Understanding local structure in ranked datasets

Julia Stoyanovich*
Drexel University, USA &
Skoltech, Russia
stoyanovich@drexel.edu

Sihem Amer-Yahia
CNRS, LIG, France
sihem.amer-yahia@imag.fr

Susan B. Davidson*
Univ. of Pennsylvania, USA
susan@cis.upenn.edu

Marie Jacob
Univ. of Pennsylvania, USA
majacob@cis.upenn.edu

Tova Milo†
Tel Aviv University, Israel
milo@cs.tau.ac.il

ABSTRACT

Ranked data is ubiquitous in real-world applications. Rankings arise naturally when users express preferences about products and services, when voters cast ballots in elections, when funding proposals are evaluated based on their merits and university departments based on their reputation, or when genes are ordered based on their expression levels under various experimental conditions. We observe that ranked data exhibits interesting local structure, representing agreement of *subsets of rankers* over *subsets of items*. Being able to model, identify and describe such structure is important, because it enables novel kinds of analysis with the potential of making ground-breaking impact, but is challenging to do effectively and efficiently. We argue for the use of fundamental data management principles such as declarativeness and incremental evaluation, in combination with state-of-the-art machine learning and data mining techniques, for addressing the effectiveness and efficiency challenges. We describe the key ingredients of a solution, and propose a roadmap towards a framework that will enable robust and efficient analysis of large ranked datasets.

1. INTRODUCTION

Ranked data is ubiquitous in real-world applications. Rankings arise when users express preferences about products and services, when voters cast ballots in elections, when funding proposals are evaluated based on their merits and university departments based on their reputation, or when genes are ordered based on their expression levels under various experimental conditions. A ranking represents a statement about the relative quality, or relevance, of the items being ranked.

When multiple rankings are present, these must be aggregated by the system to facilitate analysis. For example, aggregated opinions of Computer Science faculty may serve as basis for a ranking of CS departments, while aggregated user opinions may be used for content recommendation, or in support of data exploration. Finally, aggregated rankings of genes may be used for coexpression analysis, providing evidence that groups of genes are involved in a common biological pathway [1, 12, 24].

It has been observed that ranked datasets exhibit interesting structure [19], reflecting agreement of all, or a subset, of the rankers (also called *judges*) with respect to the items. We argue here that aggregation of rankings is most meaningful in presence of structure (i.e., of agreement), and, conversely, that structure must be identified before meaningful aggregation can take place. Intuitively, an aggregated ranking is only meaningful if it is representative of the rankings it aggregates.

Rank aggregation has been studied in a variety of fields. A notable example is the theory of social choice [4], where the (unattainable) goal is to arrive at a "correct" collective preference given individual choices. A well-known result, which serves as basis of modern social choice theory, is Kenneth Arrow's impossibility theorem [3]. According to this theorem, it is not possible to convert ranked preferences of individuals over three or more items into a consensus ranking, while meeting certain natural criteria. Arrow's work has been very influential, earning him the 1972 Nobel Prize in Economics. One of Arrow's collaborators, Amartya Sen, makes the following statement in his 1998 Nobel Prize acceptance speech:

> If there is a central question that can be seen as the motivating issue that inspires social choice theory, it is this: how can it be possible to arrive at cogent aggregative judgments about the society (for example, about "social welfare", or "the public interest", or "aggregate poverty"), given the diversity of preferences, concerns, and predicaments of the different individuals *within* the society?

Understanding and modeling the diversity of preferences, and being able to efficiently derive a representation of such preferences, motivates our vision.

Another domain where rank aggregation is of central importance is biology. Ranked lists are a very natural, and common, way to represent results of genome-wide studies, particularly for the purpose of meta-analysis [12]. Results from different studies may not be directly comparable, i.e., it may not be possible to normalize gene expression levels that were measured using different platforms, and under different experimental conditions. Rank aggregation has been used to, e.g., find gene co-expression networks [24], identify genes implicated in particular diseases [1, 10], and combine results of differential expression analysis obtained with different algorithms [6]. The main challenge in these applications is in picking up statistical signal from ranked datasets that are large, sparse and noisy, and in understanding robustness of this signal.

*This research was supported in part by Google Research Award "Identifying Ranked Agreement Among Raters".

†This work has been partially funded by the European Research Council (FP7/2007-2013) / ERC grant MoDaS, agreement 291071, by the Israel Ministry of Science, and by the US-Israel Bi-national Science foundation.

6th Biennial Conference on Innovative Data Systems Research (CIDR '13) January 6-9, 2013, Asilomar, California, USA.

