A Survey on Web Service Mining by Collaborative Filtering and QoS

S. Urmela¹, Dr. K. Suresh Joseph², K. Vaitheki³

¹M.Tech Scholar, ²Assistant Professor, ³Assistant Professor, Department of Computer Science, Pondicherry University, Puducherry, India

Abstract—Web service mining has become one of the promising areas of Service Oriented Architecture. Web service discovery methods include syntactic based system and semantic based system. Recommender system plays a vital role in recommending a product to customer, seldom used for web service recommendation. A variety of techniques have been proposed for performing recommendation namely, content-based, collaborative, demographic-based and other techniques. To improve performance these techniques have been combined in hybrid recommendation system. The most widely used filtering technique of recommender system is collaborative filtering. In this paper, we describe various collaborative filtering by QoS rating techniques applied to web service mining and addresses various collaborative filtering problems namely, cold-start problem, gray sheep problem, synonym problem, ramp-up problem, shilling attack, data sparsity and scalability.

Keywords—Web service mining, Service Oriented Architecture, Recommender system, Collaborative Filtering, QoS.

I. INTRODUCTION

The amount of information found in Internet is growing more and more every year, making it necessary to develop new Information Retrieval techniques. Personalized Information Retrieval system is emerging more nowadays especially when not limited to just search information but also recommend product or service to customer based on certain parameters like QoS and trust metrics thereby increasing level of users’ satisfaction.

Hence, the so-called recommender system plays a key role for their ability to recommend a product or service to customer increasing customers’ satisfaction. At present recommender system, proves to be effective in recommending music, financial services, twitter followers, in particular for search queries also.

In their simplest form recommender systems provide a personalized and ranked lists of items by predicting what the most suitable items users’ need are, based on the users’ history, preferences and constraints.

Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' or 'preference' that a user would give to an item. These ratings or preference can be collected either actively or passively. Active user profile collection includes: asking a user to rate an item or product after usage, presenting two different items or products and asking user to rate them on a scale of 10. Passive user profile collection includes: Recording users’ history, analyzing his/her products purchased, analyzing social network profiles and discovering his/her likes and dislikes, etc [1].

Since multiple Web services provide same functionality, another parameter must be introduced to be set as a deciding factor. QoS is the suitable deciding factor, set of non-functional requirements like response time, accessibility, throughput, availability, etc. Current Universal description, discovery and Integration (UDDI) provide support of Web service retrieval by functional-requirement only. Web service mining based on Collaborative Filtering and QoS is gaining importance [2].

The organization of the paper is as follows: Section 2 describes an overview of web service concepts and presents a comparative study of web service mining techniques. Section 3 presents a comparison of various recommendation techniques. Section 4 presents various collaborative filtering techniques applied to web service mining based on QoS ratings or QoS profile.

II. WEB SERVICE MINING

This section gives an overview of web service and a comparative study of web service mining techniques.

A. Web Service

Web services are server (service provider) and client (service requestor) applications that communicates over the World Wide Web (WWW) HyperText Transfer Protocol (HTTP).
Web services are widespread by their interoperability, loose coupling, reusability and extensibility with the help of its components namely UDDI (Universal Description, Discovery and Integration), WSDL (Web Service Description Language), XML (eXtended Markup Language), and SOAP (Simple Object Access Protocol). UDDI is a registry where service provider registers their services; WSDL is used for describing the services; SOAP is used to transfer the data, enables communication between service provider and service requester; XML uses custom defined tags to describe the data in a structured manner [3].

Web services, combined in a loosely coupled manner to enable service requester to perform complex operations. Simple Web services interact with each other to deliver composite services [4][35]. Several models are available to web service developers for programming. These models fall into two categories, supported by the IDE:

- REST-based: REpresentational State Transfer, a way to create and communicate with web services. Resources having URIs are manipulated through Hyper Text Transfer header operations [34].
- SOAP/WSDL-based: Web service interfaces in traditional web service models, having URLs are exposed through WSDL documents. Message exchange between service provider and service requester is in SOAP, a type of XML document [5].

