Personalization Services in CADAL

ZHANG Yin, ZHUANG Yueting, WU Jiangqin

School of Computer Science, Zhejiang University, Hangzhou 310027, China

E-mail: zhangyin98@cs.zju.edu.cn; yzhuang@cs.zju.edu.cn; wujq@cs.zju.edu.cn

***ABSTRACT* — CADAL is a part of universal digital library project supported by the Chinese government, 16 Chinese and American first-class universities that have taken part in the collection of resources and the construction of technical environment.. Finding useful information and knowledge in the digital library is a time consuming process. Personalization Services help individuals and communities address the challenges of information overload.**

This paper shows the architecture of the entire Personalization Services for CADAL, and techniques exploited by us to construct it, such as query expansion, relevance feedback, user modeling, collaborative-based and content-based filtering methods. The system keeps track of user interests in different domains by automatically analyzing users' query in our search engine and browsing behaviors in the website. Questionnaires are presented to users to explicitly give out the ratings about specific items, based on which the system predicts the potential interested items for the individuals.

***Index Terms*— Information System, Information Service, Information Retrieval**

I. Introduction

The Million Book Project [1] is a part of a larger universal digital library initiative by the computer scientists and information experts at CMU. Building a Universal Digital Library (UDL) to contain all existed books step by step will realize the dream of sharing all human knowledge. First challenge for UDL is to make the one hundred million books with text and images online and globally accessible. To meet the challenge, China and US parties initiated the China-US Million Book Digital Library Project [2].

The State Development and Reform Commission, Ministry of Education and Finance Ministry of China had agreed to support the China-US Million Book Digital library with a project "Chinese American Digital Academy Library (CADAL)" as a part of the Project 211 in the tenth Five-year Plan. CADAL is led by Zhejiang University and the Graduate school of Chinese Academy of Science. Now more than nine hundred thousand digitized books were collected and uploaded to the Library of Zhejiang University. People from world wide can online access this large digital collection through the website "www.cadal.zju.edu.cn". It is a key problem for the individuals and communities to find useful information from the large number of books in CADAL at right time and place. Therefore, personalization services are constructed to alleviate this problem in CADAL portal.

Web personalization[4] is the process of customizing a web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web Context, namely, structure, content, and user profile data. In this paper, we focus on the personalization, not layout customization.

Web data can be classified in four categories [9]:

1) Content data are presented to the end-user appropriately structured, such as text, images

2) Structure data represent the way content is organized. They can be either HTML/XML tag or hyperlinks connecting one page to another.

3) Usage data represent a website's usage, such as visitor's IP address, access time, access path and other attributes that can be contained in web access log.

4) User profile data provides information about the user of the website. A user profile contains demographic information (such as name, age, country, education etc.), as well as information about the user's interest and preference. Such information is acquired through the registration form or questionnaires, or can be inferred by usage data.

Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs taking advantage of the user's navigational behavior. The steps of a Web personalization process include: 1) the collection of Web data, 2) preprocess these data, e.g. transformation and modeling, 3) the analysis of the collected data, 4) the determination of the actions that should be performed. The way that are employed to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering, and Web usage mining. So far, we have implemented the basic rule-based filtering method, the content-based filtering methods exploiting user feedback, and collaborative filtering method based on cluster smoothing as personalization services in CADAL Portal. The query expansion and relevance feedback techniques were employed to construct the function of personalized search. In the future, the personalization techniques based on web usage mining will be

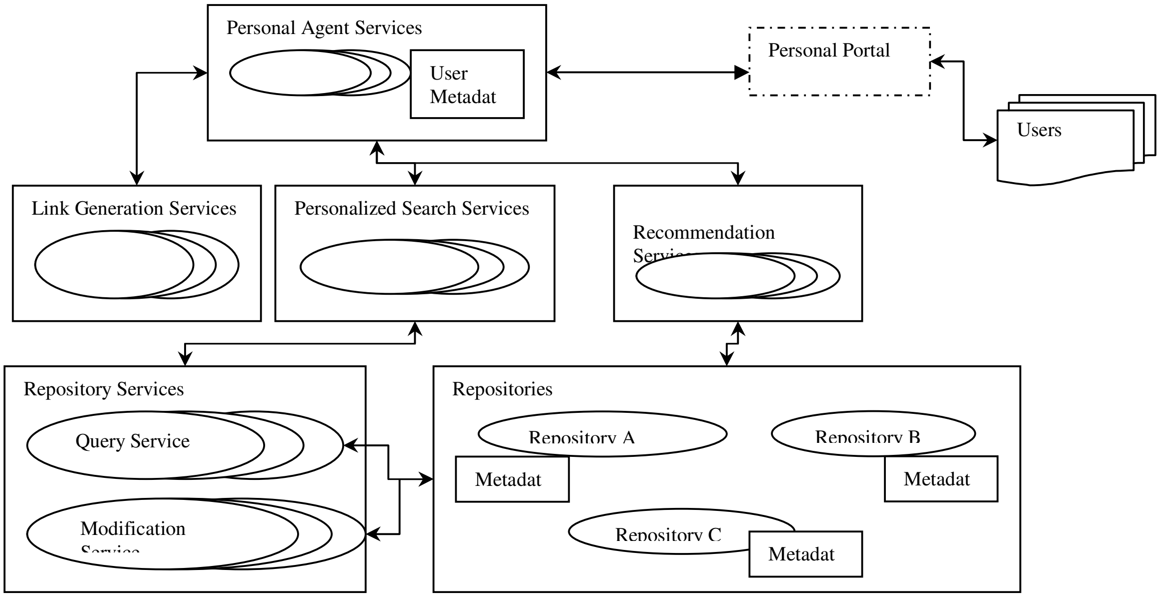


Fig 1 the architecture of personalization services in CADAL

investigated, especially when the more and more logging information is available.