This paper proposes a framework for effective modeling and efficient identification of local structure in ranked datasets. This structure is of independent interest, and can serve as basis for rank aggregation, enabling novel kinds of data analysis in economics, biology, and beyond. It is precisely through locality that we intend to model the diversity of preferences of members of a society to which Amartya Sen refers. And it is through locality (and the associated *dimensionality reduction* techniques) that we intend to address the challenge of picking up statistical signal in complex high-dimensional biological datasets. We now illustrate local structure in ranked data with an example.

Consider Jane, a resident of Williamsburg, Brooklyn, who is an active user of a popular shopping website. Jane maintains a detailed profile of her likes and dislikes of styles, looks, designers, and products. She also rates and ranks products based on her current preferences. Many of Jane's friends are also avid users of the site. Jane and her friends look to the site to provide them with recommendations of boutiques, designers, and events.

Like most users', Jane's interests are diverse, as reflected in her fashion sense. Jane represents an increasing fraction of Williamsburg residents who are financial analysts by day and hipsters by night. A natural way to produce a ranking of items for Jane is to retrieve all items that were ranked by Jane's friends, and by other users with interests similar to Jane's, and to compute an aggregate ranking. However, this may not produce an informative result, for the following reasons:

- *Domain diversity*: Different sets of items may be ranked by different groups of users. For example, one group of users may rank only high-end stilettos, while another may rank only tennis shoes. In this case, it is not helpful to aggregate all rankings into a single list, while aggregating rankings within groups may be appropriate.
- *Diversity of opinion*: Even when the same set of items is ranked by two groups of users, these groups may express conflicting opinions over the items, e.g., ranking them in the opposite relative order. Jane's hipster friends will strongly dislike Armani suites and Jimmy Choo shoes and will strongly like vintage clothing and TOMS shoes, while Jane's co-workers will have the opposite preference. Aggregating rankings in which items are ranked similarly, producing two ranked lists, one for hipsters and one for yuppies, will give rise to two consensus rankings and is desirable, while aggregating divergent rankings into a single list is not.
- *Locality of agreement*: Users may rank some items in common, but not others, and they may rank some items similarly, giving rise to a consensus ranking, and others — dissimilarly, producing divergent rankings. User opinions should thus be aggregated only over the items on which they agree. In our example, Jane's ranking of evening clothes should be aggregated with rankings by her hipster friends, and her ranking of business attire — with rankings by her yuppie friends.

Figure 1 illustrates local structure in ranked data, with fruits and vegetables representing items. Understanding nutritional choices is another application domain for the techniques described here, one that is no less important than shopping, and is a bit easier to represent pictorially.

Outline. In this paper we present our vision for a framework for the analysis of local structure in ranked datasets. We outline the challenges (Sec. 2), describe our ongoing work on effective modeling (Sec. 3) and efficient identification (Sec.4) of local structure in ranked datasets, and briefly survey related work (Sec. 5).

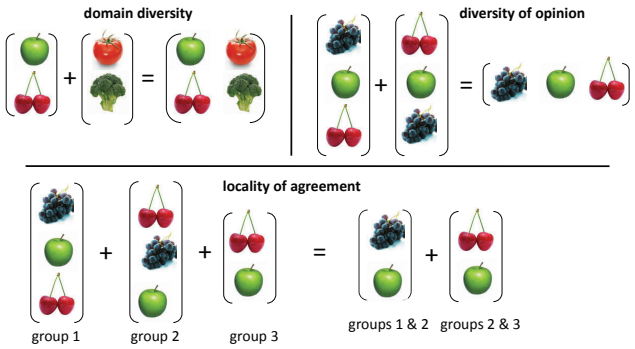

Figure 1: An illustration of local structure in ranked data.

2. CHALLENGES AND ROADMAP

The following inter-related challenges must be addressed to make modeling ranked data practical.

Challenge 1: size and sparseness. Typical applications deal with hundreds, or even thousands of items (genes, books, movies, restaurants), giving rise to a potentially intractable space of possible rankings. This is because there are $m!$ possible ways to order m items. Furthermore, because the space is so large, we expect the set of actual observations to be sparse. Thus, when reasoning over the full space of possible rankings, we cannot expect to derive sufficient statistical power from the observations when m is large. That is, size and sparseness make it difficult to learn robust models from the data.

Challenge 2: handling partial rankings. Observed rankings will be partial (or incomplete), as it is unreasonable to expect that, e.g., each user states a preference with respect to each item, or that expression levels of all genes are measured in each experiment. Notably, even if it were possible to compel the user to provide a ranking of all items, this would likely lead to noisy data, because, from the user's point of view, not all items may be directly comparable. One option is to remove partial rankings from considerations, however, this was shown to decrease model quality [7, 15].

