Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Application of Nonparametric Techniques to Collaborative Recommender Systems

Barbara D. Broome
University of Maryland - Baltimore County, bbroom1@umbc.edu

Malcolm S. Taylor
US Army Research Laboratory, mtjusho@aol.com

Victoria Yoon
University of Maryland - Baltimore County, yoon@umbc.edu

Application of Nonparametric Techniques to Collaborative Recommender Systems

Barbara D. Broome
University of Maryland Baltimore County
BBroom1@umbc.edu

Malcolm S. Taylor
US Army Research Laboratory (Ret.)
mtjusho@aol.com

Victoria Yoon
University of Maryland Baltimore County
yoon@umbc.edu

ABSTRACT

The introduction of the World Wide Web dramatically impacted our fundamental notion of information sharing, providing unparalleled awareness of both the power of information access and the penalty of information overload. Today's research on Semantic Web techniques focuses on the next step, a Service Oriented Architecture supporting automated sharing of services as well as data. Personalized service/source recommendation tools, utilizing user preference data, would be extremely valuable in tailoring information access to the user. Much can be learned from the Recommender community about incorporating preference data into the retrieval process. However, it is critical that rigorous statistical techniques be maintained in combining results across data and service sources that are not under the control of a single developer. In this paper we explore the extension of nonparametric techniques to the development of Collaborative Recommenders and its impact on establishing a generalized recommendation service within a Service Oriented Architecture.

Keywords

Recommender, preference, nonparametric statistics, Service Oriented Architecture.

INTRODUCTION

The introduction of the World Wide Web dramatically impacted our fundamental notion of information sharing, providing an unparalleled awareness of both the power of information access and the penalty of information overload. Today's research on Semantic Web techniques focuses on the next step in information sharing, Service Oriented Architectures (SOAs) that support the automated exchange of data and services in a networked environment in ways that promote not just access, but software-controlled access (Hendler, Berners-Lee and Miller, 2002; Curbera, Khalaf, Mukhi, Tai and Weerawarana, 2003). Personalized service and source recommendation tools potentially can provide an extremely valuable niche within the SOA, that of tailoring information and service access to the user, thereby avoiding information overload.

RECOMMENDER SYSTEMS

Over the past ten years, considerable research has focused on Recommenders, defined as "any system that produces individualized recommendations as output or that has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options (Burke, 2002)." Those efforts have been categorized both by the techniques they employ and by the data they access (Sarwar, Karypis, Konstan and Riedl, 2000; Burke, 2002). *Collaborative Systems* base their recommendations on the previous rankings of other users for objects in the information space (Goldberg, Nichols, Oki and Terry, 1992; Herlocker, Konstan, Borchers and Riedl, 1999; Vucetic and Obradovic, 2004; Herlocker, Konstan, Terveen and Riedl, 2004) while *Content-Based Systems* base their recommendations on the attributes of objects in the information space (Liu, Yu and Meng, 2004; Pazzani, 1999; Burke, 2002; Burke, Hammond and Cooper, 1996) or on attributes of the users (Pazzani, 1999; Burke, 2002). Hybrid approaches exist that, for example, use content-based data to restrict the information space while incorporating collaborative data for the ranking algorithm (Balabanovic and Shoham, 1997; Burke, 2002). In this paper we focus on collaborative techniques applied in both Collaborative and Hybrid Systems. This allows us to address a single data type (preference data) that is critical to the personalization process.

