Generative user models for Adaptive Information Retrieval

Yoichi Motomura
Electrotechnical Lab.
Tukuba, Ibaraki JAPAN

Kaori Yoshida
Kyushu Institute of Technology
Iizuka, Fukuoka JAPAN

Kazunori Fujimoto
NTT Communication Science Lab.
Seika-cho, Kyoto JAPAN

Abstract

For information retrieval (IR) tasks, user models are used to estimate user's true intention and demand. Unfortunately, most user models are constructed in a specialized form that is not applied to other systems or domains. This specialization makes difficulty in sharing user models as common resources for developing information retrieval systems and for researching cognitive characteristics in various users. In order to solve this problem, we need a general user modeling method. In this paper, a user model based on probabilistic framework is proposed. We call this model as *generative user model*.

The generative user model represents user's mental depth by latent (hidden) variables. It also has visible variables that mean word set and qualifier of each word as a subjective probability distribution. The model can handle uncertainty of user's subjectivity by probabilistic framework. Recent statistical studies for such latent models give learning algorithm. Our generative user model can be constructed from dataset taken by information retrieval tasks. As an example, we also introduce two different kinds of information retrieval systems, ART MUSEUM (Multimedia Database with Sense of Color and Construction upon the Matter of ART) and DSIU (Decision Support for Internet Users). The generative user model is applied to these systems. The properties of the model and interactive learning mechanism are shown.

1 Introduction

Nowadays, many variety of users want to find out their desired information. However, users should be familiar with how to choose suitable keywords to search. Because, knowing true demand of users is not easy task for information retrieval systems. Ambiguity and incompleteness of users' input, and difference of users' tendency cause this problem. Thus, users are forced to adapt to each system uniformly, and users should know which input is suitable for getting their demand on each system.

Such users' inconvenience can be solved by an adaptive user model, which can estimate true demand with taking into account individual character of each user. Many researches have been studied to realize such adaptive user models. Most of them are specialized with specific tasks and they have their own representation. As the result, such models can not be applicable for other tasks. Then knowledge of user models studied in particular domain can not be generalized for much wider purpose any more.

One aim of our research is providing a general framework of adaptive user models for many kind of information retrieval systems. We use general representation with probabilistic framework and word set of natural language for constructing user models. In this paper, we propose the generative user model based on a probabilistic network like Helmholtz machine[2]. This model can handle uncertainty of user's subjectivity and can adapt to each user by statistical learning like the EM algorithm.

Another aim of our direction is evaluation and analysis of user friendly information retrieval systems from user-adaptive point of view. We have developed two different information retrieval systems, ART MUSEUM and DSIU. ART MUSEUM is an interactive database system which treats full color painting and its artistic impression. The system has the mapping between artistic impression words and images. The system was published on the internet (http://www.etl.go.jp/etl/taiwa/ArtMuseum). Recent two years, over six hundred users accessed and retrieved images using impression words on the demonstration page. The DSIU system is a recommendation system for internet users. The DSIU handles information on the internet instead of human consultants and generates advice involving explanations. For instance, a user who wants to buy an electric product like a digital camera, selects query words (excellent image-quality, good portability, and so on). The system then

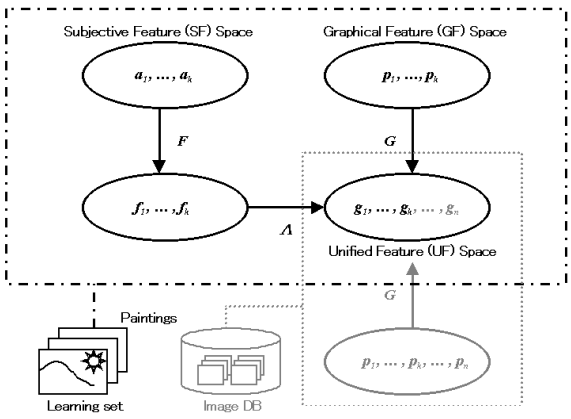

Figure 1: ART MUSEUM [4]

recommends some products and provides logical explanations as the reason of the recommendation.

The generative user model proposed in this paper is applied to these two different domains of ART MUSEUM and DSIU. We also discuss learning method of the generative model as interactive adaptation.

In the section 2, we introduce our information retrieval systems, then re-define with probabilistic interpretation. In the section 3, we propose the generative user model, and an example of interactive adaptation to users. The property of the generative model is also discussed in the section.