The rest of paper is structured as follows: First we introduce the related works to our personalization implementation. Section III shows the personalization architecture in CADAL project and defines the specific services. Section IV shows the keyword expansion and relevance feedback techniques exploited in personalized search. Section V describes the personal web portal and basic rule-based filtering techniques. Section VI shows personalization services based on Information filtering techniques. The paper ends with conclusion and remarks on further work.

II. RELATED WORKS

Personalization services in CADAL portal have been built with reference to many other published papers and works done by predecessors. [3] shows the architecture of personalization service in distributed environments for e-learning. Our current implementation in CADAL is still lack of ontology support services that existed in ELENA [3]. [4] summaries the web mining techniques in web personalization service, but focus on the web-usage data mining in detail. Since the logging data is being collected, the recommendation services in CADAL are mainly based on the contents of digital items and user ratings on items now. [5] [6] propose the practical solutions for personalized search, which are implemented in CADAL. We have made use of the filtering techniques described in [7] [8]. These filtering techniques can obtain a good trade-off between the performance and system scalability.

III. THE PERSONALIZATION ARCHITECTURE IN CADAL

The personalization architecture in CADAL benefits from the semantic web technologies: the metadata description of digital resources provided in the various repositories follows the Dublin Core standards [10]. Services which carry out personalization functionality like personalized search or recommendations, as well as other required support service can be described in OWL-S [11], and are accessible via WSDL [12] and SOAP [13]. The seamless integration and the flow of results between services and the presentation of results to users are shown in fig.1. These services are composed to serve the users through the personal portal. In the following, the services defined in this figure are to be described.

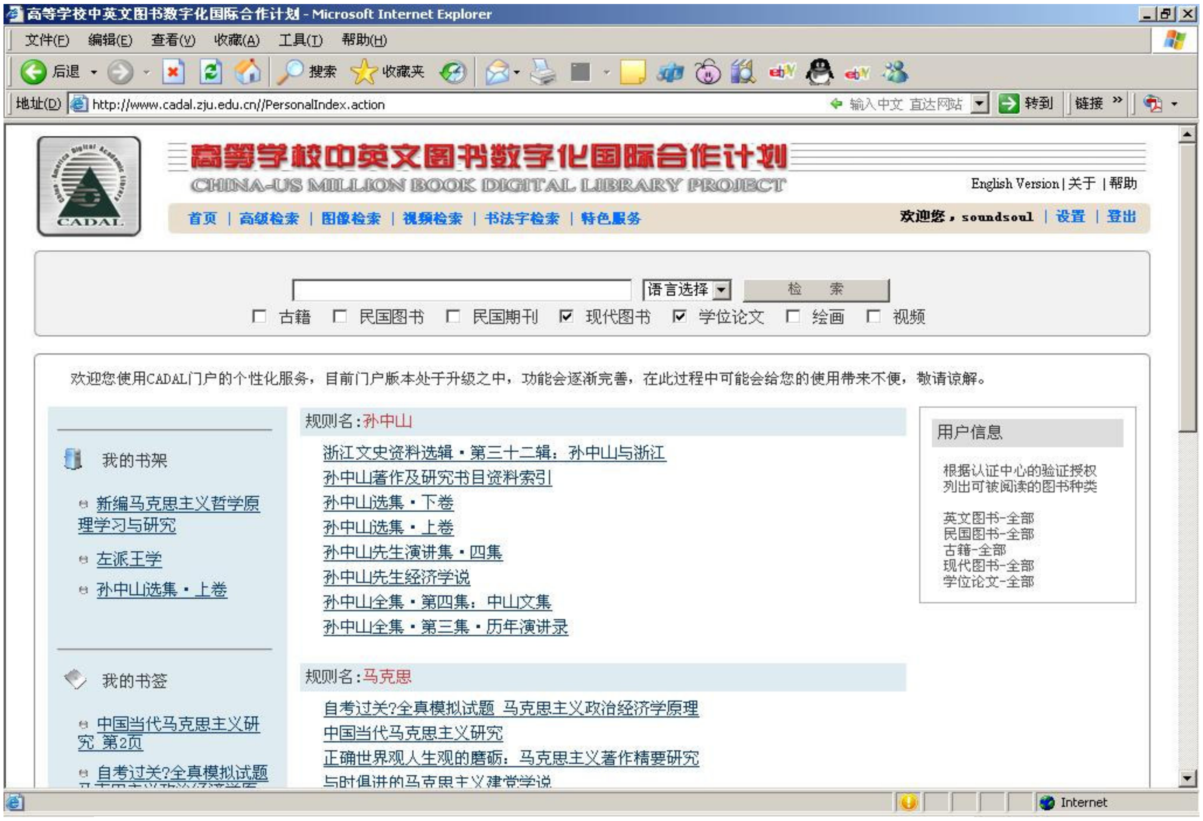
A Personal Agent Service

The central component of the personalization architecture in CADAL is the personal agent service which finds and integrates the other service described in the following subsections to help users to find appropriate digital books or other multimedia information in CADAL huge volumes of data.