In general, the data to which collaborative techniques are applied, i.e., user-generated ranking data, is *ordinal* in nature. The distinction between nominal, ordinal, interval, and ratio data is an important one. Rankings (like 1=poor, 2=good, 3=excellent) partition the data into mutually exclusive categories, as with *nominal* data. In addition they provide an indication of order which is transitive in nature. However, unlike *interval* and *ratio* data, the concept of distance between rankings is *not* preserved. (For example, User 1 gives a "poor" rating to Movie A, which he hates. He rates Movie B "good," because he did not enjoy it but thinks it has redeeming social value. Movie C is almost rated "excellent," but because of its genre is downgraded to "good." Movies B and C have the same rank; but for User 1, they are not the same distance from Movie A.) The observation that the data is ordinal is significant because the data type determines what the developer can legitimately do in applying statistical techniques to generate and evaluate algorithms to predict future rankings. In the case of ordinal data, we are restricted to the use of nonparametric statistical techniques, rather than the more familiar parametric techniques that may be applied to interval and ratio data.

Despite the ordinal nature of the data, many Collaborative Recommender algorithms employ parametric techniques. The impact of this is that these methods cannot be broadly applied in developing a general purpose recommender "service." While the algorithms may serendipitously work with a particular data set, there is no sound statistical foundation that allows us to apply those techniques to rank data in general or to combine ranking algorithms across data sets. That is, a given parametric method may work for the restaurant ranks on which it was developed, but there is no reason to assume it will apply equally well to CD ranks, or book ranks, or movie ranks, or even to a different set of restaurant ranks.

NONPARAMETRIC TECHNIQUES

The transition to nonparametric techniques may impact a number of modelling and evaluation decisions. Applications of regression techniques require methods beyond ordinary least squares regression. The median is often a more robust measure of central tendency than the mean. Predictions that imply more precision than can be derived from ordinal data must be rounded off. Tests of hypotheses must be compatible with ordinal data. Similarity measures used for weighting must be free from unfounded assumptions about the data.

Consider, for example, the use of correlation coefficients to weight a series of collaborative predictions of a user's possible interest in an item. One general approach (Herlocker et al., 1999; Vucetic et al., 2004) for predicting user a's preference for item c (p_{ac}) follows the form

$$p_{ac} = \sum_{j=1}^{n} \omega_{jc} f_{jc}(r_{aj}) \qquad \textbf{Equation 1}$$

where r_{aj} indicates how user-a rated item j, f_{jc} is a function that predicts user a's rating of item c based on the user's rating of item j, and w_{jc} represents the corresponding weight. This resembles a nearest-neighbor algorithm in which the weights, or correlation coefficients, are used to indicate proximity. Because the data are ordinal, it is important to choose a nonparametric correlation coefficient.

Correlation is an expression of relationship. This relationship, however, is not a simple functional relationship of the form y=f(x) in which for every value of x there exists a unique corresponding y, but a statistical relationship where both x and y are random variables. As a consequence, the measure of correlation is itself a random variable, with a distribution depending upon the joint (bivariate) distribution (x, y). The focus here is on the relationship between paired observations (x_1, y_1), (x_2, y_2),…, (x_n, y_n), where (x_i, y_i), i =1, 2, …, n, are realizations of the corresponding random variables–as distinct from mathematical variables. This distinction is important because correlation, being a statistical relationship, requires knowledge of the underlying (x, y)-distribution in order to interpret the significance of a specific computed value. In particular, values of the Pearson product–moment correlation coefficient,

$$\rho_{XY} = Cov(x, y) / \sqrt{V(x)V(y)}, \qquad \textbf{Equation 2}$$

a frequently encountered and widely accepted measure of correlation cannot otherwise be interpreted. Tables of percentiles of ρ_{XY} (Pearson and Hartley, 1976) carry an implicit assumption that the distribution of (x, y) is bivariate normal.

(Herlocker et al., 1999) recognized some of the potential difficulties associated with use of the Pearson correlation coefficient. Noting that the expression Equation (2) or its sample estimate equivalent,

$$r_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \qquad \textbf{Equation 3}$$

has its genesis in linear regression modeling, they comment that "When these model assumptions[1] are not satisfied, Pearson correlation becomes a much less accurate indicator of similarity." and continue, "It is not uncommon for these model assumptions to be violated in collaborative filtering data."

