Likability-Based Genres: Analysis and Evaluation of the Netflix Dataset

Andrew M. Olney (aolney@memphis.edu)
Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152 USA

Abstract

This paper describes a new approach to defining genre. A model is presented that defines genre based on likability ratings rather than features of the content itself. By collecting hundreds of thousands of likability ratings, and incorporating these into a topic model, one can create genre categories that are interesting and intuitively plausible. Moreover, we give evidence that likability-based features can be used to predict human annotated genre labels more successfully than content-based features for the same data. Implications for outstanding questions in genre theory are discussed.

Keywords: Genre; topic model; Netflix; likability;

Introduction

Many web sites, e.g. Amazon, allow users to rate items along several dimensions, the most common being likability or overall satisfaction. These ratings allow other users to roughly estimate their own probable satisfaction with the item, leading to better item selection and better satisfaction with the web site itself. Moreover, the same rating information can be exploited by a website to make personalized recommendations for the user producing the ratings. In theory, highly accurate recommendations might influence the user to purchase additional products, again leading to greater profitability for the web site in question.

This process of tracking ratings and using ratings to make personal recommendations often falls under the classification of "recommender system" or "collaborative filtering," and is a widely studied problem in the data mining/machine learning field (Resnick & Varian, 1997). To assist the development of new and better algorithms, some companies like Netflix have released large datasets containing hundreds of thousands of ratings by hundreds of thousands of users (*The Netflix Prize Rules*, 2010). These datasets can be analyzed in multiple ways, and an interesting perspective is to view them as a kind of graph or social network. By viewing users as nodes and items as edges, we can study how users are related to each other through item connectivity. Conversely, we can study how items are related to each other through users who have rated them. Another way of looking at this second scenario is as "mass criticism" wherein each user is afforded the same status as a critic, and the mass action of all critics determines not only the overall value of the item (through ratings) but also the association of an item with other items (through connectivity).

In film theory, criticism and genre theory are likewise intertwined (Stam, 2000), creating relationships between the value of film and its taxonomic place. Intuitively, a film might be called, "a good comedy" or "a poor horror," in the sense that the genre defines a kind of rubric or context by which the film is evaluated. Genre theorists often attempt to go beyond such normative characterizations to consider genre in terms of sociocultural effects between film, audience, and author. However, even in a more elaborated perspective, there are a number of outstanding issues in genre theory, which can loosely be divided into problems of definition and problems of analysis.

Problems of definition in genre theory include circularity and the monolithic assumption (Stam, 2000). The problem of circularity arises when one tries to define a genre in terms of features like those given in Table 1.

Table 1: Genre Features (Adapted from Chandler (1997)

Feature	Example
Time	Films of the 1930s
Author	Stephan King
Age of audience	Kid movie
Technology	Animated
Star	Sylvester Stallone
Director	Quentin Tarantino
Structure	Narrative
Ideology	Christian
Culture of origin	Bollywood
Subject matter	Disaster movie
Location	Western

A feature based analysis requires first assembling all the films representative of that genre and then analyzing their features. However, gathering the films requires knowing their genre in the first place, otherwise how would one know which films to assemble? A second problem of definition is the monolithic assumption, in which a film is assumed to belong to one and only one genre. While the monolithic assumption in some ways makes the task of genre definition simpler, it nevertheless ignores genres that are part of our public discourse, e.g. "romantic comedy."

Genre theory is also plagued by problems of analysis. Some questions with regard to genre analysis of film are as follows (Stam, 2000). First, are genres real or imagined? In other words, are they merely analytic constructs, or do they have some status in the world. Second, are the number of genre categories finite or infinite? Third, are genres timeless or are they culture-driven and therefore trendy? Finally, are genres universal, or are they culturebound? As questions about genre, these four questions are inherently tied back to the definition of what genre is. Therefore to answer them, we must first define genre.