The personal agent service is exposed through the personal web portal which can be browsed through the internet browsers by user. A user can view the interface of personalized search and the recommended results by recommendation services on the personal portal. Fig 1 is the screen snapshot of the entry of personal portal in CADAL.

B Link Generation Service

A link generation service provides personalized semantic relations for a digital item in accordance with


高等学校中英文图书数字化国际合作计划 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.cadal.zju.edu.cn//PersonalIndex.action
输入中文 直达网站 转到 链接
CADAL
高等学校中英文图书数字化国际合作计划
CHINA-US MILLION BOOK DIGITAL LIBRARY PROJECT
English Version | 关于 | 帮助
首页 | 高级检索 | 图像检索 | 视频检索 | 书法字检索 | 特色服务
欢迎您，soundsoul | 设置 | 登出
语言选择
检 索
Internet

□ 古籍 □ 民国图书 □ 民国期刊 ☑ 现代图书 ☑ 学位论文 □ 绘画 □ 视频

欢迎您使用CADAL门户的个性化服务，目前门户版本处于升级之中，功能会逐渐完善，在此过程中可能会给您的使用带来不便，敬请谅解。

我的书架

- 新编马克思主义哲学原理学习与研究
- 左派王学
- 孙中山选集・上卷

我的书签

- 中国当代马克思主义研究 第2页
- 自考过关?全真模拟试题

规则名:孙中山

浙江文史资料选辑・第三十二辑：孙中山与浙江
孙中山著作及研究书目资料索引
孙中山选集・下卷
孙中山选集・上卷
孙中山先生演讲集・四集
孙中山先生经济学说
孙中山全集・第四集：中山文集
孙中山全集・第三集・历年演讲录

规则名:马克思

自考过关?全真模拟试题 马克思主义政治经济学原理
中国当代马克思主义研究
正确世界观人生观的磨砺：马克思主义著作精要研究
与时俱进的马克思主义建党学说

用户信息

根据认证中心的验证授权
列出可被阅读的图书种类

英文图书-全部
民国图书-全部
古籍-全部
现代图书-全部
学位论文-全部

Fig 2 the entry page of the personal portal

the information in a learner's profiles. These relations can show the context of an item (e.g. the metadata about the digital book), or they can show other books related to this browsed book. CALIS project is another China academic digital library in line with CADAL. The link generation services dynamically generate the links for the browsed book to the union catalog service and intra-library loan system constructed by CALIS.

C Personalized Search Service

The personalized search service is based on the query expansion techniques that extend a user query by additional restrictions and variables based on user feedback. Personal agent service keeps track of the user navigation process and underlying strengths the query through identifying the concepts that occur most frequently in the metadata of relevant digital book. The query expansion techniques will combine the collected concepts and user's query to limit the number of irrelevant results and rank the most related result at the top.

The query expansion service is to be enhanced by ontology techniques [14] when the comprehensive Chinese ontology collections are constructed.

D Repository Service

In general, repository services provide access to any kinds of repository which is connected to a network. All books in CADAL are scanned to tiff image and encoded in djvu format to publish to the public. All books are classified in publish time, publish type and authoring language as six repositories such as modern books, dissertation & thesis, ancient books, minguo books, minguo journals and English books.

A repository service maintains a link to a metadata store which stores the metadata edited by manual in accordance with Dublin Core metadata standard.

Repository service can be two kinds: query service and modification service. The repository can be asked to return references to the digital books matching the given query, to create the references to the new digital book and its metadata, to delete the reference to the digital book, and to modify the metadata of the digital book.

E Recommendation service

The recommendation services analyze the user's preference data about the digital books collected by personal agent service when the user answers the questionnaires listed in the related info page of the browsed book. Then recommendation services determine what books are to be recommended to the user based on value of the similarity computed by the filtering methods. The similarity values of top 300 recommended items are to be added into user profiles. Users can see these recommended items sorted by the similarity value after they login the web portal. The recommendation services now periodically update the recommendation items for the specific user at the middle night every day in order to exploit the sparse computing time at night and save the computing time to afford intensive service at day.

F User Metadata

User metadata or user profile is composed of the information that the personal portal collect through the registration form or the questionnaires filled by user itself, the list of favorite books and bookmarks in reading books, the preference rules set by user according to which the personal portal display the monitored results on the entry page of personal portal, and the recommendation items recommended by the collaborative and content-based recommendation methods.

IV. Personalized search

Many users often send one or two keywords as a query to the search engine, the results obtained with them are not always satisfactory. These results can be improved by expanding the query with additional search items. Queries can be expanded in different manners. With manual query expansion, users indicate which item should be used for expansion. With automatic expansion, a system selects the terms for expansion.

Query expansion depends on the natural language processing techniques and relevance feedback methods. Explicit user relevance feedback is based on users' indicating which results of a search are relevant. Based on this evaluation, terms from the relevant documents are used for query expansion either automatically or indicated by user. Implicit relevance feedback deduces the evaluation from the user behavior without asking the user for the feedback. The terms often are automatically used to modify the user query. Magennis and van Rijsbergen [15] find that for automatic query expansion the optimal number of required expansion terms could be as low as six additional terms. Belkin et al. [16] compared automatic query expansion and manual query expansion for the TREC-8 interactive task and found no differences in performance and preference by the users. Later, White et al. [17] argued in TREC-10 that implicit feedback can substitute for explicit feedback. Moreover, in a real setting, users seldom request the query expansion. It is therefore our purpose to use the implicit feedback for automatic query expansion so as to not burden a user with additional tasks.

