Approximate Query Answering with Frequent Sets and Maximum Entropy

Heikki Mannila
Nokia Research Center
P.O. Box 407
FIN-00045 Nokia Group, Finland
Heikki.Mannila@nokia.com

Padhraic Smyth
Information and Computer Science
University of California
Irvine, CA 92697-3425
smyth@ics.uci.edu

We describe an approach to finding approximate answers to Boolean queries on 0/1 data. The basic idea is to build a probabilistic model for the data set and answer the queries on the basis of that model. The method consists of the following steps:

1. Count generation: scanning the dataset initially to generate some summary count information.

2. Model building: given a query, use the count information to construct an approximate probability model for the variables occurring in the query.

3. Approximate querying: answering queries using the probability model.

To instantiate this scheme, we combine two general and useful techniques: the summary information provided by *frequent sets* [1] and the probabilistic estimation principle of *maximum entropy* [2].

Our approximation method is as follows. Given the table r, we first compute the collection $\mathcal{C}$ of frequent sets of r for some suitable threshold σ. This is the summary of the data from which we compute the approximate answers. Given an arbitrary query φ over the table, let S be the set of attributes that occur in φ. We find from $\mathcal{C}$ all frequent sets that are included in S, and construct the maximum entropy distribution on S using those frequent sets as constraints. Then, we evaluate φ on the maximum entropy distribution and give the answer as the approximate answer.

This approach is useful for any type of queries: as we construct the distribution on S, we can compute φ on that distribution regardless of what the actual form of φ is. The complexity of the method is independent of the size of the data (after the initial computation of the frequent sets), linear in the number of frequent sets contained in S, and exponential in $\mid S \mid$, the number of variables occurring in the query. Thus the method is useful for any size of data and very tolerant of the number of constraints. The main limitation of the method is the exponentiality in the number of variables occurring in the query, limiting its application in practice to queries involving no more than 10 variables.

Data	Query	No.	Maxent	Ind.	Incl/
Set	Size	Queries			Excl
Product site	4	1000	0.10	0.23	0.61
	6	5955	0.12	0.37	1.26
TV shows	4	1868	0.34	0.48	1.26
News site A	4	263	0.30	0.57	1.26
	6	1705	0.22	0.75	1.98
News site B	6	405	0.04	0.09	0.13

Table 1: Performance of different query approximation methods across different data sets expressed as the percentage average absolute error in frequency between the true query size and the estimate.

Empirical results on a number of sparse binary data sets (Table 1), using synthetically generated queries, show that the method produces very accurate approximations, significantly outperforming the simple variable-independence method and the use of the inclusion-exclusion formula.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'93)*, pages 207 – 216, Washington, D.C., USA, May 1993. ACM.

[2] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480, 1972.