Personalize Approach to Question Answering
Question answering personalization is an emergent research area with the intention to make possible for the user to have control over the rendering of answers according to their topic of interest. This chapter presents a personalized approach to question answering based on end user modeling. The personalization of the retrieved data is done using implicit user information and interest area. As the customized data is refined using attributes and values, we implement several similarity metrics. These metrics consider both semantic and syntactic user information. Our evaluation with respect to a baseline QA system gives encouraging result in personalization.

6.1. Introduction

The excessive amount of information available on Web from which relevant documents have to be searched is a key issue to Information retrieval area and thus to its sub discipline of Question Answering (QA) which is designed to find concise answers in response to the user’s question. Refining such a large amount of information for answer extraction becomes a cumbersome task given the representation of data that does not meet user’s aim and requirement. To satisfy user’s requirement, personalization is appropriate remedy to design an adaptive and in some manner intelligent system to improve its usability.

The scope of personalization has been so far not much investigated in QA systems. However, some of the initial efforts were made by Maybury et al. (2002) and TREC (Voorhees, 2003) to present roadmap for personalized question answering in open domain. In general, personalization to QA system can be viewed as User Modeling problem which analyses user’s characteristics and preferences to re-rank the extracted answer list.

The concept of personalization has been used in variety of context in Web applications such as personalized information retrieval (Teevan et al., 2005; Pitkow et al., 2002), recommender systems (Ardissono et al., 2001; Magnini et al., 2001, Miller, 2003) and learning environments (Person et al., 2000; Romero et al., 2003; Linton et al., 2003). Although such applications have few common information processing mechanisms with Question Answering, but a small number of efforts has been inherently made so far
in context of personalized QA systems. Therefore, we first try to explore the model of personalization in broader context of information retrieval and then will focus on significant work that has been so far carried out with respect to the personalized QA.

Most of the personalization approaches primarily rely on building user’s profile which aims to determine user interest and enhance the relevancy of data accordingly. The information for constructing a user profile can be collected explicitly or implicitly. Explicit information is provided directly by user while implicit information can be inferred while observing user’s activity log (Baeza-Yates et al., 2011). Therefore, the personalization techniques can be sub divided into two complementary mechanisms which are (1) the user information gathering, used to identify the user’s inclination and (2) the implication of the collected data to suggest the closest content to the user interest. In the former case, user profile can be used to enhance question representation and filter results at the user interaction phase (Koutrika et al., 2005) while in later case essential information is inferred from user’s previous transaction extracted from system log file. However, the construction of user profiles can be dynamic or static in nature subject to the adapting user changing interest or based on pre stored static information respectively (Teevan et al., 2005). Additionally, short-term and long-term interests might be distinguished in user profiles when taking time into consideration as presented by Kim et al. (2003), Mobasher (2007) and Perkowitz and Etzioni et al. (1998).

Despite being such an extensive work carried out in the field of personalized information retrieval, question answering still has not advocated much of work in personalization. Hickl&Harabagiu(Hickl et al., 2006) compared a novice and expert model of the user while retrieving answer from an interactive however, the prime focus of the proposed research is other than personalization aspect just relying on simple two class modeling rather than modeling individual user profiles. A personalized QA system for the closed domain of business analysis is proposed by Thai et al. (2006), with the aim of taking into account context and recent user queries. In the work presented by Zhang et al. (2006) personalized services are enabled through modeling users’ profiles in the form of Ontology. Quarteroni et al. (2009) has addressed the issue of personalization by the integrating User Modeling component within the question answering model. The User
Model is able to filter and re-rank results based on the user’s reading level and interests. A comprehensive framework forming a unified model for future Question Answering has also been developed by combining personalized and interactive Question Answering. Unfortunately, the personalization aspect does not appear to be much implemented in question answering systems yet.

The earlier personalization approaches have contributed to the improvement of question answering system but there still require more sincere efforts which not only improve the relevancy of response list but with less complexity. Here, we proposed a personalization approach to the question answering system in restricted domain of higher education. A QA system has advantage of having some essential information at the earlier stage of question classification other than textual content due to inherent processing requirements. Therefore, we have put our efforts, to implement personalization with the help of various informative clues identified at earlier phases of question answering such as focus word and other keywords associated to focus word along with user’s previous transaction information.

6.2. User Model Design

Here we present a personalized approach to Question answering system. Our hypothesis is based on dynamic and repetitive process for designing a multifaceted user model which is used to illustrate and define different facets of user. The model discussed here comprises of three dimensions: user profile, text based model and entity model. If we suppose U as a set of user, then model for a user u ∈ U can be represented as:

\[ F_u = D_t \cup D_n \cup P \]  \hspace{1cm} (6.1)

Where,

\[ D_t: \] keywords employed by the users for their textual research

\[ D_n: \] entities visited by the user

\[ P: \] user profile
User profile is an important aspect of personalization approach. A user can be known or unknown to the system. A known user possesses an account and is identified with his various personal detail already provided to the system while an unknown user does not have so. Therefore, user modeling has to deal with different level of gap of information in both the cases. An unknown user modeling requires construction of user profile in real time and mapping to the most resembled prototype while a known user modeling requires initiation and update of the respective model in real time.