[5] proposes a related implicit measure that we believe can provide a good indication of the user interests: examining the links followed or ignored by the users. If a user follows a links, something of interest must appear in the metadata description of the browsed digital book. If the user ignores a link, nothing interesting is presented. When a link was followed, we consider the metadata information about this book routed by this link as relevant but not the contents of this book since the user has not yet read this book. This method doesn't intrude on user privacy, nor does it require any additional user effort.

A Keyword Expansion

Single keyword is usually ambiguous, or too general. Moreover, they can occur in vast quantities of documents, thus making the search return the hundreds of hit, most of which is irrelevant to the intended user query. Giving the additional keywords can refine the search providing the considerable improvement in the retrieval accuracy. We extract the words that mostly co-occur with the user query in its intended meaning in the large number of metadata descriptions of digital books. One of the characteristic of good refinement words is that they be domain specific. In this section, we present the method for automatically finding appropriate keywords to constrain and refine search for relevant books.

The Trigger Pairs Model [6] was used to identify the most similar pairs of words. If a word S is significantly correlated with another word T , then $\left(S,T\right)$ is considered as a "trigger Pair", with S being the trigger and T the trigged word. When S occurs in the document, it triggers T , causing its probability estimate to change, i.e. when we see the word S appearing at some point in a text, we expect the word T to appear after S with some confidence. The manual information (MI) that considers the words order is a measure of the correlation and used to extract trigger pairs from large corpus. The mutual information is given by the following formula:

$$\mathrm{M\,I}\left(s,t\right)=\mathrm{P}\left(s,t\right)\log\frac{\mathrm{P}\left(s,t\right)}{\mathrm{P}\left(s\right)\mathrm{P}\left(t\right)} \tag{1}$$

To evaluate the method, we use the all metadata descriptions about the digital book of 100M bytes and set the maximum distance between S and T to 300.

The trigger pair method can provide several candidate refinement keywords. An additional question is, how many and which ones to use under any given circumstances. For a search with only one keyword, the top several triggers to the keyword are used to expand the search. But for a search with more than two keywords, the choice becomes more complicated. The following algorithm is used for keyword expansion based on the trigger pairs.

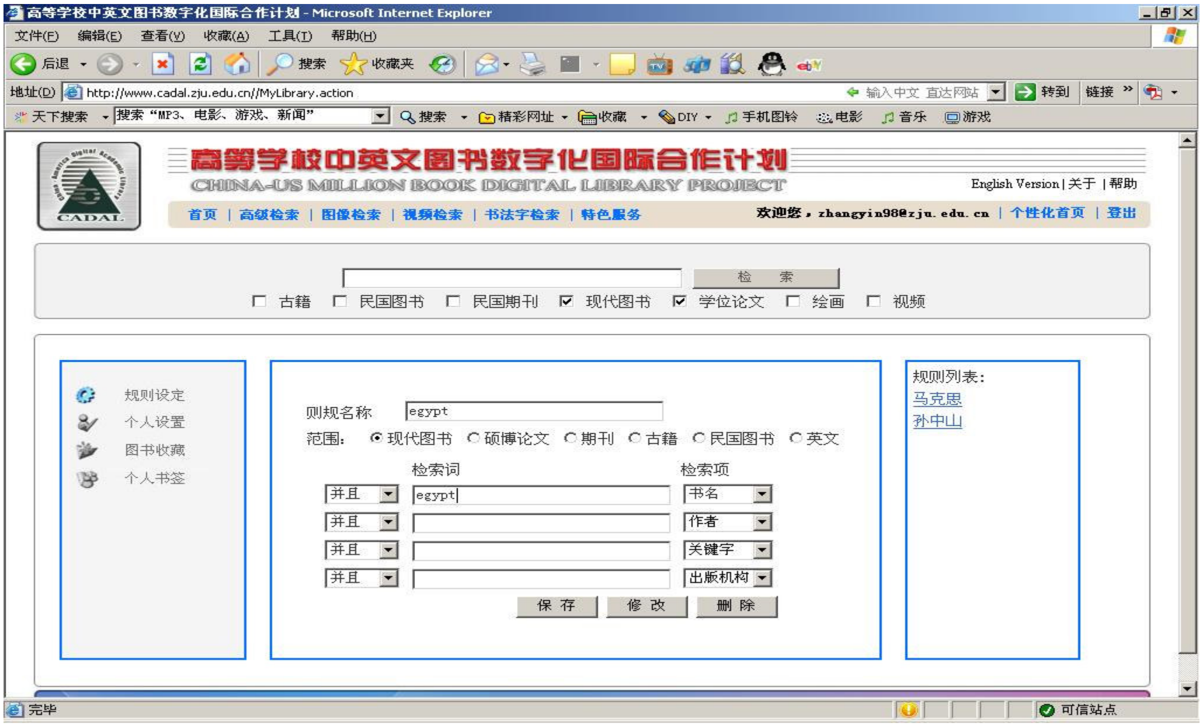
We define that the keywords are $K_1, K_2, \ldots K_m$, and the expected number of refinement words is N . Initialize $n=m$, S is the empty set.

1. $S_1=\{s_{11},s_{12},\ldots,s_{1i}\}\rightarrow K_1$ is the trigger set to K_1 . $s_{11},s_{12},\ldots,s_{1i}$ are sorted in decreasing order of the mutual information.