In this paper, we analyze the information implicit in user

ratings to build a model of genre. Our study focuses on the ratings from the Netflix dataset, which we incorporate into a probabilistic topic model (Griffiths, Steyvers, & Tenenbaum, 2007). Moreover, we show how the extracted genres can be used to predict human annotated genres with better performance than typical features used by genre critics. That a content-free analysis, based purely on likability ratings, can predict genres is surprising and provocative. We argue that the ability of a likability-based analysis to predict genre with more success than a traditional feature-based approach suggests that likability ratings not only represent a new way of considering genre, but they also represent a significant force in shaping genre categories, a force that is possibly more significant than the content itself.

Study 1: Modeling

Method

The data used in this study consisted of the Netflix dataset, which is freely available online (*The Netflix Prize Rules*, 2010). The dataset has a collection of information applicable to both training a model as well as evaluating the model using the Netflix API (*The Netflix Prize Rules*, 2010). In this study and succeeding studies, only the training data was used. The training data consists of two logical components. The first is a master file which lists for each movie a unique id, along with the title and release year for the movie. The second component is a folder which contains, for each movie id, the set of ratings given to that id by various users. Each rating is a triple consisting of user id, rating, and date of rating. Each rating is an integral number from 1 to 5. There are 17,770 movies in the dataset, 480,189 users, and 100,480,507 ratings. The dataset is sparse, meaning that not every user has rated every movie.

Topic models (Griffiths & Steyvers, 2002; Griffiths et al., 2007), also known in other communities as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), are a class of generative statistical models typically applied to text. Topic models use "bag of words" assumption, making them somewhat similar to methods such as latent semantic analysis (Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007), however there are significant differences. Rather than reduce the dimensionality of the data according an optimal least-squares approximation, topic models use a probabilistic model that assumes the data was generated by an underlying process involving hidden variables. Thus while LSA expresses the data along latent dimensions, i.e. singular vectors, which have no clear semantic interpretation, topic models express the data according to the topics that generated the data, and these topics are expressed as a collection of semantically related words, i.e. the words that are most probable given a topic.

More specifically, the standard topic model makes the following assumptions. For each document, there is an associated distribution of topics. Each of these topics has an associated distribution of words. Thus to generate a document, one first probabilistically samples a from the distribution of topics, yielding a particular topic. One then probabilistically samples from the distribution of words associated with that particular topic, yielding a word. This process can be repeated to generate more words and more documents. Thus a topic model specifies how to generate the observed data; however a model may be fitted to existing data using probabilistic inference. Briefly, this is accomplished by randomly initializing the model and then using Gibbs sampling to reestimate the model's parameters, iteratively, until the model converges. For more details see Griffiths, Kemp, and Tenenbaum (2008).

Though topic models have primarily been applied to text in the cognitive science community, the model itself is agnostic to the underlying data it represents, so long as that data has a form consistent with the assumptions of the model. One generalization of these assumptions would be as follows: data consists of a set of samples, each sample has a distribution of topics, and each item in the sample is generated from one of these topics. It doesn't matter whether the samples are documents or whether the items are words. Using this intuition, it is fairly straightforward to map the Netflix dataset into a form consistent with the topic model. Indeed there are alternate mappings (Rubin & Steyvers, 2009), but in what follows we will only consider one.

Our mapping is as follows. Each customer is a mixture of genres, and each genre is a distribution over movies. To transform the existing Netflix dataset using this mapping, we collect all of the movies seen by a customer. The number of stars given that movie is represented by the number of times that movies label appears. For example, if a customer had only rated the movie "Whale Rider" and gave it three stars, then the customer would be represented as (Whale Rider, Whale Rider, Whale Rider), analogous to a document containing the same word three times. Under the assumptions of this mapping and the underlying topic model, each star in a customer's rating can generated by a different genre. For example two stars of "Whale Rider" might be generated by the drama genre, and one star might be generated by the foreign film genre.