However, the construction of user model is based on implicit communication with user. The communication is termed as implicit because user doesn’t expected to give his views directly rather it has been deducted as an outcome of his navigation interest about an entity by the system itself. To quantify this communication, following measures have taken into consideration:

- The similarities that may occur between attributes and entities of user interest.
- The deduction of user interest from all its navigation.

6.3. Similarity Measure
Since we are mainly focusing on content based personalization, therefore we have tried to achieve the same through attribute values similarity and entity similarity.

6.3.1. Attribute values similarity
These attributes can occur individually such as in the form of numeric, Date or strings or can also be observed in the form of compound sets as numeric intervals or string sets. We define each type separately:

- **Numeric values**: These values occur distinctly, so it is required to have a proximity function to divide their variation intervals into set of similar values e.g., institution rank, fees. Each value which occurs in neighbourhood can be considered but with proper estimation of neighbourhood’s width as the width is directly proportional to the value. Higher the value, greater is the neighbourhood width. We can define neighbouring values as:

\[
\text{Sim}(v) = \{ \beta, \beta[\epsilon(\alpha - \epsilon), (\alpha + \epsilon) ] \} \quad (6.2)
\]
Where, ε defines the neighbourhood width between the values β and α.

- **String values**: These values also occur distinctly but their proximity function unlikely to the numeric values can be defined domain specific knowledge. In view of the fact that we are working in restricted domain, we already have worked on frequently occurring string attributes and collected them separately. As for now, we are restricting ourselves to adjectives and few other most frequent domain specific terminologies e.g., first, SC/ST, NRI, national etc. The neighbouring values can be determined by referring to the different instances head of pre stored domain specific adjectives.

- **Time/Date**: Similar to the numeric values, the proximity function to determine the neighbourhood values to date/time can also be defined using the proper width for each instance e.g., year, month, day, hour etc.

\[
\operatorname{Sim}(t) = \{s, s[\epsilon(\alpha - \epsilon), (\alpha + \epsilon)]\}
\]

(6.3)

Where ε defines the neighbourhood width in terms of year, month, day or hour etc. according to the values β and α.

### 6.3.2. Entity value similarity

The similarity measure between two entities depends on the similarity of concept they belong. However, in open domain, similarity between entities has been estimated based on the aggregation of their attribute values, but here in our case we can directly deduce domain dependent similar entities referring to the domain knowledge. Since, for our varying requirements of QA system design such as answer classification stage, we have by now accumulated the segregate collection of various domain specific terminologies viz., course, designation or organization name etc. which can be efficiently used to measure similarity between entities.

For rest of the entities, we will refer WordNet as resource and use semantic similarity by the Wu & Palmer’s Measure (1994) given as:

\[
S_t(Q_a, Q_b) = \frac{2 \times \text{depth}(Q_a, Q_b)}{\text{len}(Q_a, Q_b) + 2 \times \text{depth}(Q_a, Q_b)}
\]

(6.4)

Where,
Len($Q_a, Q_b$) : the length of the shortest path from synset $Q_a$ to synset $Q_b$ in WordNet.
also ($Q_a, Q_b$): the lowest common subsumer of $Q_a$ and $Q_b$
depth ($c_i$): the length of the path to synset $c_i$ from the global root entity, and
depth(root) = 1.

6.4. User’s Interest

The notion of inherent information collection is adopted for identifying user’s interest. It is assumed that when a person asked question about an entity or concept, he has obviously voted for the same. Therefore we aim to apply metrics to infer user’s interest. Here, suppose that we have a user $u$ who visits an item $y$ and a concept $C$ to which $y$ belongs.

6.4.1. Interest Pointer

User’s profile can be primarily considered for measuring the user interest towards accessed items. Following metrics could be used:

- **Access rate**: It represents the number of traversal to a given entity or concept. The number can reveal the inclination of the user towards an entity or group of entity and can be used to identify most frequently visited concept.

- **Access rate as focus word**: It represents the number of visits to an entity as focus word of the previously asked questions. This is much obvious measure for estimating user interest specifically. More times the user visit an item as focus word, more interesting the topic is considered to the user.

- **Explicit item rating**: The concept of head word discussed so far, not all the time goes correct with our assumption as like any other approach integrated classification mechanism too has limitations and if the mechanism gives erroneous result then their requires an explicit measure by the user to balance the effect of the previous one to determine proper rating.

6.4.2. Interest Prediction

- **Visited entities interest**: The key measures used here to figure out user interest are based on the access rate and access time spent towards an item or concept. To account
for these estimations, all previous saved user transaction along with focus word description has been used. If the user u improvises a specific rating to a previous visited item, the interest degree $I_e$ towards that item y can be estimated as:

$$I_e(u, y) = \frac{1}{2} \left( \frac{n_e(u, y)}{N(u)} + \frac{n_f(u, y)}{n_e(u, y)} \right) \times V(u, y) \quad (6.5)$$

Where

$n_e(u, y)$: Number of times user u asked about y

$N(u)$: Total number of user question

$n_f(u, y)$: Number of times user u asked about y as focus word

$V(u, y)$: Average rate assigned by u to y such that $0 \leq V(u, y) \leq 1$. The default value is 0.5.