B. Web Service Mining

For using web service, the role of web service discovery methods is the core element for finding the desired services. Web service discovery is the process of retrieving web service based on requester queries and service descriptions.

In general web service mining can be categorized into non-functional requirement (syntactic) based system and functional requirement (semantic) based system. With large number of web services available, retrieval based on keyword or tags alone proved to be an ineffective technique. Firstly, large number of web services might be obtained by keyword based search. Secondly, identical web services results in poor precision. As a result it leads to unusable discovered web services in complex business environment. In recent times, solutions to overcome this problem have been motivated [2].

<table>
<thead>
<tr>
<th>Table 2.1 Comparison web service mining</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Techniques</strong></td>
</tr>
<tr>
<td>Searching based on non-functional</td>
</tr>
<tr>
<td>parameters</td>
</tr>
<tr>
<td>Keyword-based search</td>
</tr>
<tr>
<td>Trust model based approach</td>
</tr>
<tr>
<td>Schema matching approach</td>
</tr>
<tr>
<td><strong>Pros</strong></td>
</tr>
<tr>
<td>Widely-used technique</td>
</tr>
<tr>
<td>Reliable technique</td>
</tr>
<tr>
<td>Satisfies service requester query</td>
</tr>
<tr>
<td>implicitly</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 2.1 depicts web service mining comparison based on syntactic and semantic based system.
### Table 2.2
Comparison web service mining based on QoS

<table>
<thead>
<tr>
<th>Approach</th>
<th>Work done</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Web service composition based on dynamic QoS prediction [6] [37]         | - Web service composition based on global and dynamic QoS decomposition and prediction  
- Reliable approach                                                    | - Relying on Global QoS constraints can’t increase the accuracy of web service prediction method.  
- Only few QoS parameters namely Response time, reliability, cost and availability are considered. |
- In initial sub-process, WSDL documents meeting QoS requirements are matched and outcome is a list of web services satisfying user’s QoS request.  
- In second sub-process, output list is sorted based on weights of QoS metrics of the service consumer. | - Existing WSDL language is unreliable to track QoS attribute values.  
- Existing matchmaker algorithm is unable to take correct decision based on QoS parameters, thereby delivering incorrect web services with QoS constraints. |
| Hybrid approach end-to-end QoS constraints for Web service composition [7] [37] | - Hybrid solution that combines global optimization with local selection techniques. | - Handling global constraints, can lead to poor performance rendering inappropriate for applications with dynamic and real time requirements.  
- Only few QoS parameters are considered. |
| QoS Prediction approach-User Clustering and Regression Algorithm [8] [37] | - An approach for web service selection which can provide the approximate QoS value for users, and support finding the optimal web service.  
- Firstly, it clusters the users based on location and network condition, then according to the QoS historical statistics. | - Clustering based on users QoS historical statistics/ratings given by users can’t be taken into account because a service provider, acting as malicious users can rate its target service provider with poor ratings.  
- Only few QoS parameters are considered. |
| WSPred: Time-aware personalized QoS prediction framework [9][33][37]    | - From past user’s usage scenario, data is collected and a feature model is built.  
- Personalized web service selection based on previous user’s QoS prediction. | - This approach is time-consuming, resource-consuming.  
- Only few QoS parameters are considered. |
| Ranking of Web service based on QoS Using Associative Classification [10] | - Web service selection by ranking QoS parameters with service consumer preferences and then ranking based on semantic matching.  
- This work involves 3-process namely, preprocessing of QoS attributes, service selection by local classification and ranking of web services based on functional aspects. | - Implementation of utility value dependence for each web service is not a reliable approach because the utility value threshold has to be set high enough for reliable classification of web services based on CBA algorithm. |
- The QoS attributes are optimized and ranked by the algorithm produced.  
- Rank value will be stored in the database named, QoSDB.  
For each query, the QoS manager process the request and list the matched web services useful for setting QoS preferences. | - Only 5 QoS attributes namely availability, throughput, response time and cost are considered.  
- Constant QoS ranks are assigned, which leads to poor search results - needs to be changed for each type of web service. |
| Agent based Model for Web service discovery with dynamic QoS [12]       | - The WSA has four components: Service Publisher, Verifier and Certifier, Retrieval Agent, Quality Analyzer and Web Service Storage (WSS). An agent service is used to facilitate service registry access.  
- The agent performs the interaction with the UDDI. The WSA assists clients in selecting web services based on a set of QoS parameters. The broker is a web services performing a collection of QoS functionalities. | - Only 4 parameters namely, response time, throughput, price and availability are considered.  
- Web service agent is not a reliable approach, because agents are not a trustworthy component.  
It can be easily overridden by malicious users. |
A. Recommender System