The inference algorithm to fit our model to the Netflix data is identical to that used in typical topic models. However, given the large size of the dataset and the widespread availability of multi-core processors, we have created and make publicly available our code for fast parallel topic models in the C# language [1]. Inference parameters were as follows. The number of topics was 50, the prior for topics appearing in a document (α) was 1, and the prior for words appearing in a topic (β) was 0.01. The α and β smoothing parameters are typical (Steyvers & Griffiths, 2007). The model was run for 200 iterations.

Results

An initial inspection of the genres found by the model reveals intuitive categories, as displayed in Table 2. The intuitive

[1] http://andrewmolney.name

Table 2: Selected Genres.

Genre 1	Genre 2	Genre 3	Genre 4
Bowling for Columbine	The Mummy Returns	Spirit: Stallion of the Cimarron	My Big Fat Greek Wedding
Fahrenheit 9/11	Bad Boys II	Brother Bear	Sweet Home Alabama
Whale Rider	Face/Off	Treasure Planet	How to Lose a Guy in 10 Days
Super Size Me	Behind Enemy Lines	The Lion King 1 1/2	Pretty Woman
Hotel Rwanda	Tomb Raider	Stuart Little 2	Legally Blonde
Maria Full of Grace	The Fast and the Furious	Garfield: The Movie	Two Weeks Notice
City of God	Rush Hour 2	Spy Kids 2	When Harry Met Sally
The Motorcycle Diaries	Gone in 60 Seconds	Home on the Range	Bridget Jones's Diary
Spellbound	XXX: Special Edition	Scooby-Doo 2	13 Going on 30
Rabbit-Proof Fence	The Mummy	SpongeBob SquarePants	The Wedding Planner

appeal of these genres is consistent with word-based topics presented in the topic model literature (Steyvers & Griffiths, 2007). Each genre list is rank ordered by probabilistic membership. Therefore the first ranked film in each genre is the most probable film given that genre, and so on. This ranking is derived from the ϕ matrix of the topic model (Steyvers & Griffiths, 2007).

Consistencies in Table 2 are evident. For example, Genre 1 could be considered documentaries or biographically inspired independent films. Genre 2 consists of action films that veer towards the fantastic. Genre 3 is made up of animated films directed at children. And Genre 4 lists romantic comedies. However, inconsistencies are also apparent. For example is "Bad Boys II" really as fantastic as a film about mummies? Or are Michael Moore films really that much like "Whale Rider"? Under this critical view, what can be gleaned from Table 2 is somewhat mixed. On the one hand, it is clear that some sense of genre can be driven by likability ratings alone. On the other, it is unclear to what extent these ratings-driven genres correspond to typical film genres. Without a correspondence-based evaluation, it is unclear whether the genres in Table 2 represent strong coherent categories or an observer bias towards any category that might make them coherent.

Study 2: Correspondence-based Evaluation

Method

To carry out a correspondence-based evaluation of our model, it is necessary to find a large existing dataset with human annotated genres for each movie. Fortunately such a dataset exists and is freely available: the Internet Movie Database (IMDB). IMDB contains an enormous amount of information for a given film, ranging from the director and year of release to less commonly known information such as the art department. Including amongst the hundreds of pieces of information associated with each movie is a set of 28 genres, listed in Table 3.

Each film in IMDB is associated with one or more of the genres in Table 3. For example, the biopic, "Ray," based on the story of Ray Charles, is labeled with Biography, Drama, and Music. How these genre labels were generated for IMDB is not clear, and interrater reliability for these genres is not available. The task of correspondence is then to match up every film in the Netflix dataset (which contains all the likability ratings) with the genres in the IMDB dataset. Unfortunately, this is less straightforward than it might first appear. The Netflix dataset is intentionally sparse, including only title, year, and ratings for each film.

Table 3: IMDB Genres.