This issue can be circumvented through the use of correlation measures that are free from stringent distribution assumptions. We limit consideration here to nonparametric (distribution free) procedures based on representation of the data by ranks. These approaches provide an additional benefit of accommodating non-numerical data that may be arranged in a preference ordering, e.g., L < K < J.

Spearman's rho

In the development of Spearman's rho, the x- and y-values from the paired observations (x_1, y_1), (x_2, y_2),…, (x_n, y_n) described above are transformed into their associated ranks. That is, the values $x_1, x_2, \ldots, x_n$ are ordered according to magnitude: $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ and then, letting the notation r(*) denote rank, we have $r(x_{[i]}) = i$; i = 1, 2, …, n. The same procedure is then applied to the values $y_1, y_2, \ldots, y_n$. In the case of tied ranks, e.g. $x_{[i]} = x_{[i+1]}$, a value equal to the average of the ranks that would have been assigned, had there been no ties, is given to each tied value.

Spearman's rho is formally defined as

$$rho = \sum_i [r(x_i) - (n+1)/2][r(y_i) - (n+1)/2]/[n(n^2-1)/12]. \qquad \textbf{Equation 4}$$

This expression can be shown to be algebraically equivalent to computing Pearson's r_{xy} on the assigned ranks, for the case in which there are no ties (Conover, 1971). In other words, it is a rank transformation of the Pearson r_{xy} correlation coefficient. A modified form of Equation (4) to accommodate tied ranks is available.

This simple transformation carries a number of valuable benefits. The most important is that the distribution of Spearman's rho, unlike the Pearson r_{xy} correlation coefficient, does not depend upon the joint distribution (x, y). This in turn allows Spearman's rho to be used as a test statistic to question whether the random variables x, y are independent (uncorrelated) and to attach a level of significance to the conclusion. This level of significance, in the modeling of collaborative filtering systems, provides a rigorous method of assigning similarity weights between paired users responsible for providing the x- and y-values (x_1, y_1), (x_2, y_2), …, (x_n, y_n) under study.

APPLICATION OF NONPARAMETRIC CORRELATION TECHNIQUES TO RECOMMENDERS

The GroupLens project at University of Minnesota maintains a set of 100K movie ratings collected over a 7 month period. Using these data (Herlocker et al., 1999) established a test set of about 500 ratings for which rating predictions were made and then compared to the actual values. Two measures of accuracy were used, Mean Absolute Error and Receiver Operating Characteristic. They demonstrated that Spearman's rho can give slightly better accuracy than Pearson's r when used as a similarity measure in the model:

$$p_{ac} = \bar{r}_a + \left[\sum_{h=1}^{m} (r_{hc} - \bar{r}_h)\omega_{ah}\right] \div \sum_{h=1}^{m} |\omega_{ah}| . \qquad \textbf{Equation 5}$$

This model predicts active user-a's rating of movie-c, where $\bar{r}_a$ is user-a's average rating, $(r_{hc} - \bar{r}_h)$ is the difference between user-h's rating of movie-c and his average rating, and ω_{ah} is the correlation coefficient for the ratings of user-a and user-h. In a sense, other users in the system are telling user-a whether movie c is better than average or not. User-a listens to

[1] Including normality assumptions on the joint distribution (x, y) of random variables x, y.

their advice and adjusts his rating accordingly. But first, he weights their advice, based on how closely their ratings are correlated with his own. We refer to this model as *User Adjusted Generic*.

Following that same framework and using that same data source, we plan to extend the effort to additional models, again comparing the effect of Spearman's rho weights versus Pearson's r. For example, we consider a model which is similar to Equation 5, but uses medians in place of means. We refer to this as the *Median User Adjusted Generic* model.

$$p_{ac} = Md(\bar{r}_a) + \left[\sum_{h=1}^{m}(r_{hc} - Md(\bar{r}_h)\omega_{ah}\right] \div \sum_{h=1}^{m}\left|\omega_{ah}\right| \qquad \textbf{Equation 6}$$

Next, there are a number of other nonparametric correlation measures, for example, Kendall's tau, Bell-Doksum's Z, Somers's d, and Gamma, to name a few. We plan to consider several of these alternatives as potential measures of similarity.

