendation purposes are largely underexplored in the literature, and we argue that it is important to revisit the research community's conceptualizations of recommender systems goals. In this paper, we discuss a purpose-driven approach to recommender systems, provide illustrations of the possible recommendation purposes from the consumer and provider viewpoints, and outline challenges with respect to the operationalization of such alternative problem formulations in academic research.

2. FROM GOALS TO METRICS

2.1 Traditional Task-Oriented Perspective

In Herlocker et al.'s seminal paper on the evaluation of recommender algorithms [4], the possible system purposes (in

the context of which an evaluation is made) are indirectly described with the help of abstract *tasks* of a recommender. The tasks include "Annotation in Context", "Find Good Items", "'Find All Good Items", "Recommend Sequence", and "Just Browsing". *Annotation in Context* corresponds to the rating prediction problem, and *Find Good Items* to item ranking. *Find All Good Items* is a special case where the goal is not to miss any relevant item, and *Recommend Sequence* is a task that is, for example, common for next-track music recommendation. *Just Browsing* finally refers to situations where there is no imminent "buying decision" on the user's side and other factors than accuracy, e.g., user interface aspects or content-richness, are important.

To evaluate the performance for a number of aforementioned tasks, mainly a variety of standard accuracy measures are used. In recent years a number of additional ways to quantify recommendation quality aspects beyond accuracy have also been more widely used, including different forms of assessing the diversity, novelty, or serendipity of the recommended items. The *Recommend Sequence* task typically requires specific algorithms and datasets, e.g., for next-basket recommendation, but can also be evaluated with accuracy measures like recall, while measuring the success of a system in support of the *Just Browsing* task is underexplored [4].

2.2 Toward Purposeful Evaluation

A question that is rarely asked explicitly in recommender systems research, is: *What is a good recommender system?* (Or: *What is a good recommendation?*) According to today's established task-oriented framework for evaluating algorithms, one could argue that the answer to this question is straightforward: a good system is one that provides a low RMSE or whatever other *computational metric* (like MAE, F-measure, NDCG, MAP, hit rate) researchers decide to apply for their problem. The choice of the metric(s) ideally should be aligned with an operational *system task* that the proposed algorithm is being designed for.

For example, when addressing the rating prediction problem, metrics like RMSE or MAE are used, and when addressing the item ranking problem, NDCG or MAP are employed. A general observation is that the vast majority of recommender systems research takes an *operational perspective* on recommendation by focusing on one of the standard, well-defined operational system tasks.

As mentioned in Section 1, having standard operational tasks (such as rating prediction using user-item matrix completion approaches) provides a lot of benefits to the research community. At the same time, it is important to note that operational system tasks often represent a significant oversimplification of real-world situations. For example, the operational task "find good items" on a movie streaming website may correspond to very different *recommendation purposes.*[1] Furthermore, it is likely that these purposes may differ dramatically depending on whether the recommender system's *overarching goal* is considered from the consumer's or the provider's point of view. In particular, the "find good items" task on a movie streaming website can be representative of having the "entertainment" (or "satisfying emotional experience") purpose from the consumer's viewpoint, while the same "find good items" task might represent the "increase user engagement" purpose from the provider's viewpoint.

[1]See Table 1 for a number of representative recommendation purpose examples from the consumer's and provider's viewpoints.

Table 1: Examples of Recommendation Purposes

Consumer's Viewpoint (i.e., value for the consumer)
Help users find objects that match their long-term preferences: Corresponds to the usual assumption that the main value of a recommender is to help users find relevant items in larger item sets. Recommendations could be limited to a subset of the items, e.g., new or trending ones.
Actively notify consumers of relevant content: Proactively point users to new items through push notifications or newsletters, minimizing the user effort to check the site.
Show alternatives: Recommend substitute products in the context of a reference item. A standard mechanism on e-commerce platforms.
Show accessories: Recommend complementary products in the context of a reference item, e.g., with the goal of cross-selling. Also common in e-commerce settings.
Help users explore or understand the item space: Help the user understand the space of options, possibly leading to higher choice confidence.
Remind users of already known items: Provide users with reminders of repeated purchases of consumables. Or, present a list of recently viewed items that the user did not purchase so far; reminders then also serve as navigation shortcuts to a reduced choice set.
Improve decision making, e.g., in terms reduced decision time or higher choice satisfaction: E.g., provide the user with a limited choice set for an estimated purchase intent, provide explanations and interactive control.
Establish group consensus: Provide recommendations that balance the interests of different group members.
Help user explore: Provide a convenient way for users to browse the catalog without immediate shopping intent.
Entertainment: Provide a satisfying emotional experience when visiting the site.
Provider's Viewpoint (i.e., value for the provider)
Change user behavior in desired directions: Guide customers to other product categories, drive the demand from top-selling items to the long tail, leverage the persuasive potential of recommenders for up-selling purposes.
Create additional demand: Point users to other relevant items (e.g., accessories) to achieve cross-selling and advertisement effects.
Increase (short term) business success: Promote items with high margins or items that are in stock.
Enable item "discoverability": Increase the visibility of new items or niche products that would otherwise not be easily found through search or catalog browsing.
Increase activity on the site: Make users stay longer on the site, e.g., thereby increasing ad revenue.
Increase user engagement: Increase customer loyalty and trust via a personalized service. Increase switching costs to other services as preferences are already known.
Provide a valuable add-on service: Use the recommendation service as a differentiating factor from other competitors.
Learn more about the customers: Utilize the collected preference information to better understand preferences and trends of the consumers. Provide mechanisms for users to explicitly state their preferences.
Generate impression of dynamic, constantly updated site: Dynamic recommendations contribute to the liveliness of the site. Updating the site with editorial content can be costly and less attractive for users.