Challenge 3: efficiency. Size of the state space and sparseness of the observed data bring efficiency considerations front-and-center. Efficiency here refers both to the running time of data analysis algorithms and to requirements in terms of sample size. Partial rankings exacerbate this problem, since reasoning about them typically corresponds to reasoning over their (many) completions [2, 8, 19]. Because of efficiency considerations, most datasets analyzed in the statistics and machine learning literature are limited to $m \leq 20$, and usually even $m \leq 5$ items. Notable exceptions are the work of Lebanon and Mao [15], who demonstrate scalability to over 1500 items for particular classes of rankings, and the work of Lu and Boutilier [17], who handle pair-wise preference data and show scalability to about 200 items.

We see potential for addressing these challenges in leveraging important insights developed in the data management community. In particular, the approach we envision will (1) declaratively specify properties of the desired model; (2) use semantic information encoded by attributes to define and navigate the state space; (3) focus on scalability and efficiency.

Locality is key! Our focus on local structure in ranked data is both novel from the point of view of modeling, and will alleviate the efficiency concerns discussed above. Ours is a point of view of dimensionality reduction, a family of techniques that have been successful due to their ability to reduce the size of the state space. *The main technical novelty of the approach discussed here is pre-*

cisely in the development of rank-aware dimensionality reduction methods. Looking for multiple local models, rather than fitting a single global model to the data, effectively breaks up the problem into smaller sub-problems. Such models have the potential of fitting the data better, are more computationally tractable, and have reasonable sample size requirements.

3. MODELING LOCAL STRUCTURE

Modeling ranked data. Consider a dataset in which m items are being ranked. We use permutations to represent rankings of items. A permutation is a bijective function $\pi : \{1, \ldots, m\} \rightarrow \{1, \ldots, m\}$, associating with each item $i \in \{1, \ldots, m\}$ a rank $\pi(i)$. $\pi(i)$ denotes the rank of item i and $\pi^{-1}(i)$ denotes the item at rank i. Consider items Apple ($id = 1$), Broccoli ($id = 2$), Cherry ($id = 3$), Tomato ($id = 4$). A ranking in which Apple is liked best, followed by Cherry, then by Tomato, and finally by Broccoli, corresponds to permutation $\pi(1) = 1$, $\pi(2) = 4$, $\pi(3) = 2$, $\pi(4) = 3$. We represent this permutation using the vertical bar notation $\pi^{-1}(1)|\pi^{-1}(2)|\pi^{-1}(3)|\pi^{-1}(4)$ as $1|3|4|2$.

There are $m!$ possible permutations of m items, and we denote by $\mathcal{S}_m$ the space of permutations. A dataset Y will consist of n observations, drawn iid from the probability distribution associated with $\mathcal{S}_m$. The central question in modeling ranked data is: what are the assumptions regarding the probability distribution over $\mathcal{S}_m$?

A natural starting point is to assume that all rankings in $\mathcal{S}_m$ are equally likely: $H_0 : Y \sim Uniform(\mathcal{S}_m)$. We may then consider the sample Y and either support or reject the null hypothesis. Several basic statistics have been developed for testing uniformity, with stringent sample size requirements.

Another possibility is that there exists a *modal* ranking $\sigma \in \mathcal{S}_m$, around which the observations in Y cluster. To represent this case, i.e., to give semantics to clustering, we need a notion of *distance* over the space $\mathcal{S}_m$. A variety of distances have been proposed in the statistics literature [19]. Perhaps the most commonly used is the Kendall distance, which, for a pair of permutations π and π', counts the number of discordant pairs, i.e., the number of pairs of items that appear in the opposite relative order in π and π'. For example, Kendall distance between $1|3|4|2$ and $1|4|2|3$ is 2.

Distance-based models are popular in statistics and in machine learning, and particular attention has been paid to *Mallows models* [18], which are essentially distance-based models parameterized by a modal ranking $\sigma \in \mathcal{S}_m$, a dispersion parameter $\lambda \in \mathbb{R}_+$, and a distance function with metric properties, e.g., Kendall distance. Distance-based models are commonly used for clustering, e.g., for identifying k modal rankings and dispersion parameters that provide the best fit to the data, for a given k. For example, given a sample Y in which a third of the rankings correspond to $\pi = 1|2|3|4$, another third to $\pi' = 4|3|2|1$ ($d_{Kendall}(\pi, \pi') = 4$), and the remaining rankings are within Kendall distance 1 from either π or π', a clustering algorithm based on the Mallows model will have no trouble discerning the structure.