$S_2=\left\{s_{21},s_{22},\ldots,s_{2j}\right\}\rightarrow K_2$ is the trigger set to K_2 .

…

$S_m=\{s_{m1},s_{m2},\ldots,s_{mk}\}\rightarrow K_m$ is the trigger set to K_m



高等学校中英文图书数字化国际合作计划 - Microsoft Internet Explorer

地址(D) http://www.cadal.zju.edu.cn//MyLibrary.action

□ 古籍 □ 民国图书 □ 民国期刊 ☑ 现代图书 ☑ 学位论文 □ 绘画 □ 视频

则规名称 egypt

范围: ● 现代图书 ○ 硕博论文 ○ 期刊 ○ 古籍 ○ 民国图书 ○ 英文

	检索词	检索项
并且	egypt	书名
并且		作者
并且		关键字
并且		出版机构

规则列表:
马克思
孙中山

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退
搜索
收藏夹
输入中文 直达网站
转到
链接
天下搜索
搜索 "MP3、电影、游戏、新闻"
搜索
精彩网址
收藏
DIY
手机图铃
电影
音乐
游戏
CADAL
高等学校中英文图书数字化国际合作计划
CHINA-US MILLION BOOK DIGITAL LIBRARY PROJECT
English Version | 关于 | 帮助
首页 | 高级检索 | 图像检索 | 视频检索 | 书法字检索 | 特色服务
欢迎您，zhangyin98@zju.edu.cn | 个性化首页 | 登出
检 索
规则设定
个人设置
图书收藏
个人书签
保 存
修 改
删 除
完毕
可信站点

Fig 3 the user rules setting interface

2. $S = S \bigcup \left(\forall \left(S_p, S_q, \ldots, S_r \right) \left(S_p \bigcap S_q \bigcap \ldots \bigcap S_r \right) \right)$, and $\left(s_p \bigcap s_q \bigcap \ldots \bigcap s_r \right)$ is one of the combinations of n sets out of m. The words in the S are sorted in decreasing order of mutual information.

3. If $|S| \geq N$, let the top N words in S be the refinement words and stop.

4. Otherwise, let $n = n - 1$, continue step 2.

This method can improve the recall rate of the search and provide disambiguation information for ambiguous query word.

B Relevance Feedback Implementation

When a user needs to find information regarding a particular topic he or she starts searching by typing keywords and click the search button on the entry page of web site. A connection to a search engine is established and the results of the first search are displayed to the user. These results are the first 20 records of the digital books with their title, authors, and the link to the full text displayed. Our algorithm never modifies the first user query, since the user feedback is not yet available and we do not predefine a set of document as relevant. Instead, the relevant and non relevant contexts are built on the fly for each search session for each user. These contexts are based on the titles, descriptions and other metadata of the digital book. The system attempts to expand each second and subsequent query.

A search session is a set of consecutive searches by a user to find information on the one or more topic. When a user follows a link to a digital book, the title and other metadata was categorized as relevant, and the links not followed are categorized as non-relevant. The category having followed links contains the implicit positive feedback and represents the relevant context. The user keywords are also added to this category. The category having ignored links contains the implicit negative feedback and represents the non relevant context. The words and their occurrence frequency are retained for both contexts separately. Since user engage in the multiple searches, the contexts change with every search, so both are continuously updated during a search session. We use the implicit feedback to differentiate the relevant and non relevant digital books in the top-ranked results. Implicit feedback often increases the proportion of relevant and non-relevant contexts. Xu and Croft [18] found that the proportion of relevant documents in the top-ranked documents affects the results, with a higher proportion resulting in better performance.

The trigger pairs of the user query are to be replaced by the most frequent words in the relevant context. If the number of results of modified query is fewer than 10, the original user query is used to search the relevant books.

V. RULE-BASED PERSONALIZATION SERVICE

Fig 3 shows the individual-customized rules setting interface through which individuals can specify the Boolean combination of the specific rules. One rule tells the system that it should filter the digital items according to the keywords that are found in the title, subject, creator, publisher, or description of digital items. On entering the entry of personal portal, the individual can see the filtered results based on the combination rules. In the middle area of fig 2, the number of filtered results of each combination is limited up to 20. When user specifies the multiple kinds of combination of rules, the portal displays the filtered results of each combination in the top-down order. In the current implementation, AJAX techniques [22] are used to asynchronously to communicate to the web application server. Therefore, the sooner the server completes the filtering process according to one combination, the sooner the results of this filtering are displayed. When the user encounters the favorite items or stop reading at some place in the book being browsed, it can indicate the portal to remember the favorite books and bookmarks in the book being browsed, which are both displayed in the entry of personal portal.

VI. PERSONALIZATION SERVIC BASED ON INFORMATION FILTERING TECHNIQUES

In CADAL portal, a content-based filtering method is implemented. User profile is represented as a vector of indicative keywords extracted from the contents of all digital books. When users return more and more relevance feedback, the recommender system will retrain the discriminative model to obtain the right model parameter to better predict the interest of user to item. Besides relevance feedback, users can answer the questionnaires of rating the digital item. Users can select one of five ranks with one the worst and five the best. A hybrid collaborative filtering method is implemented to learn a cluster model from the rating data to predict the preference of users to unseen items.

