Journal of Computational Information Systems 8: 20 (2012) 1–8
Available at http://www.Jofcis.com

User Personalization Mechanism in Agent-based Meta Search Engine ⋆

Qingshan LI*, Yanxin ZOU, Yingcheng SUN

Software Engineering Institute, Xidian University, Xi'an 710071, China

Abstract

Traditional Meta search engines are not well considered user's interests, unable to provide users with personalized service. This paper presents a user personalization mechanism in Agent-based Meta search engine. The mechanism can create and update user profile by mining user behavior during information retrieval, it can well reflect user's current interests and be able to update the user's interests dynamically. Use user profile in filtering the search results, can make the search results meet user's requirements personalized.

Keywords: Meta Search Engine; Personalized; Agent; User Profile

1 Introduction

Search engine collect numerous web sites to provide global cyber source and retrieval mechanism, it help users to obtain the information they need. Because of its universal nature and the rapid growth of Web information, it still can't meet the query request of different background, different purposes and different periods, and the test of search engine's capability in Internet information indexing is more and more severe, existing search engines have low coverage and low precision on the defect, without considering the user personalization demand, unable to fully meet users' demand for effective information.

Meta search engine is capable of using plenty of member search engines, improves the search coverage; personalized service technology provide different users with different service, to meet different needs, it through the collection and analysis of users information to learn the user interest and behavior [1], while personalized service technology applied to search can satisfy user's individual needs, improve the search precision. Personalized Meta search engine combines the two advantages of two kinds of technology, is the effective way to improve the existing search engine.

⋆Project supported by the National Natural Science Foundation of China (No. 61173026), the National High Technology Research and Development 863 Program of China (2012AA02A603), the Fundamental Research Funds for the Central Universities of China, and the Defense Pre-Research Project of the 'Twelfth Five-Year-Plan' of China (513***301).

*Corresponding author.
Email address: qshli@mail.xidian.edu.com (Qingshan LI).

October 15, 2012

Agent technology is an important research direction of artificial intelligence and software engineering, which has been widely used. The application of Agent in Web information retrieval to provide users with personalized information, is based on the characteristics of Agent [2]: the autonomy of Agent allows decision-making mechanism to decide what kind of behavior to do; the learning of Agent makes Agent can learn user interest to establish user model; the society of Agent makes Agent can communicate with other Agent, provide better services to users.

User personalization mechanism of intelligent retrieval has the bright application prospect, so it attracts enterprises and research institutions to do research actively. Literature [3] through the research on user macro demand in 1999, put forward an meta search engine structure based on users' demand; 2001, literature [4] proposed the intelligent agents to achieve the meta search engine method; literature [5] first proposed in combination with Agent adopt interest spanning tree to present user interest model; 2002, literature [6] put forward to use ontology technology to record the user search intent, and combine with Agent technology, complete the related intelligent processing; in 2006, literature [7] put forward forgetting algorithm to update user interests, and the interest is divided into long-term interest and short-term interest; 2008, literature [8] establish a more detailed model on the ontology and Agent. But at present the ontology's form there is still no consensus; literature [9] using Agent technology and information filtering technology to improve the intelligence of search engine; in 2010, literature [10] presents a thought which to predict the future behavior through analysis of user behavior when using a search engine.

This paper do research on the core of users personalization of meta search engine, the user profile, based on analysis of meta search engine and individualized techniques, combined with Agent technology applied in intelligent meta search system user personalization. The integration of Agent technology and Meta search engine technology improve the intelligence of meta search engine. The main contributions of this paper is proposed a new user personalized mechanism. The other part of this article is organized as follow: The first section introduces the theoretical model of Meta search engine personalization mechanism based on Agent. Second section describes the user profile and user query analysis in detail. The third section analyzes the performance of the model. Section fourth concludes the paper and prospects for the future work.

2 User Personalized Mechanism

The research of this paper, the Agent-based search engine user personalization mainly from the following two aspects: (1) According to user's personality traits, and users' retrieved history, establishing user personalization model, the user profile. The Meta search engine can provide personalized search service according to the user's interest only after the user profile is established. The user profile is basic for Meta search engine realize the personalized search. (2) User query is usually ambiguity, and user's real intention is most directly reasoning by member engines. But the different member engines have different reasoning mechanism, which can lead Meta search engine to return a large number of irrelevant results to users. Therefore, in order to offer users personalized search, Meta search engine need to optimize user's query, and reference to user interests model to identify users query intentions as possible.