Handling partial rankings. We argued in the introduction that observed rankings will often be partial, particularly when m is reasonably large. Rankings may be partial for two reasons. First, a full set of preferences may not be available, e.g., a user who has never tasted dragon fruit cannot rank it relative to other fruit. Second, not all items may be directly comparable, e.g., a judge may prefer to give one ranking for fruits and another for vegetables.

In the machine learning literature it is often assumed that a partial ranking is a top-t ranking, i.e., that judges have ranked their t favorites out of a large number of r items. For example, [7] make this assumption. In [11] the authors work with partial observations that can be decomposed (factored). This class corresponds to top-t observations, including also the desired / less desired dichotomy (t items are preferred to the remaining items), when ties are allowed. A more general model of incomplete rankings is presented in [15], where the goal was to efficiently learn a model from heterogeneous ranked datasets. Here, heterogeneity refers to rankings being incomplete in different ways.

Leveraging attributes. Importantly, we observe that rankings of items often correlate with item attributes, such as designer and price for clothing, cuisine and ambiance for restaurants, or biological annotations for genes. That is, how items are ranked depends on what they are. Furthermore attributes of the judges such as age, income, and profession of a user in a shopping application, or a description of experimental conditions in the genetics example, may also correlate with rankings. That is, how items are ranked depends on who, or what, is ranking them.

Leveraging attributes has two important advantages. The first is computational: attributes may be used to limit the search space, and to guide its systematic exploration. The second advantage is equally as important, and is one of usability: if attributes are used to guide the search for structure, then the identified structure can be naturally described using these attributes. Returning to our example, if two consensus rankings are identified for Jane, these may be shown to her together with an explanation of the items they contain, and of the judges whose opinions were aggregated to produce them. The first ranking may be of vintage boutiques that are well-liked by Jane's neighbors in the 30-35 age group, while the second ranking may be of high-end fashion stores on Madison Avenue, well-liked by people with an annual income in the \$150K-\$200K range.

Combining multiple models. We envision a system that uses local models as building blocks. A *local model* $M(I)$ represents the probability distribution over a *subset of items* I , i.e., is defined over a projection of $\mathcal{S}_m$ onto I. Considering structure in projections of $\mathcal{S}_m$ is a natural way to address partial rankings, and to handle the related efficiency concerns. The model M may be parametric, and may correspond to, e.g., a Mallows model, or it may be non-parametric as in [15].

Multiple local models are defined over $\mathcal{S}_m$, and together form a meta-model $\mathcal{M}$. Importantly, models in $\mathcal{M}$ are defined over possibly overlapping subsets of items. For example, one model may represent the rankings of fruits, another — the rankings of iron-rich foods (which include certain fruits, vegetables and meats), yet another — the rankings of sweet foods, etc. To accommodate diversity of opinion (see Figure 1), we allow multiple models to be defined over the same set of items. So, we may have models $M_1(I_1), M_2(I_2) \in \mathcal{M}$, with $I_1 = I_2$. Another important ingredient is our representation of a *population of judges* J, encoded using a probability distribution over $\mathcal{M}$. A meta-model is associated with a population of judges J, parameterized by J: $\mathcal{M}(J)$.

4. IDENTIFYING LOCAL STRUCTURE

Efficiency is of central importance for the work discussed here, yet, as argued in the introduction, is difficult to achieve due to the size of the state space, the complexity of the models, and the need to handle heterogeneous partial rankings. Efficiency concerns in the modeling of ranked data have recently come into focus in the machine learning community, and important progress has been made in [15, 17]. Nonetheless, the state of the art is still far from being able to accommodate datasets in which the number of items is in the millions of items, which is realistic by today's standards.

To achieve scalability, we plan to use the declarative paradigm. We aim to develop a declarative framework that will incorporate, and possibly combine, different kinds of local models as building blocks. The choice of a particular model will be based on two re-

lated criteria: robustness and cost. *Robustness* refers to statistical guarantees, goodness of fit, or generalization properties. An interesting and important question, which we will address as part of this work, is how robustness of individual local models translates to robustness of the over-all meta-model, i.e., how to guarantee that the combined model is consistent. *Cost* refers to sample size requirements, and to the running time of fitting the model to the data.

We plan to leverage our recent work on rank-aware clustering [21, 22], where the idea was to identify correlations that hold between item attributes and ranking. We instantiated our model in scope of the Bottom-up Algorithm for Rank-Aware Clustering (BARAC). We demonstrated scalability on real datasets with millions of items, showing that structure can be identified in interactive time. The technique that allowed for efficient running time was based on *a bottom-up search* strategy that identified *local correlations* in subspace projections of high-dimensional datasets. We plan to build on this insight, and to consider how bottom-up techniques can be incorporated into the framework proposed here.

