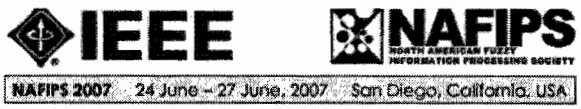


Copyright and Disclaimer

IEEE Catalog Number: 07TH8957C
ISBN: 1-4244-1214-5
Library of Congress: 2007924575

Visualization of Item Features, Customer Preference and Associated Uncertainty using Fuzzy Sets

Azene Zenebe
Department of Management Information Systems
Bowie State University
Bowie, MD 20715, USA
azenebe@bowiestate.edu

Anthony F. Norcio
Department of Information Systems
University of Maryland (UMBC)
Baltimore, MD 21250
norcio@umbc.edu

Abstract - Some of the requirements during preferences discovery or preference modeling through machine learning and data mining are: (i) understanding of features of items, e.g. genres content of a movie; (ii) understanding of patterns in customer feedback on items to explore and identify customer preferences; (iii) understanding of discovered patterns in customer preference on features of items, e.g. preference to genres of movies; and (iv) understanding of similarity among customers' preference, e.g. to form similar cluster of customers in their genre preference. An attempt is made to satisfy these requirements using fuzzy set driven information visualization technique; and movie as the item and genre as the feature are used for illustration. Visualization of features of items, patterns of customer previous feedback to these items, and the relationship between the feedback and item features along with associated measure of uncertainty are presented. The uncertainty is non-stochastic type that is induced from subjectivity, vagueness and imprecision in item features and user preference; and it is modeled using fuzzy set. The visualization of the discovered preference along various demographic features of the users is also presented. This in turn can help forming clusters of users with similar preference to various kinds of items.

I. Introduction

We propose an approach to model the non-stochastic uncertainty in item features, and customer features and preferences during preference modeling for recommender systems. The proposed approach models non-stochastic uncertainty that is induced from subjectivity, vagueness and imprecision and successfully modeled using fuzzy set. In relation to items, the uncertainty is associated to what extent (e.g. example low, medium or high) the items have some features. For instance, given a movie, to what extent the movie has drama content or is highly drama? In relation to preference, the uncertainty is associated to what extent a customer likes, dislikes, or be indifferent to an item based upon features of an item. More information on the approach can be found in [1, 2].

The use of a fuzzy set and visualization enable the development and implementation of algorithm for automatic discovery of customer preferences from data with non-stochastic uncertainty. Knowledge discovery that incorporates uncertainty is important because it puts the mining process in a more realistic settings [3].

Information visualization helps see patterns in item features and user preference as well as helps in dealing with the high volumes of data [4]. During discovery of customer preferences, understanding features of items, customer feedbacks and the relationship between the two are essential. Visualization of initial data source provides insight in the discovery process. Moreover, visualization of the discovered pattern and knowledge increases the comprehension and acceptance of the discovered relationships.

After conducting analysis of items recommendation process and item knowledge, the major requirements during preferences discovery are: (i) understanding of features of items, e.g. genres content of a movie; (ii) understanding of patterns in customer feedback on items to explore and identify customer movie preferences; (iii) understanding of discovered patterns in customer preferences on features of items, e.g. preferences to genres of movies; and (iv) understanding of similarity among customers' preference, e.g. to form similar cluster of customers in their genre preferences. These requirements are satisfied using visualization techniques. In our research and rest of the paper, we use movie as the item and genre as the feature for illustration.

This paper shows how fuzzy set driven visualization of uncertainty helps customer preference modeling (i.e. data mining) processes. The support gained from the visualization includes: understanding of features of items; understanding of patterns in customer feedback on items to explore and identify customer preferences; understanding of discovered patterns in customer preferences on features of items, e.g. preferences to genres of movies; and understanding of similarity among customers' preference to various kind of items.

This paper is organized as follows. Section II presents the modeling of non-stochastic uncertainty in items recommender systems using fuzzy set. Section III presents the dataset and preprocessing done on movies. Section IV presents the results of the visualization. Finally, conclusion and future research directions are presented in Section V.

II The Representation Method

For an item described with multiple attributes, more than one attribute can be used for a recommendation. Moreover, some attributes can be multi-valued involving overlapping or non-mutually exclusive possible values. For example, movies are multi-genres and multi-actors [5]. These values of multi-valued attributes in an item can be represented more accurately within a fuzzy set using membership function than within a crisp set.