Recommender systems are systems that recommend products or services based on users’ past behaviour or consumption patterns. Recommender system is broadly classified as Content-based, Knowledge-based, Collaborative Filtering and demographic-based [14] [32]. Table 3.1 depicts comparison of various recommender systems.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Description</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>Recommends based on information about items</td>
<td>Incapable to find users' interest towards products or services, Impossible to find product quality</td>
</tr>
<tr>
<td>Knowledge-based</td>
<td>Recommends products or services based on inference about users' preferences</td>
<td>User profile maintenance is tedious</td>
</tr>
<tr>
<td>Collaborative Filtering</td>
<td>Recommends by collecting and analyzing users’ past behaviour, preferences</td>
<td>Suffer from grey sheep problem, cold-start problem, shilling attack, etc. Malicious users’ rating</td>
</tr>
<tr>
<td>Demographic-based</td>
<td>Classification based on demographic categorizations</td>
<td>Possibility of wrong personal categorization</td>
</tr>
<tr>
<td>Hybrid Recommender</td>
<td>Combines the above techniques or individual approaches of each technique</td>
<td>Inappropriate combination of recommender technique leads to poor recommendation or decreased precision result</td>
</tr>
</tbody>
</table>

B. Collaborative Filtering

Collaborative filtering technique aims to recommend a product or service to targeted user based on other users’ ratings towards the product or service. In general collaborative filtering can be classified into 2 broader areas namely; Memory based Collaborative Filtering and Model-based Collaborative Filtering. Table 3.2 depicts various Collaborative Filtering techniques.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Technique</th>
<th>Description</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory-based Collaborative Filtering [15] [32]</td>
<td>User-based</td>
<td>Recommends product or service based on user-user similarity</td>
<td>Data Sparsity, Scalability, Doesn’t generalize data, Overfit, Gray sheep problem, Cold-start problem, Ramp-up problem</td>
</tr>
<tr>
<td>Item-based</td>
<td></td>
<td>Recommends product or service based on item-item similarity</td>
<td></td>
</tr>
<tr>
<td>Model-based Collaborative Filtering [16] [32]</td>
<td>Bayesian Network</td>
<td>Recommends product or service based on Bayesian network model</td>
<td>Inflexible quality of predictions, Synonyms Problem, Cold-start problem Ramp-up problem</td>
</tr>
<tr>
<td>Rule Association</td>
<td></td>
<td>Recommends product or service based on assumption or association</td>
<td></td>
</tr>
<tr>
<td>Clustering model</td>
<td></td>
<td>Recommends product or service based on clustering users’ or items</td>
<td>Calculated similarity among users by user ratings mean</td>
</tr>
<tr>
<td>Hybrid Collaborative Filtering</td>
<td>Similarity fusion [17][32][38] (user-based + item-based)</td>
<td>Recommends product or service by analyzing personality types among users</td>
<td>Different ratings evaluated-malicious attack</td>
</tr>
<tr>
<td></td>
<td>Personality diagnosis [18][32] (memory-based + model-based)</td>
<td>Recommends product or service by analyzing personality types among users</td>
<td>Analyzing personality type by weighted switching technique is time-consuming</td>
</tr>
</tbody>
</table>