- **Concept Interest:** It is worth stating that each candidate answer for each concept in the user profile represents the user from different viewpoint. The previous question answer session helps in identifying user’s specific preference for a concept. For example, a user might only be interested in specific parts of a concept. Therefore, our answer class taxonomy could be considered greatly while re-ranking retrieved answer list. Our proposed QA system has divided answer space into sub-domains to carve the system effort in searching an answer to the considered domain. The processing has also defined some constraints to be satisfied for classification as described by Singh et al. (2015b). Since the constraints were based on domain specific terminologies, therefore information regarding entities will help us in predicting the user inclination towards a sub domain or concept. In order to predict a personalized classification, we consider the entities as constraints to concept. With $y_i \in C$, the interest degree of u towards C is:

$$I_c(u, y) = \sum_{i=1}^{n} I_e(u, y_i) \quad (6.6)$$

Where

$n$: The number of entities $y_i$ referred by u related to the concept constraint.
L_e(u, y_i): The interest degree of u towards y_i.

- **Features value interest:** After classifying the domain information in form of broader concepts or sub-domain, there is requirement to order the entities along with their attributes which are more likely to get investigated according to their relevance. Nevertheless, personalized answer could be provided only if previous user transaction hold adequate information about a user’s interest. In such case not only entity but respective attributes should also be considered while re ranking results according to user’s interest. Consideration of attribute or entity solely does not solve our purpose of determining user’s inclination since an attribute to an entity maybe an entity itself. Also, we have been provided with focus word or head word information at earlier phase of question classification using integrated approach as proposed for the QA system by Singh et al. (2015a) that can be effectively used as an attribute to an entity. The information regarding head word or focus word information of user’s previous transaction not only gives important clue to deduce user's interest implicitly but also filter noise owing to other entities present in the question. As long as, QA system is provided with the head word information the system can easily deduce the user’s biases towards a concept or entity and suggest instances accordingly. However, focus word alone is not capable of defining the whole context; therefore the entities or attributes associated directly to focus in previous user transaction can be considered to order entities within the same concept according to their relevance. The attribute’s interest degree with respect to an entity has to be calculated within the concept.

Consider an attribute F_i ∈ C and, S_e the set of all entities associated with F_i as focus word within the concept whose co-occurrence frequency is equal to v_k, where 0 ≤ v_k ≤ n_d(u, F_i), the number of times user u asked about F_i as focus word

\[
I_a(u, F_i, v_k) = \frac{1}{n} \left( \sum_{j=1}^{n} I_e(u, X_j) \right)
\]

(6.7)

Where

I_e(u, X_j): is the interest degree of entity X_j associated with attribute F_i ∈ C
Chapter 6 Personalized Approach to Question Answering

\[ n \] : total number of entities associated with \( F_i \in C \)

Now, from above discussion we can estimate user preferences in terms of interest degrees. Therefore, total Interest count can be given as:

\[ I(u, x, f) = I_c(u, x) \ast I_a(u, F_i, v_k) \quad (6.8) \]

Where

\( v \) : is the maximum count for co-occurrence of entity x with attribute f as focus word

\( I_c(u, x) \): entity interest degree within the concept by the user

\( I_a(u, F_i, v_k) \): interest degree of attribute with respect to entity inside concept

**6.5. Relevance Computation**

For any question that has been posed to QA System, it has been already provided with focus word information at question classification stage. Therefore, keywords that occur just before or after focus word could be prime source while estimating for personalization. Primarily, considering only bigrams that necessarily constitutes focus word will reduce computational overhead while examining for attribute and entity relationship. However, the question whose focus word couldn’t be extracted properly at classification stage, all keywords bigram will be considered for estimating personalization.

For a given question, consider the set of retrieved answers or snippets as \( A= \{ a_1, a_2, a_3, \ldots a_m \} \), the keywords for each answer \( a_m \) are stored in array as \( x_1, x_2, \ldots, x_n \). The interest score that has been assigned to the particular answer can be estimated by the following equation:
\[ I_a(u, a_m) = \begin{cases} \sum_{i=1}^{n} I_c(u, x_i) + I_{j-1}(u, F_j, v_{j-1}) + I_{j+1}(u, F_j, v_{j+1}), & \text{if } F_j \text{ is known} \\ \sum_{i=1}^{n} (I_c(u, x_i) + I_i(u, x_{i+1}, v_k)), & \text{otherwise} \end{cases} \] (6.9)

Where \( n \) is the number of keywords array for particular answer or snippet, \( j \) is the index of focus word in the answer.

Finally, the answer or snippet which has highest interest score is considered higher in the ranked list. Re-ranking the search results according to the score is the final step for the proposed question answering personalization approach.

### 6.6 Summary

This chapter summarizes task of personalization based on end user modeling for question answering. Content based measures such as entities and their attributes appeared in user’s previous questions have been taken into account for building user profile.