A Content-based filtering method

The purpose of user profile learning is to find a classifier with least generalization error on future data using the training data available, thus the answer usually depend on the data set. At the early stage of filtering, we collect the very few training data, thus a low variance algorithm that is insensitive to the training example could be a better choice. When enough data are collected through the interaction with the user later, a low bias algorithm that closely approximates the best solution may work better. We implemented the LR_Rocchio algorithm [7] that combines the classical Rocchio algorithm[19] and logistic regression statistical algorithm in order to handle the whole filtering process.

At a certain point in the adaptive filtering process, suppose that we have t training documents with user's indication of relevance.

$$D_t = (X,Y)_t = \left[(x_1,y_1),(x_2,y_2),\ldots,(x_t,y_t)\right]$$

Where x_i , $i = 1$ to t , is a vector that represents the relevant and non-relevant documents indicated by users in a K dimensional vector space indexed by K keywords. $y = 1$ if the document x is relevant, otherwise $y = -1$. The recommender system recommends the top-ranked items sorted in decreasing order of the posterior probability of relevance of document x based on the training data: $P(y = 1 \mid x, D_t)$.

A widely used profile updating methods in the information retrieval community are the different variations of the increasing Rocchio algorithm, which can be generalized as:

$$Q' = \alpha \cdot Q + \beta \frac{\sum_{x_i \in R} x_i}{|R|} - \gamma \frac{\sum_{x_i \in NR} x_i}{|NR|} \qquad (2)$$

Where Q is the initial profile vector, $Q' = (w_{r1},\ldots,w_{rk})$ is the new profile vector, R is the set of relevant documents, and NR is the set of non-relevant documents.

When filtering, the Racchio algorithm only provides a score indicating how well the document matches the user profile. The score is calculated by measuring the distance between the document vector and the user profile Q' using the cosine formula. The system will deliver the document x to the user if and only if its score is above the dissemination threshold, the decision rule is: $(w_{r1},\ldots,w_{rk})\, x \geq \textit{threshold}$

Let w_{r0} =-threshold, $w_R^T = (w_{r0}, w_{r1},\ldots,w_{rK})$. We make x a new $K+1$ vector with the first dimension corresponding to a pseudo-feature of constant value 1, the above equation can be rewritten as:

Deliver if and only if $w_R^T x \geq 0$

The Rocchio algorithm is a simple heuristic algorithm that empirically works well.

Logistic regression is one widely used statistical algorithm that can provide an estimation of posterior probability $P(y \mid x)$ of an unobserved variable y given an observed variable x . A logistical regression model estimates the posterior probability of y via a log linear function of observed document x :

$$P(y = \pm 1 \mid x, w) = \frac{1}{1+\exp\left(-y w^T x\right)} \qquad (3)$$

where w is the K dimensional logistical regression model parameter learned from the training data.

The Bayesian-based learning algorithms often begin with a certain prior belief $p(w)$ about the distribution of the logistic regression model parameter w .A Gaussian distribution $p(w) = N(w; m_w, v_w)$ is often used as the prior distribution of the logistic regression weights, where m_w is the mean of the Gaussian distribution in the K dimensional parameter space and v_w^{-1} is the $(K+1)\cdot(K+1)$ covariance matrix of the Gaussian distribution. If all items in the covariance matrix v_w^{-1} are zero, $p(w)$ is a non-informative prior: all values of w have the same probability. A classifier learned with a non-informative prior usually over fits the training data. The LR_Rocchio algorithm set a Bayesian prior of the logistic regression model parameter using the user profile calculated by Rocchio algorithm.

Let $w_R^T = \left(w_{r0}, w_{r1}, w_{r2}, \dots, w_{rk}\right)$ be the profile vector calculated by the Rocchio algorithm. We use the same representation for logistic regression as Rocchio for documents: the same set of keywords with the same weighting schema, plus a pseudo-dimension, which is always 1. The probability of relevance of a given document x based on logistic regression model w is:

$$P\left(y=1 \mid x, w\right) = \frac{1}{1+\exp\left(-w^T x\right)} \qquad (4)$$

Since the goal is to minimize classification error, the system using the logistic regression will deliver the document x if and only if $w^T x \geq 0$.

A Gaussian distribution $N\left(m_w, v_w^{-1}\right)$ for logistic regression encodes the belief that the true decision boundary is around the one defined by m_w . Instead of setting $m_w = \left(0, 0, \dots, 0\right)$, we set m_w the same as the boundary found by Rocchio algorithm, which is better than the commonly used non-informative prior or zero mean Gaussian prior. A prior m_w that encodes Rocchio's suggestion about decision boundary can be learned via constrained maximum likelihood estimation:

$$m_w = \arg\max_w \sum_{i=1}^{t} \log\left(\frac{1}{1+\exp\left(-y_i w^T x_i\right)}\right) (5)$$

Under the constraint: $\cos\left(w, w_R\right) = 0$

The resulting logistic regression parameter m_w maximizes the likelihood of the data under the constraint that it corresponds to the same decision boundary as the Rocchio algorithm. The solution is in a simple form that can be calculated efficiently:

$$m_w = \alpha^* \cdot w_R$$

Where α is a scalar:

$$\alpha^* = \arg\max_\alpha \sum_{i=1}^{t} \frac{1}{1+\exp\left(-y_i \alpha w_R^T x_i\right)} \qquad (6)$$

The solution can be found quickly using gradient descent algorithm.

In practice, the Rocchio algorithm influences heavily on the logistic regression algorithm at the early stage of relevance data collection. The LR_Rocchio algorithm automatically manages the trade-off between bias and variance based on the amount of the collected data.