A membership function in fuzzy set theory is deliberately designed to treat the vagueness and imprecision in the context of the application [6]. The type of function that is suitable depends on the application context, and in certain cases the meaning captured by fuzzy sets is not too sensitive to the variations in the shape [7]. In practice, triangular, trapezoid, Gaussian, S-function, and exponential-like function are the most commonly used membership functions. Moreover, in practice, suitable membership function's shape is assumed a priori and its parameters are determined by domain expert or using machine learning techniques[7]. The membership function (1) is developed using domain knowledge and heuristic described next; however other membership with similar characteristics as (1) can also be considered.

Let an item I_j ($j = 1 \ldots M$) be defined in the space of an attribute $X = \{x_1, x_2, x_3, \ldots. x_L\}$, then I_j can take multiple values such as $x_1, x_2, \ldots$, and x_L. The membership function of item I_j to value x_k ($k = 1 \ldots L$), denoted by $\mu_{x_k}(I_j)$, need to be obtained. Hence, a vector $X_j = \{(x_k, \mu_{x_k}(I_j)), k = 1 \ldots L\}$ is formed for I_j, where $\mu_{x_k}(I_j)$ can be interpreted as the degree of similarity of I_j to a hypothetical (or prototype) pure x_k type of the item; or as the degree of presence of value x_k in item I_j.

We use movie as the domain and movie genre as the attribute to make the proposed method operational and apply the heuristic. According to Cooper-Martin [8] in a movie marketing application, most movies are selected for pleasure and expenditures of time. And users choose movies based on what they like and enjoy. Furthermore, users use subjective features of movies such as "funny", "romantic" and "scary" (all are a kind of movie genres) to select movies more than objective features such as the director, theatre location and admission price which are useful but are less important.

The reasons why genre is considered as the major attribute for the representation of movies are: movie genres describe the content of movies, and movies are multi-genres [5]; and analysis of descriptions of main film genres shows that genre g_1 of movies (e.g. action) and genre g_2 of movies (e.g. adventure) are overlapping in terms of their subject matter and other movie's attributes[5]. Based on the result of the findings in [8], movies highly liked by users can be grouped into similar categories by subjective features of movies such as genre and MPAA rating.

Given the definition of a movie in the space of genre (G), a movie can have one major genre denoted by x_1 and one or more minor genres x_2, x_3, and so on, in the decreasing order of their degrees of presence in a movie. The degree of membership of movie I_j ($j = 1 \ldots M$) to genre x_k ($k = 1 \ldots N$) is denoted by $\mu_{x_k}(I_j)$. Hence, for I_j, we can form a vector $G_j = \{(x_k, \mu_{x_k}(I_j)), k = 1 \ldots N\}$.

Different approaches have been used to measure genre content of a movie, e.g. [9]. We followed an approach based on fuzzy set theory. To determine the degree of genre presence in movies, we use the genre rank orders available and take the following heuristic approach in the absence of other better methods of determining the genre content of a movie (for example automatic content analysis for determination of genre contents of a movie):

Step 1: Sort x_k in descending order of $\mu_{x_k}(I_j)$. In IMDB (www.imdb.com*) the genres of movie I_j are presented in the order of significance. For example, movie 'King Kong (2005)' has Action as a major genre, and Adventure as the 1st minor, Drama the 2nd minor, Fantasy as the 3rd minor, and Thriller as 4th minor genres.

Step 2: Assign higher degrees of membership value to more important genres of a movie. For instance,

If I_j has only one genre, then $\mu_{x_k}(I_j) = 1$ and $\mu_{x_k}(I_j) = 0$ for all k=2 to N.

If I_j has two genres, then $\mu_{x_k}(I_j) = 1$, $\mu_{x_k}(I_j) = 0.7$ and $\mu_{x_k}(I_j) = 0$ for all k=3 to N.

If I_j has three genres, then $\mu_{x_k}(I_j) = 1.0$ and $\mu_{x_k}(I_j) = 0.50$, $\mu_{x_k}(I_j) = 0.20$ and $\mu_{x_k}(I_j) = 0$ for all k=4 to N; and so on.