Table 2: Proposed Framework: From Goals to Metrics

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	Overarching Goal	"Personal Utility": Happiness, Satisfaction, Knowledge, ...	"Organizational Utility": Profit, Revenue, Growth, ...
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space • ...	• Change user behavior in desired directions • Create additional demand • Increase activity on the site • ...
Operational Perspective	System Task	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • Find suitable accessories • Retrieve novel but relevant items • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., precision, recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item "discoverability" (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), ...	

This has a couple of important implications. First, the consumer's and provider's purposes are not always aligned. In this example, more movie watching (good for provider) may not necessarily mean more satisfaction (good for consumer). Therefore, the two viewpoints may often need to be modeled and evaluated separately. And second, the traditional computational metrics that are associated with standard operational tasks – e.g., the NDCG ranking metric for the "find good items" task – may be reflective of neither the consumer's nor the provider's viewpoint. For example, better preference-based ranking, as measured by NDCG, does not automatically guarantee more consumption/engagement nor more satisfaction. Thus, it is important for the recommender systems community to (a) move beyond the standard operational perspectives and explicitly take real-world *strategic perspectives* (purposes and goals) into account, and also to (b) design a more comprehensive set of operational system tasks and corresponding computational metrics for different recommendation purposes and viewpoints.

2.3 Proposed Framework

To summarize, we propose the following framework for understanding and evaluating recommender systems utility as shown in Table 2:

- *Overarching goals*: Goals describe the general underlying personal or organizational motivation of using or providing a recommender system.
- *Recommendation purposes*: Purposes capture the specific utility of the service for a consumer or provider, i.e., the expected measurable effects or value of deploying the system. Examples are given in Table 1.
- *System tasks*: These represent operational, algorithmic tasks to be accomplished by the recommendation system.
- *Computational metrics*: Metrics provide numeric quantifications of the extent to which the recommender is able to accomplish a given system task.

As an illustrative example, consider an online music service provider, whose overarching goal is to increase revenue via improving customer long-term loyalty. One of the possible purposes of the recommendation system could, therefore, be to maximize the number of monthly subscription renewals by keeping the users active and involved in the service. This, for example, could be achieved through recommendations that help the user discover new artists. An operationalized system task could then be to find artists that the user most probably does not know yet, e.g., because they are newcomers but which match his or her musical tastes. Corresponding computational metrics could therefore consider aspects of accuracy (preference match), serendipity (surprise and entertainment), item novelty, as well as the monthly subscription numbers (e.g., from an A/B test).

While the discovery of new artists might also be in the interest of consumers, the consumer's and provider's goals are not necessarily aligned. Assume that an e-commerce platform aims to increase the profit on the site as an overall goal. The purpose of the recommender could then be to lure the consumers to high-margin catalog items, and the algorithmic task is to find such high-margin items that are still a reasonable match for the consumer's estimated preferences. The consumers, possibly under the impression that the recommended items are the most relevant (or the most popular) and, therefore, "safe" choices, might as a result continuously make sub-optimal buying decisions, which can lead to consumer dissatisfaction in the long run. Therefore, from the perspective of the computational metrics, the best solution may require a delicate balance of optimizing potential profits while minimizing the resulting accuracy losses.

3. DISCUSSION / FUTURE DIRECTIONS

Examining the 63 long papers from the last two years of the ACM RecSys Conference, we observe that about 85% of all papers use at least one of several possible accuracy measures for rating prediction and item ranking. About 20% of papers aim at improving algorithms in terms of diversity, novelty, and other alternative quality measures. About another 20% look at domain- or business-oriented measures like click-through rates. This shows that the field has a largely agreed-upon, shared understanding of recommendation problems. However, this also illustrates that we are perhaps too strongly focused on a certain set of computational metrics and a limited set of established system tasks.