2 Information retrieval systems

2.1 ART MUSEUM

ART MUSEUM system is an art gallery which treat full color painting and its artistic impression. The system has the personal index on unified feature (UF) space, which are derived from graphical feature (GF) space on color and construction and subjective feature (SF) space on words which express artistic impression (see Fig. 1). Each function in the Fig. 1 is explained in the following.

The ART MUSEUM has 200 full color paintings in its image database now. We picked 50 paintings as a learning set up from image database by cluster analysis. A user answers his artistic impression on each painting of learning set as the weight vector of adjective words. We can construct an UF space by canonical correlation analysis. It can be considered that neighboring paintings on the UF space give the similar impressions. Besides, the words which represent similar impression distribute neighboring points. Sense retrieval describes the neighboring paintings of a mapped weight vector according to the words presented by a user on UF space. We describe details on each space and the retrieval mechanism below.

2.1.1 Subjective feature space

Let us explain how to select textual domain data in ART MUSEUM system. Usually, art critics view paintings from several viewpoints, such as motif, touch, composition and coloring of paintings. Chijiwa reported that the dominant impression derived from paintings is coloring. This report suggests that there is a reasonable correlation between coloring and impression. Thus, we selected the 30 adjective words which represent the impression from coloring of the paintings. Some users answered their artistic impression on 50 paintings (learning set) using 30 adjective words.

We wanted to decrease the number of adjective, because the query should not give heavy load to users. Therefore, we removed 20 adjective words which were not used frequently and not learned properly. Currently, the ART MUSEUM system uses 10 adjective words as a textual domain data. In the same way, some users answered their artistic impression on 50 paintings (learning set) using mentioned 10 adjective words as the weight of the adjectives $\vec{a}_k$ to each painting $k \in$ the learning set (see Fig. 3).

2.1.2 Graphical feature space

Let us explain about a pictorial domain data. The system has the coloring feature of each paintings in the database. The coloring feature is parameterized by the distribution and autocorrelation of the RGB values. In short, we can parameterize the coloring of a painting as follows.

1. Divide a painting into 32×32 sub-pictures to approximate the combination and the arrangement of colors.
2. Calculate the distribution of the RGB intensity value in the sub-pictures.
3. Calculate the local autocorrelation of RGB intensity as the GF vector $\vec{p}_i$.

The local autocorrelation features are obtained by scanning the image with the local masks and by com-

puting the sums of the products of the corresponding sub-pictures. In a way, the features computed from the image show not only coloring but also something like touch of paintings. Because the features include very local and detail information.

2.1.3 Unified Feature Space

Let us show the algorithm for constructing an UF space. We cannot directly compare the subjective words in a query and the graphical features of a painting, since they are on the different domains.

We may expect that there is a reasonable correlation between the set of words and the parameterized with the graphical feature. The ART MUSEUM system can analyze such correlation between the different domains. We will regard the correlation as the personal view model for the user. We can construct an UF space on this model to compare the subjective words and graphical feature. The algorithm to construct an UF space is represented by following equations.

1. Answer his artistic impression of the paintings (learning set) using a set of impression words.
2. Apply the canonical correlation analysis to the result of the questionnaire by the user. The linear mappings F and G make their correlation maximum: $\vec{f}_k = F\vec{a}_k, \vec{g}_k = G\vec{p}_k$.
3. Calculate the UF vectors of paintings in the database from the following formula, $\vec{g}_i = G\vec{p}_i$.

It can be considered that the neighboring paintings have similar impression and similar graphical feature on UF space. We will refer to the UF space of $\vec{g}_i$ as the personal index of the user model. Note that we do not have to assign the adjectives $\vec{a}_i$ to every painting in the database. Once the system has learned the linear mappings F and G, it can automatically construct the UF space only from the GF vectors. Accordingly, whenever we put new paintings into the existing image database, it is not necessary to construct the UF space.

2.2 Decision Support for Internet Users

DSIU systems handle information on the Internet instead of humans and generate advice involving explanations of the information [6, 7]. As an example to show the DSIU process, we consider a user who wants to buy an electric product, e.g., a digital camera.