Based on the heuristics illustrated for a movie, the possibility for item I_j to take different values of X varies, and the membership function should meet the following four criteria: 1) assigning higher degree of membership to major values than minor values; 2) assigning 0 to values that are not associated with the item; 3) degrees of membership should be normalized to the range of [0,1]; and 4) the same value of X at same rank positions between different items should have varying degrees of membership values if the number of values of X associated with the items are different. We represent this type of heuristic with a Gaussian-like fuzzy set membership function, as shown in (1).

* The IMDB is a large database consisting of comprehensive information about past, present and upcoming movies. "IMDb History," vol. 2004: Internet Movie Database Inc., n.d. accessed on May 2006 from www.imdb.com

$$\mu_{x_k}(I_j) = r_k / 2^{\sqrt{\alpha^*|L_j|(r_k-1)}} \quad (1)$$

where N=|L_j| is the number of values of X associated with I_j and r_k ($1 \le r_k \le |L_j|$) is the rank position of value x_k, and $\alpha > 1$ is a parameter used as a threshold to control the difference between consecutive values of X in I_j. Moreover, α is the only parameter that needs to be determined.

For example, with α set to 1.2 (after various trails), movies M_1=Copycat 1995: Crime/ Mystery/Thriller/Drama, and M_2=Grudge 2004: Horror/Thriller/Mystery with their crisp (I_1 and I_2) and fuzzy (G_1 and G_2) representations are shown in Table I. In crisp representation, all genres that exist in a movie have equal degree of presence or content. However, in fuzzy set based representation, genres that exist in a movie have different degree of presence or content.

TABLE I
MOVIES REPRESENTATION IN SPACE OF GENRES

	Crime	Horror	Mystery	Thriller	Drama
I_1	1	0	1	1	1
G_1	1	0	0.44	0.35	0.29
M_2	0	1	1	1	0
G_2	0	1	0.41	0.47	0

The presented heuristic that leads to the membership function in (1) is developed based on the analysis of the movie dataset, literature on movies[8] and trials conducted on α. This heuristic, for instance, assumes that two genres will not have equal degree of presence in a movie. This assumption is logical because a movie cannot have exactly the same "content" of two or more genres. In future research, studying various membership functions and finding optimal α through evolutionary computing is needed. However, an ideal solution would be to find the degree of presence of each genre in a movie by analyzing the content of the movie using automatic content analysis technologies yet to be well developed and available. We strongly think the representation scheme is not sensitive to variation in membership functions provided that the functions satisfy the real properties of the item under consideration, e.g. for a movie, the distribution of genres in a movie satisfies the four criteria established from the domain analysis.

The representation scheme can be extended to recommender systems based on a combination of multiple attributes. For example, one can use movie genre describing the content of movies as the first attribute and actresses/actors as the second attribute. The actors in a movie can be represented in a vector A={a_1, a_2, … a_k} for K actors. The degree of role or importance of an actor or actress a_k in a movie m_i can be represented by degree of membership associated with the fuzzy variable 'degree of role or importance'. That is, A_j ={(a_k, $\mu_{a_k}(m_j)$), for k=1 to K}, where $\mu_{a_k}(m_j)$ can be determine heuristically. Furthermore, the representation scheme presented for a movie can be generalized and applied to any item with similar characteristics as the movie. A few examples are Music, TV shows, Restaurants and Books.

III. Movie and Customer Feedback Dataset and Pre-processing

The benchmark dataset from MovieLens at GroupLens research project of the University of Minnesota (movielens.umn.edu) along with additional data extracted from the Internet Movie Database (www.imdb.com) is used in this study. The dataset includes movie attributes, customer ratings, and customer demographic information. It consists of 100,000 ratings (1-5) from 943 customers on 1682 movies; and each customer has rated at least 20 movies. In the dataset, movies are described with: movie id, movie title, release date, video release date, IMDb URL, and 20 genres including action, adventure, animation, children's, comedy, crime, documentary, drama, fantasy, film-noir, horror, musical, mystery, romance, sci-fi, thriller, war, western, family, and others.

Genres in the MovieLens dataset are represented with binary values, which do not reflect the true content of movies in the genre space. Therefore, we use the representation scheme, described in Section II, by incorporating information about movie genres from the Internet Movie Data Base. For example for a user 5, Table II shows the representation of some of the rated movies.