Since different users have different requirement, the search information model has become increasingly difficult to adapt to the rapid growth of the Internet information resources. To know users' search intent, we must first understand users' personality characteristics, so user profile is the core of personalized search. Because every time users are expected to get their

own need simply and quickly, and do not want too much interaction, judgment or feedback in search process, thus most search engine mining the history of user access history to obtain users description information. Users access history are usually not very rich, or even difficult to obtain, this involves search engine access mode limitations, the search engine processing technology ability at present and the individual user privacy safety etc. The author proposed a kind of Agent-based intelligent Meta search engine user personalized model, which according to analysis user browsing history, mining the potential user behavior characteristics and potential interest, then construct user interest model. Its structure can be divided into six parts as shown in Fig. 1.

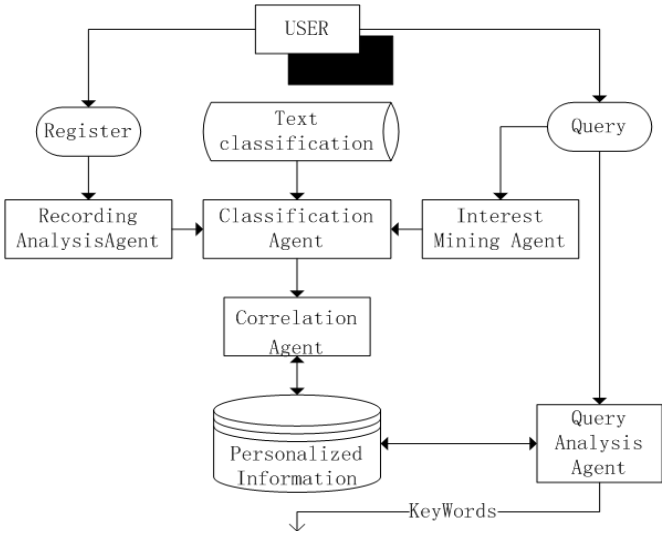

Fig. 1: Agent-based meta search engine user personalized model

- Text classification corpus exists on the server, is a model for constructing and updating user interest model. In practical application of the system, the classification Agent use text classification corpus to classify web documents.

- Recording analysis Agent start when user register intelligent meta search engine, its function is to dig historical information on user' browser, get user current interest related webpages, give classification Agent to do next processing, creation of user interest model.

- Interest mining Agent start when the user uses the intelligent meta search system, recorded the user actions after the system given query results, according to user action, such as access time, collection, download page giving different weights. In which mining user most interested webpages, handed over to the classification Agent, used to update the user interest model.

- Classification Agent is the core of user personalization, its function is to analyse the webpage links which recorded by record Agent or interest mining Agent. Grasping these links page, dig inside keywords, is used to create or update the user interest model. First, Agent will page segmentation, using the improved naive Bayes algorithm to classify different pages, mining the same category pages, then according to a given category accounted for the total number of documents arrangement in number of words, according to the TF get the keywords, keywords as the interest of the user word stored.

- Correlation analysis Agent mining in the category the word appears when get a particular word, association rule mining algorithm. For a given key words, system extended related words for association expansion, another function is using a forgetting factor to update the user interest model.
- Query analysis Agent search word is present in the user interest model when user search, then expand the word, then handed over to the following Agent, for further processing.

3 User Profile and Query Intent Analysis

Agent-based intelligent Meta search engine user personalized mainly around the following two points: the establishment and update of user profile, the user query intent analysis. After the research of the typical interest model, we designed a new user profile which can forgetting learning, update and optimization the user interests.

3.1 User profile model

The user profile we designed is composite of multiple interest feature vector, each user interest vector is present by a four tuple representation, such as Eq. (1):

$$k_i, c_i, d_i, w_i \tag{1}$$

k_i indicates word, i.e. the user's interest word; c_i indicates classification; whereby the category which user interest word belongs to; d_i indicates update date, mainly in order to facilitate the forgetting factor calculate weights of the words; w_i indicates weight, the value when it last update.

Amount of user interest vocabulary is limited, it is better present the user interested words, literature [5] points out that the number of user interest feature word between 40 to 60 is more appropriate. The classification of interest words is determine by its document category, the advantage is good for polysemy better distinction.