B collaborative filtering method

The major approaches of collaborative filtering are classified in two kinds: memory-based and model-based approaches. The basic idea of memory-based approaches is to compute the active user's predicted vote of an item as a weighted average of votes by other similar users or K nearest neighbor. The Pearson Correlation Coefficient (PCC) algorithm [20] is the one of the most popular memory-based approaches. Two popular model-based algorithms are the clustering for collaborative filtering and aspect models [21]. Clustering techniques work by identifying the groups of users who appear to have similar preferences. Once the clusters are created, predictions for an individual can be made by averaging the options of the other user in that cluster. We implemented a hybrid model algorithm using the cluster-based smoothing [8]. The algorithm is:

1. Create the user clusters C using the k-means method.
2. Given the user u_a , and i rated items, an item t and an integer K , the number of nearest neighbors. Choose s users into G from groups that are most similar to user u_a .
3. Calculate similarity $sim\left(u, u_a\right)$ for each u in G in which the rating of the user u is the combination of $R_u\left(t\right)$ and $R_{Cu}\left(t\right)$.
4. Select the top-K most similar users as neighbors.
5. Predict the rating of the item t for u_a by the behaviors of the K nearest neighbors.

Let $T = \left\{t_1, t_2, \dots, t_m\right\}$ be a set of items, $U = \left\{u_1, u_2, \dots, u_n\right\}$ be a set of users. u_a is the user for whom we provide the recommendations for items that user has not seen before. Let the $\left(u_1, i_1, r_1\right), \dots, \left(u_k, i_k, r_k\right)$ be all the ratings that users give. Each triple $\left(u_i, i_i, r_i\right)$ indicates the item i_i is rated as r_i by the user u_i . For each user u , $R_u\left(t\right)$ denotes the rating of item t by user u and $\bar{R}_u$ denotes his average rating. Assuming that users could be clustered into N groups, then the clustering results of the users $U = \left\{u_1, u_2, \dots, u_n\right\}$ are represented as $\left\{C_u^1, C_u^2, \dots, C_u^n\right\}$, the Pearson correlation-coefficient function is taken as the similarity measure function. The similarity between user u and user u' is defined as :

$$sim_{u,u'} = \frac{\sum_{t \in T(u) \wedge T(u')} \left(R_u(t) - \bar{R}_u\right) \cdot \left(R_{u'}(t) - \bar{R}_{u'}\right)}{\sqrt{\sum_{t \in T(u) \wedge T(u')} \left(R_u(t) - \bar{R}_u\right)^2} \sqrt{\sum_{t \in T(u) \wedge T(u')} \left(R_{u'}(t) - \bar{R}_{u'}\right)}} \qquad (7)$$

At the early stage of system running, the collected rating data is sparse. To fill the missing values in data set, clusters are explicitly exploited to smooth the sparse data. Based on the clustering results, we apply the smoothing techniques to the unseen rating data. Let's first define a special rating value as follows:

$$R_u(t) = \begin{cases} R_u(t) & \text{if user u rate the item t} \\ \hat{R}_u(t) & \text{else} \end{cases}$$

Where $\hat{R}_u(t)$ denotes the smoothed value for user u 's rating to the item t .

Given a user u , $C_u \in \{C_u^1, C_u^2, \ldots, C_u^n\}$ refers the cluster user u belongs to. The following equation is used to calculate $\hat{R}_u(t)$:

$$\hat{R}_u(t) = \bar{R}_u + \Delta R_{Cu}(t)$$

Where $\Delta R_{Cu}(t)$ is average deviation of rating for all users in cluster C_u to item t , which is defined as:

$$\Delta R_{Cu}(t) = \sum_{u' \in C_u(t)} \left(R_{u'}(t) - \bar{R}_{u'}\right) \Big/ \left|C_u(t)\right| \tag{8}$$

Where $C_u(t) \in C_u$ is the user set in user cluster C_u that have rated item t. $\left|C_u(t)\right|$ is the number of users in cluster C_u who have rated the item t.

We make use of the user cluster to limit the number of neighbors similar to the user in preference to increase the system scalability. The centroid of cluster is represented as the average rating over the cluster. The similarity between the cluster C and user is defined as:

$$sim_{u_a,C} = \frac{\sum_{t \in T(u_a) \wedge T(C)} \Delta R_C(t) \cdot \left(R_{u_a}(t) - \bar{R}_{u_a}\right)}{\sqrt{\sum_{t \in T(u_a) \wedge T(C)} \left(\Delta R_C(t)\right)^2} \sqrt{\sum_{t \in T(u_a) \wedge T(C)} \left(R_{u_a}(t) - \bar{R}_{u_a}\right)^2}} \tag{9}$$

After calculating the similarity, the users in the most similar cluster are taken as the candidates that need to be recalculated similarity with the active user on the smoothed data. After smoothing, the rating data contains two parts: original data and group data. The different weights are placed on the two parts when calculating the similarity between the cluster users and the active user. The confidential weight w_{ut} for the user u to item t is defined as:

$$w_{ut} = \begin{cases} 1-\lambda & \text{if user u rate the item t} \\ \lambda & \text{else} \end{cases}$$