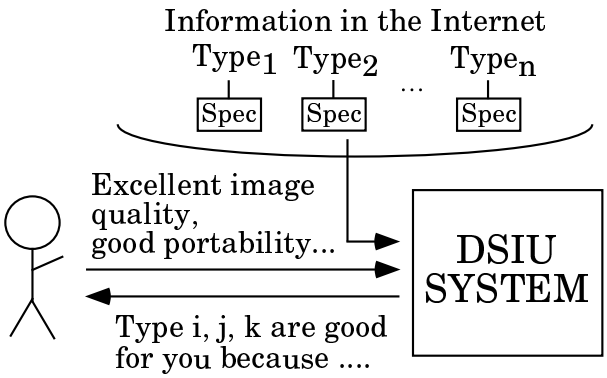

Figure 2: DSIU[6]

First, the user inputs his or her preferences for digital cameras, e.g., excellent image-quality, good portability, and so on. DSIU systems then recommend some digital cameras and provide logical explanations for the recommendation to the user, e.g., CCD pixel sizes are important for image-quality and the camera has 1.41 megapixel CCD, which is comparatively superior specification. The role and importance of CCD are understood by the user while such explanations are provided although the technical term "CCD" may not be familiar to the user initially. As a result, the user can make a decision based on his or her thinking and understanding. Figure 2 shows the information flow of a DSIU system.

2.3 Generalization of IR tasks

In the above subsections, we show two different kinds of information retrieval systems. We give definition for general information retrieval tasks based on keyword set and probability distribution.

At first, let's denote W as a set of keywords $w_1, \cdots w_n$. For each word, we can introduce qualifier $f(w_i)$ that means strength or degree of word w_i. Then, by normalizing $P(w_i) = f(w_i)/\sum_i f(w_i)$, we have probability of word w_i and probabilistic distribution $P(W)$. When we have large amount of documents, the probability $P(w_i)$ is obtained as relative frequency of each word. In this case, this probability becomes objective in the document source. We have opposite interpretation also when the qualifier is given by particular user's subjective impression. In this case, the probability becomes subjective.

As the next step, we expand the probability as conditional one. Let's denote X as the object that is user's

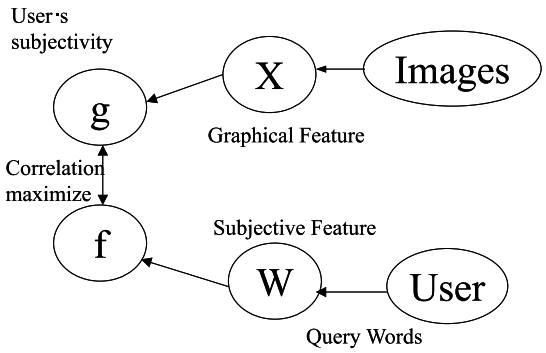

Figure 3: Abstract model of ART MUSEUM

demand. We can introduce a conditional probability distribution $P(W|X)$. This conditional probability distribution can represent both of objective characteristic of X by frequency in documents or user's subjective impression about X from user's questionnaire. Obviously, a user's $P(W|X)$ may not be the same as other users' models.

When a user gives word set W, and qualifier of each word (by symbol like *very, slightly and so on* or numeric value), both ART MUSEUM and DSIU find objects which satisfy the user's demand. Actual retrieval mechanism of these systems are deterministic. ART MUSEUM's one is based on feature extraction of pattern recognition and canonical correlation analysis. DSIU's one is based on grammatical structure and semantics of documents distributed in the internet. However, probabilistic interpretation of these systems is also established from a theoretical point of view. Especially, when it is necessary to evaluate uncertainty of user's feeling and to generate hypothesis according to possibility, probabilistic representation is useful.

Searching process is regarded as finding X maximizing $P(X|W)$ that means likelihood of X given W. Obviously, the distribution $P(X|W)$ should be based on each user's subjective impression. If the IR system finds the result only from uniform knowledge which does not depends on user's subjectivity, the result may not satisfy each user's demand. This problem is analyzed the difference of the system's $P(W|X)$ and user's $P(W|X)$. Then adaptive information retrieval system should have an internal model of the user, $P(W|X)$ and should adapt it to each user.