This paper uses the improved naive Bayes algorithm based on word frequency to classify documents. Naive Bayes algorithm is the most widely used one of two kinds of classification model. Naive Bayesian classification model estimates the parameters required for small, less sensitive to missing data, the algorithm is relatively simple. Naive Bayes algorithm runs need training base, here the text classification corpus of Sogou lab, the nine categories are: IT, finance, health, sports, tourism, education, recruitment, culture and military, each category contains 1990 documents. Because the author is based on the use of the word frequency of the naive Bayesian classifier, so first of all to the text classification corpus preprocessing, in accordance with the classification of the extraction of entries and number, and stored as a naive Bayesian classification based on word frequency training base. Eq. (2) is a universal naive Bayes formula.

$$C_{NB} = argMax(P(c_j) * \prod_{1}^{c} P(x_i|c_j)) \tag{2}$$

C_{NB} indicate the classification of document $P(c_j)$ is the Prior probability of class j, $P(x_i|c_j)$ is the class conditional probability of characteristic quantities x_i which belongs to class c_j. Find the

probability of document in all categories; take the maximum as the classification.

$$P(c_j) = \sum_{k=1}^{v} TF(X = x_i, C = c_j) / \sum_{m=1}^{W} \sum_{k=1}^{V} TF(X = x_k, C = c_m) \tag{3}$$

Eq. (3) is prior probability formula based on the frequency of naive Bayes, v indicates the number of feature words, i.e. the number of words in class c_j of training set, $TF(X = x_i, C = c_j)$ represents the number of attribute x_i appeared in class c_j, indicates number of all classes. The type first respectively calculates a classification of all the features of word frequencies and, then value is divided by all the classification characteristics of word frequency and the quotient, as classified by the prior probability.

$$P(x_i|c_j) = (TF(X = x_i, C = c_j) + 1)/(v + \sum_{k=1}^{v} TF(X = x_k, C = c_j)) \tag{4}$$

Eq. (4) is class conditional probability formula based on word frequency of naive Bayes, +1 in the formula is to prevent the class conditional probabilities become 0. The class conditional probability here is improved; consider a feature in the classification of documents in many cases, the improved formulas such as the Eq. (5).

$$P(x_i|c_j) = TF(k_i, d) * (TF(X = x_i, C = c_j) + 1)/(v + \sum_{k=1}^{v} TF(X = x_k, C = c_j)) \tag{5}$$

$TF(k_i, d)$ is the number of keyword k_i from unclassified document appeared in document d. We calculate the interest term weight in user interest model method is also adopted $TF(k_i, d)$, namely TF algorithm. The method for calculate the weight of Key words in document include Boolean algorithm, word frequency and $TF - IDF$ algorithm, in the general case, $TF - IDF$ has the highest accuracy, but $TF - IDF$ algorithm is guaranteed to appear low frequency words with higher weights, this conclusion is in the number of types of document classification income, webpage links are crawling the document first classified in the mechanism we design, therefore, keyword weights are in their category of document sets are calculated, this does not apply to $TF - IDF$ algorithm. So we use the TF algorithm to calculate weights, calculation formulae such as Eq. (6):

$$w_i = TF(k_i, d) = \sum_{d=1}^{n} \frac{f_{id}}{\sum_{v=1}^{m} f_v} \tag{6}$$

w_i is the weight of keyword, k_i indicates keyword, d is the document the word belong to, the range of d is from 1 to n. f_{id} is the number of keyword wi appeared in document d. m is the number of words in document d, from 1 to m, f_v indicates the number of word v appeared in document d, so $\sum_{v=1}^{m} f_v$ indicates the number of all words appeared in one document. The algorithm first computes number of occurrences of a keyword in a document, then divided by the number of documents all entries appear and get the key words in this document frequency, then add all the keywords word frequency of other document from the category the key word belong to, finally get the keyword weight value. Because the user interest model space is limited, so users should model key words are adjusted to suit the user interest, changes over time, and ensures that users

pay close attention to most current period interest term weight value to maintain the highest. So here introduce a forgetting factor [7], such as Eq. (7):

$$F(k_i) = e^{-\frac{log2}{hl}(now-created)} \tag{7}$$

now represents the current date, created shows the last updated date of an interest word in interest model, *hl* indicates days of half-life, user's interest is forgotten half after *hl* days, we selected 7 days as its half-life. By introducing the eigenvalue algorithm, can effectively achieve the dynamic update user interest.