Documentary	Animation	Family	Sport
Crime	Drama	Mystery	Action
Sci-Fi	Comedy	Short	Game-Show
Romance	Fantasy	Adventure	Music
Thriller	Biography	History	Musical
Horror	Adult	War	Film-Noir
Reality-TV	Western	Talk-Show	News

IMDbPy is the Python-based software library used for manipulating the IMDB data (*IMDbPy*, 2010). IMDbPy provides a search capability for querying a particular title. This search capability purposely returns more than single title in order to accommodate alternate title forms. Using IMDbPy, a correspondence requiring an exact match of both year and title yields only 8,283 exact matches out of a possible 17,770. Relaxing the exact match requirement so that years match and titles match up to the colon yields an additional 1,082 matches.

Inspection of the data reveals that failures to match have a variety of reasons. First, typographic conventions differ between datasets, such that a foreign film may have its original title spelling in one dataset and an Anglicized title in another, e.g. "Character" and "Charackter." In addition, year information may be off by one between the two databases. Sequels and series are a particular problem, such that one database may precede the name of an episode with the name of the series, whereas the other does not. Some errors also exist in the matched films. It is possible, though rare, for two films to be released in the same year with the same name. For

example, "Ray," the biopic of Ray Charles, appeared in the same year as a genre short of the same name. Finally, because to the inconsistencies with series naming conventions and the partial match strategy described above, some within-genre mismatches can occur, e.g. "Star Trek: Insurrection" and "Star Trek: First Contact." However, the distribution of genres is very similar in both the matched set and the original set, as shown in Table 4. Additionally, the correlation between the proportional distributions for original and matched sets is .978.

Table 4: Proportion of Genres.

Genre	Matched	Original
Action	0.14	0.12
Adult	0	0.02
Adventure	0.04	0.04
Animation	0.04	0.05
Biography	0.03	0.02
Comedy	0.24	0.2
Crime	0.06	0.05
Documentary	0.08	0.1
Drama	0.21	0.19
Family	0.02	0.02
Fantasy	0.01	0.01
Film-Noir	0	0
Game-Show	0	0
History	0	0
Horror	0.05	0.04
Music	0.02	0.02
Musical	0.01	0.01
Mystery	0.01	0.01
News	0	0
None (missing)	0	0.05
Reality-TV	0	0
Romance	0.01	0.01
Sci-Fi	0.01	0.01
Short	0.01	0.03
Sport	0	0
Talk-Show	0	0
Thriller	0.02	0.01
War	0	0
Western	0.01	0.01

Once the 9,249 films were paired, the WEKA toolkit (Hall et al., 2009) was used to build two sets of predictive models. The first set uses as features only the distribution of topics associated with each movie, a row vector. For example, position 1 would be the probability that a movie belongs in genre 1, position 2 to probability a movie belongs in genre 2, and so on for all 50 genres. The second set of models uses as features a collection of information from IMDB, chosen to best match the features sometimes used by film critics to determine the genre of a film, as described in Table 1. These features are listed in Table 5.

Table 5: IMDB Features.

Feature	Type
Plot	NUMERIC
Title	NUMERIC
Actor1	NOMINAL
Actor2	NOMINAL
Director	NOMINAL
Year	NUMERIC
MPAA	NOMINAL
Genre	NOMINAL

A few features of Table 5 warrant brief remarks. Plot is a plot synopsis of the film. The two actor features are the first and second named actors on the billing, i.e. the stars of the film. MPAA is the rating of the film, e.g. PG-13. The other features are self-explanatory.

Some of these features are nominal, such as actor and director names, meaning that they are associated with a fixed set of labels as is genre in Table 3. However, the IMDB plot synopsis is an arbitrary string of considerable length, e.g. 500 words, and the title is a shorter but equally arbitrary string. In order to be usable features that two films could have in common, both plot and title were transformed using term frequency/inverse document frequency such that each word in the string became its own feature. This large set of features was considerably pruned using stop words and stemming, so that only 1,420 features remained. The WEKA command line used to convert plot and title to these numeric features was "StringToWordVector -R1,2 -W100 -prune-rate-1.0 -C -T -I -N0 -L -S -SnowballStemmer -M1 -WordTokenizer".