Finally, we intend to explore the use of significance measures to better refine the prediction algorithm. We have less confidence in a correlation estimate if n = 10 than if n = 100. (Herlocker et al., 1999) proposed multiplying the correlation coefficient by n/50 if n < 50. This approach, while possibly effective for a given data set, is inappropriate in a generalized ranking service. The use of nonparametric correlation opens opportunities to explore significance weighting techniques that are not tied to specific data sets and do not rely on assumptions regarding the data's distribution, providing a statistical grounding for significance weighting that can potentially be extended to a generalized recommender.

CONCLUSION

The use of user preference is critical to avoiding information overload. Collaborative Recommenders provide a strong foundation on which to build a generalized recommender service that incorporates user preference into the choice of data and services accessed in an SOA. However, careful attention must be placed on avoiding data-related assumptions that may work in one environment, but lack the statistical rigor required to abstract up to a more general environment. When working with preference data, recognition that nonparametric techniques must be used with ordinal data is one step toward insuring that requisite statistical rigor.

REFERENCES

1. Balabanovic, M. and Shoham, Y. (1997) Fab: Content-based, collaborative recommendation, *Communications of the ACM,* 40**,** 3**,** 66-72.
2. Burke, R. (2002) Hybrid Recommender Systems: Survey and Experiments, *User Modeling and User-Adapted Interaction,* 12**,** 4**,** 331-370.
3. Burke, R., Hammond, K. and Cooper, E. (1996) Knowledge-based navigation of complex information spaces. in *13th National Conference on Artificial Intelligence,* Menlo Park, CA, AAAI Press, 462-468.
4. Conover, W. J. (1971) Practical Nonparametric Statistics, John Wiley & Sons Inc., New York.
5. Curbera, F., Khalaf, R., Mukhi, N., Tai, S. and Weerawarana, S. (2003) Service-oriented computing: The next step in Web Services, *Communications of the ACM,* 46**,** 10**,** 29-34.
6. Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992) Using Collaborative Filtering to Weave an Information Tapestry, *Communications of the ACM,* 35**,** 12**,** 61-70.
7. Hendler, J., Berners-Lee, T. and Miller, E. (2002) Integrating Applications on the Semantic Web, *Journal of the Institute of Electrical Engineers of Japan,* 122**,** 10**,** 676-680.
8. Herlocker, J. L., Konstan, J. A., Borchers, A. and Riedl, J. (1999) An Algorithmic Framework for Performing Collaborative Filtering in *Special Interest Group on Information Retrieval (SIGIR),* Berkley, CA, 230-237.
9. Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. (2004) Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems,* 22**,** 1**,** 5-53.
10. Liu, F., Yu, C. and Meng, W. (2004) Personalized Web Search for Improving Retrieval Effectiveness, *IEEE Transactions on Knowledge and Data Engineering,* 16**,** 1**,** 28-40.
11. Pazzani, M. J. (1999) A Framework for Collaborative, Content-Based and Demographic Filtering, *Artificial Intelligence Review,* 13**,** 5-6**,** 393-408.
12. Pearson, E. S. and Hartley, H. O. (Eds.) (1976) Biometrika Tables for Statisticians, Charles Griffin and Co. Ltd., Buckinghamshire, England.
13. Sarwar, B., Karypis, G., Konstan, J. A. and Riedl, J. (2000) Analysis of Recommendation Algorithms for E-Commerce, University of MN TR 00-047, 1-13.
14. Vucetic, S. and Obradovic, Z. (2004) Collaborative Filtering Using a Regression-Based Approach, Knowledge and Information Systems, *Knowledge and Information Systems.*