We hope that the proposed framework will help to facilitate the discussion about revisiting the more foundational aspects related to the purposes and goals of recommender systems and to reconsider and extend the way we operationalize the recommendation problem in academic settings. The framework also points to different possible directions for future work, e.g., in the following areas.

3.1 Challenges related to Tasks and Metrics

From the operational perspective of system tasks and computational metrics, a number of general challenges arise when we consider that recommendations can serve multiple purposes. More research is needed for understanding potential trade-off situations and how to systematically evaluate them with standardized approaches. E.g., consider the recent stream of research on diversified and serendipitous recommendations – in most domains higher diversity is achieved only with compromises on ranking accuracy. No clear understanding, however, exists on *how much* diversity is actually desirable (for a domain) or how large of a compromise on accuracy should be tolerable. As a result, this calls for more standardized multi-metric optimization and evaluation schemes in the context of a specific task.

Whenever new recommendation purposes are addressed and corresponding system tasks are designed, the question arises whether the existing metrics are appropriate to assess the system's ability to accomplish the task. Looking again at the diversity problem, using the average pairwise diversity of items using metadata features is one common way to assess the diversity level of a recommendation set. However, even for such a common metric, it is often not clear if it reflects something that is truly valuable to the consumers for a given recommendation purpose. Whenever new metrics are developed for a new task, it is important to empirically validate, e.g., through user studies or live tests, to which extent the quantitative metric is capable of truly capturing what it should it measure (i.e., the metric-task-purpose fit).

A general dilemma in this context is that certain recommendation purposes and corresponding system tasks might be very specific to a given application domain. The challenge is to identify or develop the next wave of metrics that can generalize across different recommendation problems/domains – such metrics would provide a standard set of problems for a research community to focus on. Notable examples of highly desirable metrics include metrics that could help predict online recommendation success from offline experimentation, since the way online success is measured can often be very idiosyncratic for different domains, purposes, and tasks [2, 3].

3.2 Data and Protocol Issues

Many of the sketched directions for future research require the existence of new evaluation protocols and benchmark data sets, which carry much more information than the snapshots of explicit rating databases that serve as a basis for much of today's research. As an example, consider the RecSys 2015 challenge where the tasks were to predict from click-stream data whether a visitor of an e-commerce site will make a purchase in a given session and which item will be eventually purchased. As this problem setting can be relevant for a number of e-commerce recommendation purposes, it has the potential to be further developed to a standardized *system task* in our framework. But, general principles still have to be developed with respect to appropriate evaluation measures and protocols – in the challenge, a proprietary metric was applied – or with respect to the types of data that are considered to be generally available for such a task.

3.3 Balancing Overarching Goals: Leveraging RECO-nomics

Consumer and provider goals can be competing or even incompatible, as mentioned earlier, and not much research exists in the recommender systems literature that addresses this problem. In contrast, balancing competing issues like supply and demand, buyer's surplus and seller's revenue, goals of goods/service providers and consumers, etc. are topics that have been extensively researched in the economics, marketing, and consumer research literature. As one example, in economics *allocative efficiency* represents a metric of social welfare for a given economic mechanism (say, a specific type of auction), i.e., a metric of how well the mechanism supports the interests of buyers and sellers. The recommender systems community should be looking to borrow some ideas from economics and related fields to develop recommendation approaches for key purposes and tasks that can represent *both* consumers and providers in a systematic and principled manner. Even though various economic models and simulations are based on simplifying assumptions, economic modeling in recommender systems (i.e., *RECO-nomics*) represents a highly promising research direction.

4. CONCLUSIONS

The recommender systems research field has reached a certain level of maturity, in particular in terms of its canonical, established operationalizations of the recommendation problem. We advocate that it is, therefore, the right time to step back and revisit some of the more foundational aspects of recommender systems and to reconsider the variety of purposes for which such systems are already used today in a more systematic manner. The high-level research framework proposed in this paper is intended to serve as a means to approach these fundamental aspects in a structured way by considering not only the operational but also the strategic perspective for recommender systems design and evaluation.

5. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TKDE*, 17(6):734–749, 2005.

[2] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *RecSys '14*, pages 169–176, 2014.

[3] C. A. Gomez-Uribe and N. Hunt. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM TMIS*, 6(4):13:1–13:19, 2015.

[4] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM TOIS*, 22(1):5–53, 2004.

[5] D. Jannach and K. Hegelich. A case study on the effectiveness of recommendations in the mobile internet. In *RecSys '09*, pages 205–208, 2009.