In both the first and second sets, the genre class to be predicted is the first genre listed by IMDB. This restriction is due to WEKA's inability to perform multi-class classifications, and implies that overall performance of the models is significantly lower than would be the case if any genre label associated with a movie was permitted as a correct answer.

The two differing data formats is what separates the first and second sets of models. Within each set, the same machine learning algorithms were used to predict genre. These include the following five models. First, ZeroR, which predicts the most prevalent class, e.g. Comedy. Secondly, NaiveBayes, which assumes features are independent and uses Bayes Rule to construct a classifier. Thirdly, AdaBoostM1 uses an ensemble of weak learners, in this case a decision stump, using the boosting approach (Schapire, 2003). Fourthly, J48 is a decision tree whose internal branching on attribute values is constructed to maximally discriminate amongst the training data. And finally, Ibk is an instance/prototype based classifier, i.e. k nearest neighbors where k has been set to 10 neighbors. These five algorithms were selected because they represent a cross section of the most widespread and effective machine learning techniques (Wu et al., 2007).

Each model was trained using 10 fold cross validation in

which the dataset is divided into ten bins, and the model trained 10 times, using a different bin as test data each time. Significant differences were measured using a paired samples t-test, p = .05, corrected for the variability introduced by cross validation (Nadeau & Bengio, 2003).

Results

The results of the predictive models are displayed in Table 6. Numbers shown indicate percent correct, aggregated across all genre categories. All significant differences are relative to the ZeroR model for each set.

Table 6: Results in Percent Correct.

Model	Likability Based	Content Based
rules.ZeroR	23.51	23.51
bayes.NaiveBayes	9.94	27.12
meta.AdaBoostM1	23.96	23.51
trees.J48	37.30	29.21
lazy.IBk	41.22	27.50

Interestingly there is a fair distribution of performance across all models for the first set (likability-based genres). The worst performer is NaiveBayes, worse than the ZeroR model, while the best performer is IBk-10, at 41%. All differences in this first set are significant.

Performance on the second set of models is worse than the performance on the first set. There is very little deviation away from ZeroR. All differences are significant, except AdaBoostM1, which is not significantly different from ZeroR. The best model of the second set, J48, has only 29% accuracy compared to 41% for IBk in the first set. This performance is particularly poor considering the base rate (ZeroR) is 23%.

Two important points are clear from this data. The first is that the likability-based genres are indeed strong and coherent, predicting the correct human annotated label in 41% of cases. The second is that the likability-based features are more successful at predicting the human annotated label than are the content-based features.

Discussion

Perhaps the most significant finding of both studies is that genres can be extracted from just ratings. Although the percent accuracy using just ratings is 41%, that is still a large figure given two observations. The first is that the 41% performance is based on a single genre classification, when IMDB allows multiple classifications. So 41% performance represents the lowest, most conservative figure. The second observation is that the likability-based performance is considerably higher than the content-based performance at 29%. This difference suggests that likability-based genre classification is a more accurate model of how humans classify film genres than is content-based classification.

The topic model we use makes very few assumptions, and yet the assumptions it does make are quite strong. The basic premise of the model is that people are a mixture of genres. These genres, in turn, generate the ratings observed. To claim that people are a mixture of genres, when genres are typically considered to be a property of artifacts, is a strong and radical claim. The results of the two studies presented above not only support this claim but also suggest that it should be taken seriously as a new approach to genre.

Suppose that likability-based genres are taken seriously. Are they useful, particularly in regard to existing genre studies? The current focus on film suggests that they are. Recall the complementary problems of genre definition and analysis discussed in the introduction. Using likability-based genres as a framework, these can be addressed straightforwardly.