Rankings form a partial order, more precisely, a join semi-lattice, expressing more general, more specific, or inconsistent orderings of items. For example, $1|2,3$ represents a partial ranking in which 1 is preferred to both 2 and 3, and there is no relative preference between 2 and 3. The full ranking $1|2|3$ is compatible with the partial ranking $1|2,3$, and so $1|2|3$ is a descendant of $1|2,3$ in the semi-lattice. In [15] the authors relied on the partial order to develop efficient algorithms for learning a non-parametric model from heterogeneous rankings. In our own recent work we developed an efficient method for learning a collection of local models from structured datasets with missing values [23]. The models, called meta-rules, encoded conditional independence assumptions, and were used in scope of an ensemble, called a meta-rule semi-lattice. We demonstrated that this model can be learned efficiently, and that it provides accurate probability estimates with reasonable sample size requirements and running times.

In [23] we were able to realize efficiency in terms of both sample size and running time due to lazy evaluation and sharing of computation, techniques that are characteristic of declarative approaches. Specifically, in our case lazy evaluation corresponded to *on-demand sampling* from the parts of the probabilistic space in which there was not enough data available to guarantee robustness. Sharing of computation was enabled by the partial order over the tuples (items) in the workload. We will build on the insights of [23] to develop efficient algorithms as part of a declarative framework.

5. RELATED WORK

Towards the goal of understanding the structure of ranked data, a variety of models have been developed in the statistics literature [5, 8, 9, 19]. In recent years, there has been increased interest in analyzing ranked data in the machine learning [7, 11, 14, 15, 17], information retrieval [16], and bioinformatics [1, 12, 24] communities. To the best of our knowledge, none of the previously proposed approaches directly model the local structure of ranked data.

Learning to rank is an active area of research in information retrieval [16], where the objective is to predict a ranking of unseen items based on their features. While there may be some common technical insights linking learning to rank and the approach described here, the goals of the two lines of work are essentially different. Our goal is to identify structure in datasets containing *observed items*, to explain the structure using attributes (features) of items and of judges, and to produce *multiple representative rankings*. In contrast, in learning to rank the goal is to predict a *single global ranking* for *unseen items*, based on their features. Nonetheless, we plan to explore deeper connections between our proposed approach and learning to rank as our work progresses.

Our proposed approach builds on insights from data mining, in particular on subspace clustering [13, 20]. We build on our own prior work on rank-aware clustering [21, 22].

6. REFERENCES

[1] S. Aerts et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5), 2006.

[2] M. Alvo and P. Cabilio. On the balanced incomplete block design for rankings. *The Annals of Statistics*, 19(3).

[3] K. J. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4), 1950.

[4] K. J. Arrow, A. K. Sen, and K. Suzumura. *Handbook of Social Choice and Welfare*, volume 1. North-Holland, 2010.

[5] L. Beckett. A censored ranking problem. In *Probability Models and Statistical Analyses for Ranking Data*. 1993.

[6] A.-L. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5), 2009.

[7] L. M. Busse, P. Orbanz, and J. M. Buhmann. Cluster analysis of heterogeneous rank data. In *ICML*, 2007.

[8] D. E. Critchlow. *Metric methods for analyzing partially ranked data*. Springer, 1985.

[9] P. Diaconis. A generalization of spectral analysis with applications to ranked data. *Annals of Statistics*, 17(3), 1989.

[10] O. L. Griffith et al. Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. *J Clin Oncol*, 24(31), 2006.

[11] J. Huang, A. Kapoor, and C. Guestrin. Efficient probabilistic inference with partial ranking queries. In *UAI*, 2011.

[12] R. Kolde et al. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4), 2012.

[13] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *TKDD*, 3(1), 2009.

[14] G. Lebanon and J. D. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *ICML*, 2002.

[15] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *JMLR*, 9, 2008.

[16] T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[17] T. Lu and C. Boutilier. Learning mallows models with pairwise preferences. In *ICML*, 2011.

[18] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1/2), 1957.

[19] J. I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

[20] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explorations*, 6(1), 2004.

[21] J. Stoyanovich and S. Amer-Yahia. Rank-aware clustering of structured datasets. In *CIKM*, 2009.

[22] J. Stoyanovich, S. Amer-Yahia, and T. Milo. Making interval-based clustering rank-aware. In *EDBT*, 2011.

[23] J. Stoyanovich et al. Deriving probabilistic databases with inference ensembles. In *ICDE*, 2011.

[24] J. M. Stuart et al. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302, 2003.