III. RECOMMENDER SYSTEM

A. Recommender system

Recommender systems are systems that recommend products or services based on users’ past behaviour or consumption patterns. Recommender system is broadly classified as Content-based, Knowledge-based, Collaborative Filtering and demographic-based [14] [32]. Table 3.1 depicts comparison of various recommender systems.
C. Challenges of Collaborative filtering

**Scalability:** In the entire environment where recommendation plays a vital role, involves various products or services. Hence computational complexity of recommending product or service is proportional to the number of products or services and number of active users’. However this problem can be solved by employing effective recommender algorithm.

**Data Sparsity:** The amount of products or services available on internet is large. Hence only active users’ rate these services. The problem of data sparsity is that even a good products or services have poor ratings.

**Cold-start problem:** For new users’, user profile will be initially created with no rating of targeted users’ by other users’. This approach is similar for items or services too. The problem of cold-start is that these users’ or items’ cant’ be recommended even though functionality of product or service is good.

**Gray-sheep problem:** Competitive service provider might provide poor ratings to its competitive services thereby decreasing its chance of being recommended. Similarly, competitive service provider might provide good ratings to its own services thereby increasing its chance of being recommended. Even some users’ acting as malicious users’ might provide inappropriate ratings to products or services. For new users’, user profile will be initially created with no rating of targeted users’ by other users’. This approach is similar for items or services too.

**Ramp-up problem:** Similar to cold-start problem.

**Synonym problem:** Most product or service either similar or same available seems to have different entries as services. The problem of Collaborative Filtering is that it doesn’t handle this type of association.

**Shilling attack:** Competitive service provider might provide poor ratings to targeted services or good ratings to its own services, similar to gray-sheep problem [19] [38].

IV. WEB SERVICE MINING BY COLLABORATIVE FILTERING AND QoS

We explored various research papers to study about various web service mining by collaborative filtering techniques and QoS and present their cons.

A. QoS Profile Description

The various QoS parameters which we had considered essential for web service mining are listed here,

- **Accessibility:** Accessibility is an important aspect of quality service which represents the degree a web service is capable of serving a Web service query [20].
- **Accuracy:** Accuracy represents the probability of the request being responded correctly while providers answering the client’s requests.
- **Availability:** Availability was measured by the mean ratio of the whole of times that the users can access the service successfully divided by the whole of time that the users use to request for the service.
- **Best practices:** Compatibility with WS-I Basic Profile.
- **Compliance:** Compatibility with WSDL specification.
- **Cost:** Cost of service measuring from the rate of service charge of the service providers in the same group.
- **Documentation:** Measure of documentation (description tags) in WSDL.
- **Execution time/transaction time:** The time taken by the service to execute and process its sequence activities.
- **Latency:** Latency is value obtained from subtracting response time from request time of the web service invocation.
- **Performance:** Performance is an important quality aspect of Web service, which measures in terms of throughput and latency. Lower Latency and Higher Throughput represents a good performance of web service [20].
- **Reliability:** Reliability is measured by ratio of all of the times that the users request for the service successfully divided by the whole of the times that the users request for the service in specific time.
- **Response time:** Response time is an important quality aspect of Web service, measured from the time since the users send their requests to service sever until it was responded.
- **Scalability:** Scalability is an important quality aspect of Web service, refers to the ability to consistently serve the service requesters’ request despite variations in the volume of requests [20].
- **Successability:** Successability refers to the extent to which web service provider yields successful results over service requesters’ request messages [21].
- **Throughput:** Throughput refers to the maximum number of services that a platform providing web services can process for a unit time [21].
Some related work in the field of web service mining by recommender system and QoS is discussed here.