TABLE II
MEMBERSHIP DEGREE OF MOVIE TO GENRES RATED BY A USER 5

	G_j (vector for j^{th} movie)						
Movie (I_j)	Rating	Drama (x_{j1})	Comedy (x_{j2})	Action (x_{j3})	…	xj_{19}	x_{j20}
56	4	0.68	1.00	0.44	…	0.00	0.00
79	5	1.00	0.00	0.00	…	0.47	0.00
89	3	0.68	0.00	0.00	…	0.00	1.00
..	..	..	..		…	..	..
254	2	0.44	0.00	1.00	…	0.00	0.00

IV. Visualization Of Item Feature and Customer Preferences

A. Visualization of Distribution of Features of Item: A Case of Genres in a Movie

Movie genres describe the content of movies, and movies are multi-genres [5]. An analysis of the descriptions of the main film genres shows that movies of genre g_1 (e.g. action) and movies of genre g_2 (e.g. adventure) share common subject matter and other movie's attributes[10]. Hence, it is sometimes difficult to judge whether a movie belongs completely to a specific genre or not. As a result, it induces uncertainty in the determination of the genres distribution of a movie. Fuzzy set allows us to represent this type of uncertainty in data [11]. Fig. 1 presents visualization of 1683 movies' content by 20 genres using the crisp set by computing

percentage and the fuzzy set by computing average membership function as degree of presence. It shows that there is disparity in the distribution of genres content of movies in the two representation schemes. Therefore, the fuzzy theoretic based representation results in different distribution from that of the crisp set representation.

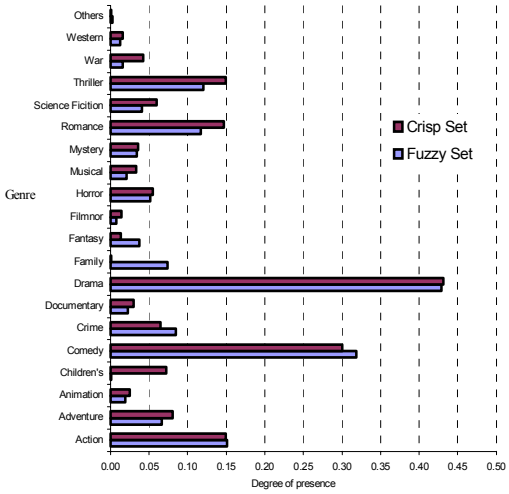
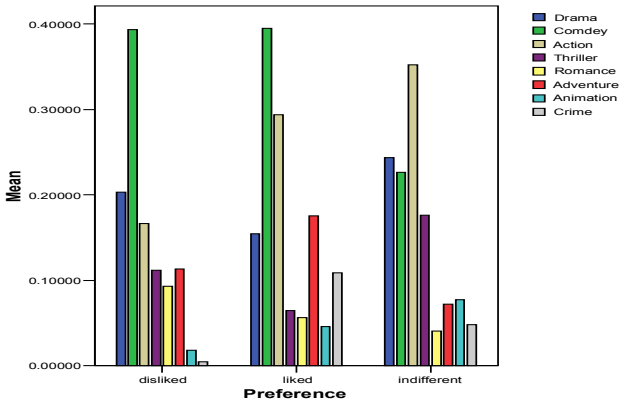


Fig. 1 Visualization of movies' genres content - distribution in 1683 movies

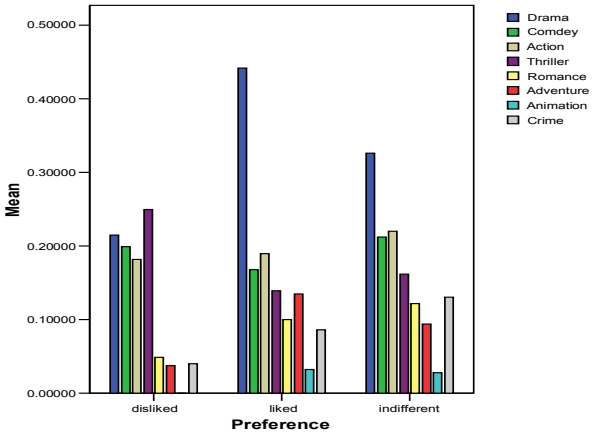
B. Visualization of patterns in customer feedback: A case of customer's movie ratings

In order to understand and explore customer genre preference, based on customer's ratings, movies are categorized into three groups: disliked (NI) with ratings of 1 and 2, liked (PI) with ratings of 4 and 5, and indifferent (II) with rating of 3. Visualization of customer ratings and mean degree of memberships of movies to various genres for selected customers are computed and presented in Fig. 2.