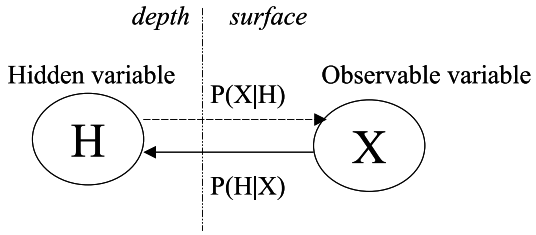

Figure 4: Helmholtz machine

3 Generative user model

In order to make a system much flexible and adaptive for any different type of users, we need adaptive user models. As the first step, we investigate a user model construction method for adaptive information retrieval by analyzing users' words in the particular situation and applying probabilistic models.

In this section, we propose a user model for adaptive information retrieval task as a variation of latent variable models[1]. In particular, we respect for constructing the general user model that can be transfered to many other systems. Thus, our model is based on general representation that does not depends on a specific system and a task.

3.1 Latent variable probabilistic model

User's subjectivity is sometimes affected mental depth of the user. It looks uncertain because of fluctuating of user's feeling and emotion. We can just estimate such mental factors only from obtaining observable information. For such coping with uncertainty, probabilistic models are useful. Especially, probabilistic models that have hidden variables are called latent variable models. Learning method for one of such models (Helmholtz machine) based on the EM algorithm is also studied[2]. The Helmholtz machine consists of two symmetric models. One model which generates observable variables (X) from latent variables (H) is called the generative model, and it is described as $P(X|H)$. Opposite model which outputs latent variables from observable variables is called the recognitive model, and it is described as $P(H|X)$.

The recognition model is learnt from data generated by the generative model, vice versa. This method is called wake-sleep algorithm[2] and similarity with the EM algorithm is reported[3].

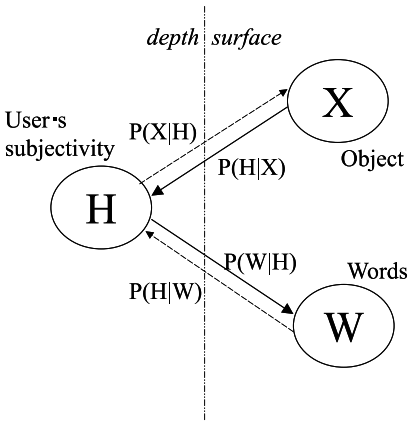

Figure 5: Generative user model

3.2 Generative user model

In this subsection, we introduce probabilistic modeling for information retrieval system. Variables introduced in this paper X and W are observable variables. Moreover, in order to realize the general user model, we introduce unobservable latent variable H that represents users' uncertain mental factors. Users' individual characters and difference of subjective feeling can be represented by H. When a user see or associates a object X, we model the user's recognition process into $P(H|X)$. Uncertainty of H can be evaluated by the entropy of $P(H|X)$. When a user describes W and qualifier of W, these depend on the user's mental state H. Thus, we model the word generating process into $P(W|H)$.

As the result, we can model the word representation when a user looks at a object X by,

$$P(W|X) = \int_H P(W|H)P(H|X)dH. \qquad (1)$$

In this model, there are two parts, $P(H|X)$ and $P(W|X)$, that reflect users' uncertain subjectivity. We define the variable H as a continuous vector. Then user's subjectivity can be compared in an Euclidean space. This is the remarkable advantage of our model. Because users' word impression, which has different dimension, is difficult to compare directly in general.

When we have subjective distribution $P(W|X = x_i)$ for each object x_i, the model $P(W|H)$ and $P(H|X)$ with hidden variable H is constructed by the EM algorithm[8].

3.3 Interactive adaptation

In the beginning, let's see an example of interactive user adaptation scheme in the ART MUSEUM.

For adapting to each user, a mapping function F and G in Fig. 1 should dynamically suit to a specific user's subjective interpretation. The mapping function is corresponding to $P(W|H)$ and $P(H|X)$ in the generative model. The interactive adaptation to each user in the ART MUSEUM is proceeded as the following. (Notation of each variables are shown in Fig.1.)

1. Calculate the average personal view model according to a certain group of users' answer.
2. Retrieve once using the average model.
3. If the retrieval results don't fit his subjective interpretation, the user can give his interpretation on the target painting j using the set of impression words.
4. Interpolate the weight vector $\vec{a}_j$corresponding to the target painting j.
5. Calculate the mapping function F and G by operating using the changed weight vector $\vec{a}_j$.

After adaptation, obtained user's personal view is applied as follows.