A. By Mohamad Mehdi et al, “Probabilistic approach-Trustworthy web service selection based on QoS” [22][36], involves a probabilistic approach for predicting the quality of a Web service based on the evaluation of past experiences (ratings) of each of its consumers. QoS ratings of services are represented using different statistical distributions, namely multinomial Dirichlet (MDD), multinomial generalized Dirichlet (MGDD), and multinomial Beta-Liouville (MBLD). Bayesian inference method is employed to estimate the parameters of the mentioned distributions, which presents a trustworthy web services to service consumer Experimental evaluation involves 3 classifier namely: classifier 1- Bayesian approach with the Beta-Liouville distribution, classifier 2-Bayesian approach with a Dirichlet and classifier 3-compare them to the state of the art naive Bayes (NB) classifier.

B. By Lin, S-Y et al, “Web service discovery-Trustworthy QoS-based collaborative filtering approach” [23], deals with a trustworthy two phase web service discovery mechanism based on collaborative filtering and QoS. In the first phase, the observer agents will collect records of user behavior, including querying and invoking web services and monitor actual QoS, and then store the profile information in the public cloud database. This phase involves 3 sub-phases namely establishing query and web services matrices, finding query similarity and calculating the relevance between query and web services. This phase mainly establishes item based (memory based) collaborative filtering. The result of phase 1 discovered services may satisfy users’ functional requirements and have correct QoS information. In the second phase, the QoS scores of the selected web services are derived from the QoS information stored in database. This phase involves 3 sub-phases namely establishing a matrix of QoS and web services, normalizing the QoS value, and calculating the QoS score. A high QoS score indicates that the web service meets the requirements of a user. Finally, the suitable web services with high QoS scores are recommended to the target users.

C. By Sheng et al, “Combining Collaborative Filtering with Content-based Features for recommending Web Services” [24][35], proposes a novel approach that dynamically recommends Web services satisfying users’ interest.

The proposed work involves a hybrid approach of both collaborative filtering and content-based recommender systems. Experimental results show that the proposed hybrid system outperforms the latter two recommendation system it terms of recommendation performance.

D. By Chen et al, “Similarity-Aware Slope One Collaborative Filtering- QoS Prediction for Web Services” [25], employs similarity-aware slope one algorithm for QoS ratings prediction. The proposed work combines both Pearson similarity and slope one measurement for QoS ratings prediction. Weight adjustment and SPC (Statistical Process Control) based smoothing is also utilized for abnormal data smoothing. The proposed work shows better precision result compared with slope-one and famous WSRec system. The work has the capacity to reduce noise in QoS ratings data.

E. By Qi Yu et al, “Collaborative QoS evaluation- QoS-aware service selection” [26][38], proposes a service selection scheme that provides automation for assessment of QoS of an unknown service providers thereby providing a reliable web service that matches service requester’s query. Relational Clustering based Model (RCM), which effectively addresses the data scarcity issue. Experimental results of RCM model on both real and synthetic datasets demonstrates that the proposed automation model can more accurately and reliably predict the QoS parameters of an unknown web service, matching service requester’s query.

F. By Yali LI et al, “Hybrid Collaborative Filtering-Web Service Recommendation” [27], proposes a hybrid method that takes into account user-based and item-based collaborative filtering algorithm, making improvement on similarity calculation by adopting Pearson Correlation Coefficient (PCC) to measure the similarity between two users or two services.

G. By Zheng et al, “Collaborative Filtering- QoS-Aware Web Service Recommendation” [28][33], proposes a Collaborative Filtering recommendation method for QoS prediction of web services, making advantage of past usage experience of service requester. Initially, a user-collaborative mechanism for collecting past Web service QoS information from different service requester is done. Finally, based on the QoS data collected, a collaborative filtering recommendation is designed for prediction of Web services with unknown QoS values. A prototype model named, WSRec is implemented and experimental results show that proposed model achieves better prediction accuracy than traditional approaches.
H. By Yechun Jiang et al, “Personalized Collaborative Filtering- Effective Web Service Recommendation” [29] [38], describes an effective personalized collaborative filtering method for web service recommendation. A key component of Web service recommendation techniques is computation of similarity measurement of Web services. Apart from the Pearson Correlation Coefficient (PCC) similarity measurement, they take into account the personalized influence of services when computing similarity measurement between users and personalized influence of services. An effective Personalized Hybrid Collaborative Filtering (PHCF) technique is developed by integrating personalized user-based (memory-based) algorithm and personalized item-based (memory-based) algorithm. Experimental results show that the method improves accuracy of recommendation of Web services significantly.