As before, the problems of definition include circularity and the monolithic assumption (Stam, 2000). The basic problem of circularity lies in a supervised approach in which a critic tries to align film features with a given genre category. A likability-based model, as an unsupervised model, avoids this problem entirely because their is no initial assumption of genre used to define the features of genre. Instead, genre emerges from genre-agnostic likability ratings. The second problem of definition, the monolithic assumption, is addressed by the structure of the topic model. Under this model, every movie has some probability of membership in every genre. Study 2 above illustrates that it is not necessary to pigeonhole a movie into a genre in order to create meaningful genres: even using a probabilistic definition of genre, one can still approximate the monolithic assumption to 41% accuracy. Pluralistic genres, like "romantic comedy," are not a special case but are represented in the same way as any other genre.

Using the likability-based definition of genre, we can also clarify problems of analysis that have been raised (Stam, 2000). First, are genres real or imagined? According to our approach, genres are only manifested through people's preferences. Therefore they do not have any status in the world except as a consensus of preferences across large groups of people. On whether the number of genre categories finite or infinite, the structure of the topic model suggests that the number of genres is completely arbitrary, and is controllable using the parameter T, the number of topics. This suggests that likability-based genres are potentially infinite. Third, on whether genres are timeless or are trendy, the likability-based model suggests that they are trendy. Any new ratings that are assimilated into the model can change the resulting genres. As long as the people making the new ratings represent a new mixture of genres, the genres will shift towards the trendy. Finally, as to whether the genres are universal or culturebound, one can speculate that they are culturebound to the extent that one culture may rate movies consistently differently from another culture. This is intuitively plausible, e.g. Bollywood movies rated in India vs. the United States, and may be accounted for in the same way as the timeless or trendy prob-

lem.

Likability-based genres also extend beyond the traditional conceptualization of genre and correspond to the notion of intertextuality. In film, intertextuality has be described as having several properties (Stam, 2000). The first overarching property is that every film is necessarily related to every other film. Second, intertextuality is an active process, so rather than "belonging" to a genre, a film dynamically relates to other films. Finally, intertextuality involves not only all other films, but potentially other arts and media. Clearly the likability-based model corresponds to each of these three properties, by being based on the connectivity amongst all movies via ratings, using an active data-driven model, and using the abstract notion of rating, which can be applied to heterogeneous items like film, music, and books simultaneously. Thus the likability-based model can apply to modern intertextual theories of media in addition to traditional notions of genre.

In summary, likability-based genres offer a novel and useful way of considering genre: people are a mixture of genres. Likability-based genres can predict a significant percentage of genres in the Netflix dataset. Moreover, likability-based genres can also be used to address fundamental problems of definition and analysis in film theory. Likability-based genres can also be extended to broader frameworks than genre, such as intertextuality. However, likability-based genres as described in this paper do not represent a complete theory. In order to understand this phenomenon fully, it is necessary to understand how the ratings themselves are generated as well as how likability-based genres manifest in other contexts.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education,through Grant R305A080594 and by the National Science Foundation, through Grant BCS-0826825, to the University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the National Science Foundation.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Chandler, D. (1997). *An introduction to genre theory.* Available from http://www.aber.ac.uk/media/Documents/intgenre /chandler_genre_theory.pdf

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). New York: Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386). Lawrence Erlbaum Associates.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211–244.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18.

Imdbpy. (2010, February). Available from http://imdbpy.sourceforge.net/index.php?page=main

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*(2&3), 259-284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Mach. Learn.*, *52*(3), 239–281.

The netflix prize rules. (2010, February). Available from http://www.netflixprize.com/rules

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Commun. ACM*, *40*(3), 56–58.

Rubin, T., & Steyvers, M. (2009). A topic model for movie choices and ratings. In *Proceedings of the ninth international conference on cognitive modeling.* Manchester, UK.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification* (Vol. 171, pp. 149–172). New York: Springer Verlag.

Stam, R. (2000). *Film theory: an introduction*. Malden, Mass.: Wiley-Blackwell.

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 424–440). Lawrence Erlbaum.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, *14*(1), 1–37.