Fig. 2 indicates the disparity in the degree of preference to the different genres for customer 5 and customer 7. For instance, customer 7 likes drama with 0.45, dislikes drama with 0.21, and indifferent to drama with 0.32 degree of membership. Using the maximum fuzzy set operator, customer 7 favors drama movies and dislikes thriller movies. Likewise customer 5 favors action and adventure movies and disfavors drama movies. The examples support the soundness of preference mining based on item features. Therefore, visualization of these patterns in other customers' feedback on items supports the assertion: given the vector of movie genres, it is possible to sort them into preferred (PX), non-preferred (NX), and indifferent (IX) and Unknown (UX) for a customer based on the customer previous movie ratings. This assertion is also supported by the literature in movie marketing [8].



(a) Customer 5: <age=33, Female, Nurse>



(b) Customer 7: <age=57, Male, Administrator>

Fig. 2 Genre preferences distribution for (a) customer 5 and (b) customer 7

As results, algorithm to determine the degree of preference to the various genres in the three categories of preference preferred (PX), non-preferred (NX), and indifferent (IX) is developed. The algorithm works as follow. First the movies rated by a customer are segmented into the three groups. Second the genres composition of each segment is analyzed to determine the genre preferences of a customer. Liked, Disliked, Indifferent, or Unknown is the possible categories of preference, and these results are stored in a vector.

An illustration of how the algorithm works is presented as follows. For user 5 and genres x_1='Drama' and x_3='Action', the algorithm groups the 134 movies rated by user 5 into the following categories: (i) NI and the mean degrees of membership of these movies to x_1 and x_3 are 0.203 and 0.167 respectively; (ii) PI and the mean degree of membership of these movies to x_1 and x_3 are 0.139 and 0.309 respectively; and (iii) II and the mean degree of membership of these movies to x_1 and x_3 are 0.242 and 0.358, respectively. For user 5, execution of the preference modeling algorithm produces vectors consisting of mean degrees of membership of each genre to PX, NX, IX, and UX denoted as (genre(x_k),

$\mu_{PX}(x_k), \mu_{NX}(x_k), \mu_{IX}(x_k), \mu_{UX}(x_k)$):
(Drama, 0.139, 0.203, 0.242, 0); (Comedy, 0.389, 0.393, 0.237, 0); (Action, 0.309, 0.167, 0.358, 0); (Thriller, 0.067, 0.112, 0.174, 0); (Romance, 0.054, 0.093, 0.038, 0); (Adventure, 0.174, 0.113, 0.078, 0); (Animation, 0.066, 0.018, 0.072, 0); (Children's, 0, 0, 0, 1); (Crime, 0.102, 0.005, 0.055, 0); (Documentary, 0, 0, 0, 1); (Fantasy, 0.100, 0.070, 0.042, 0); (Film-nor, 0, 0, 0, 1); (Horror, 0.098, 0.198, 0.087, 0); (Musical, 0.036, 0.015, 0.057, 0); (Mystery, 0.016, 0.046, 0.010, 0); (Science Fiction, 0.231, 0.068, 0.099, 0); (War, 0, 0.023, 0, 0); (Western, 0.024, 0.032, 0, 0); and (Family, 0.077, 0.171, 0.221, 0).

For de-fuzzification, maximum inference operator can be used, and the results are stored in vectors of liked, disliked, indifferent and unknown or indeterminist genres along with degree of preferences. For user 5, the inferred ordered lists of genre preferences are:

- PX={(Science Fiction, 0.231), (Adventure, 0.174), (Crime, 0.102), (Fantasy, 0.100)},
- NX={(Comedy, 0.393), (Horror, 0.198), (Romance, 093), (Mystery, 0.046), (Western, 0.032), (War, 0.023)},
- IX={(Action, 0.358), (Drama, 0.242), (Family, 0.221), (Thriller, 0.174), (Animation, 0.072), (Musical, 0.057)}, and
- UX={(Children's, 1), (Documentary, 1), (Film-nor, 1), (Others, 1)}.