J. By Chen et al, “Collaborative Filtering-Personalized Context-Aware QoS Prediction for Web Services” [31], describes a personalized context-aware QoS prediction method for web services recommendations based on the slope one approach. Proposed work considers context, which is an important factor in both recommender system and QoS prediction. Experimental results demonstrate that the suggested approach provides better QoS prediction.
Table 4.1  
Comparison web service mining by Collaborative Filtering and QoS

<table>
<thead>
<tr>
<th></th>
<th>User Based CF</th>
<th>Item Based CF</th>
<th>Hybrid CF</th>
<th>Model Based CF</th>
<th>Approach</th>
<th>QoS Evaluation</th>
<th>QoS-trust Evaluation</th>
<th>Related Attributes</th>
<th>Algorithm</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Bayesian Network</td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T A R)</td>
<td>Bayesian Network Classifier</td>
<td>Inflexible Quality of predictions Synonyms Problem</td>
</tr>
<tr>
<td>B</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td></td>
<td>√ (Metrics)</td>
<td>X</td>
<td>(P A RT T)</td>
<td>Matrix Formulation</td>
<td>Data Sparsity Scalability Doesn’t generalize data Overfit Gray sheep problem Ramp-up problem</td>
</tr>
<tr>
<td>C</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>√ (CF and Content-Based)</td>
<td>X</td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T)</td>
<td>Malicious user ratings</td>
</tr>
<tr>
<td>D</td>
<td>X</td>
<td>√</td>
<td></td>
<td></td>
<td>Slope One Method</td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T)</td>
<td>Pearson similarity and slope one method</td>
<td>Time-consuming Cold Start problem Gray sheep problem</td>
</tr>
<tr>
<td>E</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T A R)</td>
<td>Cosine Similarity</td>
<td>Gray sheep problem Data Sparsity Scalability Doesn’t generalize data Overfit</td>
</tr>
<tr>
<td>F</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td></td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T A R)</td>
<td>Pearson Correlation Coefficient</td>
<td>Malicious user ratings</td>
</tr>
<tr>
<td>G</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td></td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T A R)</td>
<td>Correlation Similarity</td>
<td>Data Sparsity Scalability Malicious user ratings Doesn’t generalize data Gray sheep problem</td>
</tr>
<tr>
<td>H</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td></td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T A R)</td>
<td>Pearson Correlation Coefficient</td>
<td>Malicious user ratings</td>
</tr>
<tr>
<td>I</td>
<td>X</td>
<td>√</td>
<td>X</td>
<td>X</td>
<td></td>
<td>√ (User Ratings)</td>
<td>X</td>
<td>(RT T)</td>
<td>Pearson Correlation Coefficient</td>
<td>Malicious user ratings</td>
</tr>
<tr>
<td>J</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Slope One Method</td>
<td>√ (Metrics)</td>
<td>X</td>
<td>(RT T)</td>
<td>Slope One Method</td>
<td>Time-consuming Cold Start problem Gray sheep problem</td>
</tr>
</tbody>
</table>

*P-Price  A-Availability  RT-Response Time  T-Throughput  R-Reliability

√ - Indicates presence
X - Indicates absence
REFERENCES


[34] Java-Stay Hungry stay Foolish: http://lorddisk.wordpress.com/category/java/


[38] http://link.springer.com/article/10.1007/s11280-012-0186-0