1. We can compare the retrieval candidates by same keywords represented by the weight vector $\vec{a}_0$ among the users. Because the system provides the retrieval candidates of each user which have delicate differences. The system's interpretation closes to each of the user's interpretation by learning.
2. We can observe each user's interpretation of any paintings which is supposed by a simulation using each user's personal view model. Because the system infers the suitable impression words for simulating the user's personal view using the inverse mapping function F^{-1} and λ^{-1}, as follows, $\vec{a}_0 = F^{-1}\lambda^{-1}G\vec{p}_0$.

This interactive user-adaptation of deterministic model can be translated as statistical learning of the generative model. The interactive adaptation to each user in the generative model is proceeded as the following.

1. A user gives query as subjective distribution of words, $P(W)$.

2. Retrieve X using a IR system.

3. Evaluate $P(W|X)$ by the generative model (Eq.1).

4. The user gives his impression about X as $P'(W|X)$.

5. Modify the model by the EM algorithm[8].

Instead of the step 2, the reverse probability $P(X|W)$ of the generative model can generate hypothesis X. When $P(X|W)$ is given by modified model $P'(W|X)$, this data generation is equivalent to *boot strapping* studied in statistical researches.

After adaptation, obtained the generative user model can be utilized in the following each case.

1. We have average user model $P(W|H)$ and adapted user model $P'(W|H)$.

2. Under the same expression of a query $P(W)$, the difference of the user's subjectivity can be evaluated by the KL divergence between $P(H|W)P(W)$ and $P'(H|W)P(W)$, where $P(H|W) = P(W|H)P(H)/P(W)$.

3. In order to interpret particular user's expression $P'(W)$ as the common expression of average users, $P(W)$, we can estimate particular user's mental state H' by $P'(H|W)P(W)$. Then we get the common expression by $P(W|H')$.

This words generation ability is prominence in the generative user model.

4 Conclusion

In this paper, we described our information retrieval systems and their tasks. In the tasks, we can introduced by subjective probabilistic distributions to represent relation among input query words, retrieved object and a user's mental state. Then, information retrieval task is generalized from probabilistic point of view. In order to reflect user's subjectivity, we proposed the generative user model that consists the conditional probability between a user's mental depth and impressed words, and the conditional probability between a user's mental depth and an object to retrieve. The generative user model can create possible hypothesis by random sampling according to the distributions. As the model can be evaluated by the interaction with users, interactive adaptation to each user is realized. The generative model is regarded as a kind of Bayesian network for user modeling[9]. We can expand the model by adding more variables and complex structure in our future work.

Acknowledgement

The authors thank to Katsunobu Ito and Isao Hara (Electrotechnical Lab.) for their helpful comments.

References

[1] J.C. Loehlin, "Latent variable models: An introduction to factor, path and structural analysis",Hillsdale, New Jersey Erlbaum,1992.

[2] G.Hinton, P.Dayan, J.Frey and R.Neal, "The "wake-sleep" algorithm for unsupervised neural networks", *Science*, 268, pp. 1158-1161, 1995.

[3] S.Ikeda, S.Amari and H.Nakahara, "Convergence of the wake-sleep algorithm", *Advance in Neural Information Processing Systems*, 11,1999.

[4] Kaori YOSHIDA, Toshikazu KATO and Torao YANARU, "Image Retrieval System based on Subjective Interpretation", em Biomedical Soft Computing and Human Sciences, 4, pp.65–74, 1998.

[5] K.Yoshida, T.Kato and T.Yanaru, "'A Study of Database System with KANSEI Information", *Proc. of IEEE SMC'99*, VI, pp.253-256, 1999.

[6] Kazunori Fujimoto and Kazumitsu Matsuzawa, "Intelligent Systems Using Web-pages As Knowledge Base for Statistical Decision Making", *New Generation Computing*, 17, 4, pp.349-358, 1999.

[7] Kazunori Fujimoto, Kazumitsu Matsuzawa and Toshiro Kita, "Do you think content on the Internet is easy to understand?", *Proceedings of the Tenth Annual Internet Society Conference(INET-2000)*, 2000.

[8] J.Binder, D.Koller, S.Russell, K.Kanazawa, " Adaptive Probabilistic Networks with Hidden Variables.", *Machine Learning* , 29, 213–244, 1997.

[9] F.De Rosis, et.al., "Modeling the User Knowledge by Belief Networks", *Journal of User Models and User-Adapted Interaction*, 2, No.4,pp.367-388, 1992.