Where λ is the tuning parameter between original rating and group rating, its value varied from 0 to 1. The system will select the top K most similar users based on the following similarity function:

$$sim_{u_a,u} = \frac{\sum_{t \in T(u_a)} w_{ut} \cdot \left(R_u(t) - \bar{R}_u\right) \cdot \left(R_{u_a}(t) - \bar{R}_{u_a}\right)}{\sqrt{\sum_{t \in T(u_a)} \left(w_{ut} \cdot \left(R_u(t) - \bar{R}_u\right)\right)^2} \sqrt{\sum_{t \in T(u_a)} \left(R_{u_a}(t) - \bar{R}_{u_a}\right)^2}} \tag{10}$$

After the neighbor selection, a weighted aggregate of the deviations from the neighbor's mean is used to generate the prediction for the active user as the following:

$$R_{u_a}(t) = \bar{R}_{u_a} + \frac{\sum_{i=1}^{K} w_{ut} \cdot sim_{u_a,u_i} \cdot \left(R_{u_i}(t) - \bar{R}_{u_i}\right)}{\sum_{i=1}^{K} w_{ut} \cdot sim_{u_a,u_i}} \tag{11}$$

VII. Conclusion

In this paper, we have described the personalization services currently implemented in the CADAL portal, and the techniques to build the respective personalization service. We build the CADAL portal as a testbed of versatile information filtering methods based on the contents of items, user ratings and customized rules.

In the future, the architecture of personalization services is to be extended to incorporate the ontology techniques like WordNet®. WordNet® is an online lexical reference system in which nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. As portal runs, we will put more effort on the web usage mining techniques to discover the user pattern from the web data.

Acknowledgement

The work has been supported by China-US Million Books Digital Library project.

References

[1] ST.Clair Gloriana, 2005, "Million Book Project vs Google Print," *Journal of Zhejiang University Science*, vol.6A,No.11,pp 1195-1200, November 2005.

[2] T.J.Huang, Y.H.Tian, C.L.Wang, X.D.Shi, W.Gao, 2005, "Towards a multilingual, multimedia and multimodal digital library platform," *Journal of Zhejiang University Science*, vol.6A, No.11, pp 1188-1192, November 2005.

[3] Dolog, P. Henze, N. Nejdl, W. and Sintek, M. 2004. Personalization in distributed e-learning environments. In *Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters* (New York, NY, USA, May 19 - 21, 2004).

[4] Eirinaki, M. and Vazirgiannis, M. 2003. Web mining for web personalization. *ACM Trans. Inter. Tech.* 3, 1 (Feb. 2003), 1-27.

[5] Leroy, G., Lally, A. M., and Chen, H. 2003. The use of dynamic contexts to improve casual internet searching. *ACM Trans. Inf. Syst.* 21, 3 (Jul. 2003), 229-253.

[6] Chen, L. and Sycara, K. 1998. WebMate: a personal agent for browsing and searching. In *Proceedings of the Second international Conference on Autonomous Agents* (Minneapolis, Minnesota, United States, May 10 - 13, 1998)

[7] Zhang, Y. 2004. Using bayesian priors to combine classifiers for adaptive filtering. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004).

[8] Xue, G., Lin, C., Yang, Q., Xi, W., Zeng, H., Yu, Y., and Chen, Z. 2005. Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Salvador, Brazil, August 15 - 19, 2005).

[9] SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., AND TAN, P.-N. 2000. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*, 2 (Jan.), 12–23.

[10] The Dublin Core Metadata Initiative. http://dublincore.org/

[11] Semantic Markup for Web Services. http://www.daml.org/services/owl-s/1.1/

[12] W3C. Web Service Description Language (WSDL) Version 2.0 Part1: Core Language. http://www.w3.org/TR/2006/CR-wsdl20-20060327/

[13] W3C. Soap Version 1.2 part 0: Primer http://www.w3.org/TR/2003/REC-soap12-part0-20030624/

[14] Brickley, D. , and Guha,R.V., Resource Description Framework (RDF) Schema Specification 1.0,2004 http://www.w3.org/TR/rdf-schema/

[15] Magennis, M. and Rijsbergen, C.J.V. 1997. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 342-332

[16] Belkin, N.J., Cool, C., Head,J., Jeng, J, Kelly,D. , Lin,S., Lobash,L., Park,S.Y., Savage-knepshield, P., and Sikora, C. 1999. Relevance feedback versus local context analysis as term suggestion devices: Ruters' TREC-8 interactive track experience. In *Proceedings of the Eighth Text Retrieval Conference*(TREC 8, Gaithersburg, MD).565-573.

[17] White, R.W., Jose, J.M., and Ruthven, I. 2001 Comparing explicit and implicit feedback techniques for web retrieval: TREC-10 interactive track report. In *proceedings of the 10th Text Retrieval Conference* (TREC 2001, Gaithersburg, MD)

[18] Xu, J. and Croft, W.B. 1996. Query expansion using local and global document analysis. In *proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Finland) . 57-64.

[19] Ault, T., Yang, Y., 2001 KNN, Rocchio and metrics for information filtering at TREC-10. In *Proceeding of the Tenth Text Retrieval Conference (TREC-10).* National Institue of Standards and Technology, special publication 500-225,2001

[20] Resnick, P., Iacovou, N., Suchak, M., Bergstrom,P., and Riedl, J., 1994, Grouplens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp 175-186, 1994.

[21] Hofmann, T., and Puzicha, J., 1999, Latent Class Models for Collaborative Filtering . In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp 688-693, 1999.

[22] Garrett, J.J., Ajax: A New Approach to Web Applications, http://www.adaptivepath.com/publications/essays/archives/000385.php