C. Visualization of inferred customer preference: A case of inferred genres preference

Visualization of inferred customer preference is useful to understand the results as well as to identify similarity among customers in their preferences. Fig. 3 presents preference of customers for various genres along with measure of uncertainty - membership degree as measure of degree of preference. These graphs show the variations in the inferred preference to genres among customers for the various genres. Some genres such as drama, comedy and actions are liked more than others such as Horror, Family and Musicals. Moreover, Fig. 4 shows the clusters of users who liked drama with various degrees – very low, low, medium, high, and very high.

Knowledge of patterns among customers' preference is important in recommender system, e.g. to form similar cluster of customers with similar genre preferences. Two users are similar if and only if they are similar in the genres they like, dislike, and indifferent to. Fig. 5 presents pattern that exists in customers' preferences to selected preferred genres by age and gender. Disparity among customer preference can be observed across the different ages and genders. Some of the interesting patterns in Fig. 5 are:

- Younger male customers liked action movies with degree of preference approximately between 0.10 and 0.75 that is greater than older male customers with degree of preference approximately between 0.0 and 0.25.
- Female customers have lower preference to Action movies than male customers.
- Younger customers like drama movies with degree of preference approximately between 0.25 and 0.75.
- Younger customers like comedy movies with degree of preference between approximately 0.10 and 0.75 that is greater than older customers with degree of preference approximately between 0.00 and 0.25.

Therefore, clustering using gender, age and genre is useful for movie recommender systems.

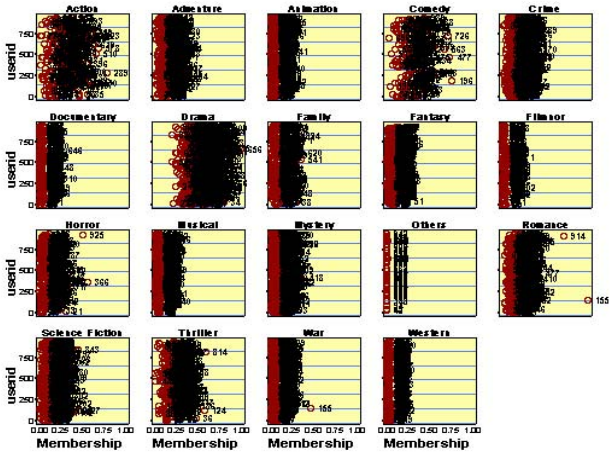


Fig. 3 Degree of membership to liked preference category for all genres

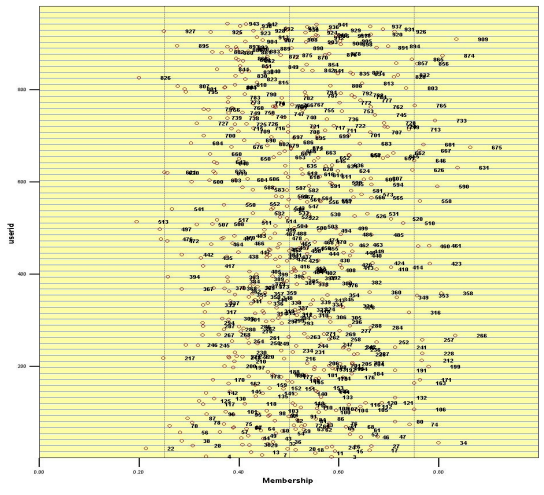
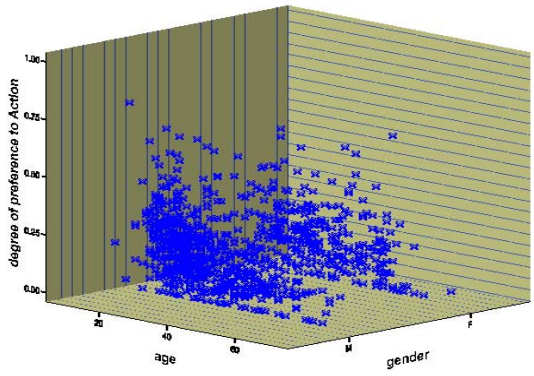
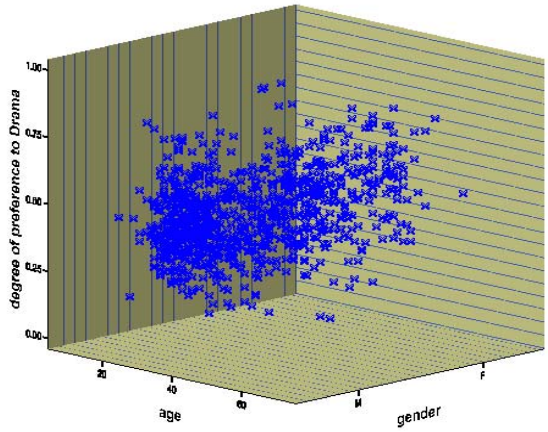


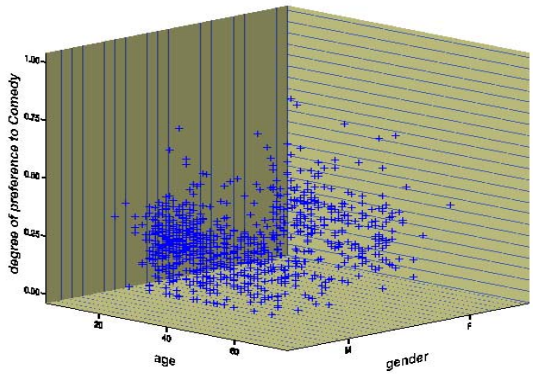
Fig. 4 Degree of membership to liked preference category for drama



(a) Action



(b) Drama



(c) Comedy

Fig. 5 Liked genres by age and gender along with measure of uncertainty for (a) action, (b) drama and (c) comedy

V. CONCLUSION AND FUTURE RESEARCH

The representation of items, e.g. movies in the genre space, using a fuzzy set creates opportunities to study the patterns in item features, e.g. genres of movies, and customer preferences, as well as the associated uncertainty measured by degree of memberships. This paper also shows how fuzzy set driven visualization of uncertainty helps customer preference modelling processes. In particular, visualization helps the understanding of features of items and patterns in customer feedback on items in order to explore and identify customer preferences. It also helps the understanding of discovered pattern in customer preferences on features of items, e.g. preferences to genres of movies; and understanding of similarity among customers' preference.

In this paper, unsophisticated visualization techniques are used. Future research will focus on the use of advanced visualization techniques to explore more the highly complex item features, user characteristics, and user preferences as well as visualization of clusters of user preferences by various demographic features including gender, age and occupation.

REFERENCES

[1] Zenebe and A. F. Norcio, "Uncertainty Identification, Representation and Measurement in User Modeling: A Methodology," Proc. Proceedings of the 14th Information Resources Management Association International Conference, Intelligent Information Systems track, 2004.
[2] Zenebe and A. F. Norcio, "Evaluation Framework for Fuzzy Theoretic-Based Recommender System," Proc. 11th International Conference on Human-Computer Interaction, 2005.
[3] Z. Chen, *Data Mining and Uncertainty Reasoning*. New York: John Wiley & Sons, Inc., 2001.
[4] A. Keim, M. Sips, and M. Ankerst, "Visual data-mining techniques," in *The Visualization Handbook*, C. D. Hansen and C. R. Johnson, Eds. New York: Elsevier, 2005, pp. 831-843.
[5] R. Altman, *Film/Genre*. London: British Film Institute, 1999.
[6] S.-M. Hsu, C. Wu, and T.-W. Tien, "A Fuzzy Mathematical Approach for Measuring Multi-facet Consumer Involvement in the Product Category Evaluation," *Marketing Research On-Line*, vol. 3, pp. 1-19, 1998.
[7] W. Pedrycz and F. Gomide, *An Introduction to Fuzzy Sets*. Cambridge, Massachusetts: The MIT Press, 1998.
[8] Cooper-Martin, "Consumers and Movies: Some Findings on Experiential Products," *Advances in Consumer Research*, vol. 18, pp. 372-378, 1991.
[9] J. A. Walter and H. Ritter, "On Interactive Visualization of High-dimensional Data using the Hyperbolic Plane," Proc. SIGKDD'02, pp. 123-132, 2002.
[10] J. Staiger, "Hybrid or inbred: the purity hypothesis and Hollywood genre history," *Film Criticism.*, vol. 22, no.1, pp. 5-21, 1997.
[11] P. Smets, "Theories of Uncertainty," in *Handbook of Fuzzy Computation*, E. Ruspini, P. P. Bonissone, and W. Pedrycz, Eds. Philadephia, PA: Institute of Physics Publishing, 1999.