3.2 Analysis of user queries

The system also use word association to expand associated interested words as query optimization method, we usually use multiple words to express the query intent when using search engines to get better results.

Association rules mining process consists of two stages: The first stage must identify all high-frequency project group from data collection, high frequency means a project group the frequency relative to all records, must reach a certain level. A k-itemset which satisfying the minimum support degree is known as frequency K-project group, generally expressed as Large K or Frequent K. Algorithm produce Large k+1 from the Large K project group, until it can no longer find longer high-frequency project group. The second stage of association rule mining is to generate Association rules. From the high-frequency project group produces association rules, is the use of a previous step frequency K project group to generate rules on minimum confidence threshold under the conditions, if a rule obtained the trust to meet the minimum confidence, call the rules for association rules.

The minimum support degree of association rules is set to 0.4, the minimum confidence is set to 0.8. In order not to affect the system performance, the design requirements in association rule up to three related words can be extended.

4 Experimental Results Analysis

In order to verify the user personalized mechanism, we design an Agent-based user personalized program, using the example establishing and updating of user profile as well as the query intent analysis. The experimental results show that the proposed mechanism is effective. First part of the experiment, we drew 50 webpage links as the user history, its around three main themes:

- Apple's mobile phone.
- The ancient Chinese text oracle.
- The United States developed automatic rifle M16.

The three themes in the use of other search will appear irrelevant information, such as query Apple will appear the fruit apple. So selecting the three related topics to be tested is able to show the user personalization module good function and performance. The number of user interest words is 40. Apple here category belongs to class IT, so on Apple mobile phone theme

webpage are classified as IT, so the key words in the webpage is classified as class IT. According to IT on the web in proportion, the system assigned to class IT 12 interest words. All interested word update date is agreed, in order to facilitate the comparison of weight, will all interested word update date for the same day, the corresponding weight will be multiplied by the forgetting factor to get the new weight value.

The words of the apple theme, according to the several maximum weighted were: iPhone, mobile phone, iPhone4, apple. This four word association expansion respectively: apple mobile, iPhone apple, phone, iPhone apple. Oracle theme keywords, were classified as culture, we can see the highest weight words respectively: Oracle, Chinese, writing, calligraphy. The M16 theme key words, highest weight words: M16, rifles, assault. As a result, in the user history analysis, system based on user interest classification is very accurate, and the words obtained with high weight significantly expressed interest of the user, and can be extended out the relevant word, has reached the design requirement.

In the user model creating test, system analysis on 50 webpage user history, capture page, segmentation, classification and association word takes a total of about 10 minutes, the time can meet the design requirements, but also need to further improve the efficiency of prototype.

The extended term effects on the keywords on popular search engines are tested in second part of the experiment. We compare the relevance of search result of the expanded words with not expanded ones. The first 50 records of the result search engine returns are being selected in the experiment.We choose Baidu, Google, Bing, Youdao, Soso, the five commonly chinese used search engine's results to test, according to three topics as shown in Fig. 2 to get the correlation.

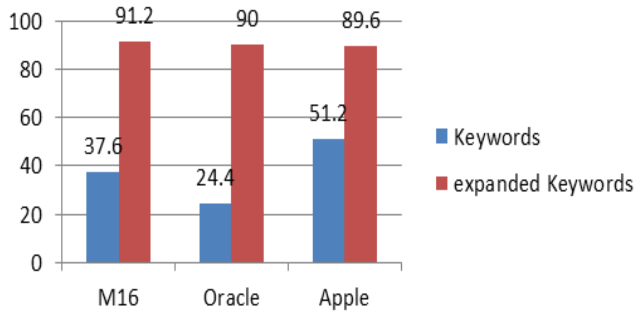

Fig. 2: Correlation contrast

Alternative keywords are expanded later got the correlation of search results greatly improved, can satisfy the demand of users. In the experiment, we also noted the relevance of search results in the expansion of the ambiguity is not strong keywords to improve is not obvious.

5 Conclusions

Based on Agent technology, mainly on the user interest model constructing and updating, as well as the user query analysis and extended to study search engine users personalized key technology. On the personalized search technology is not very mature, a lot of work is still in the research stage, so there is no perfect evaluation mechanism and large-scale application, so continue to conduct study has very important significance. This paper has made some achievements, but still need to be further improved. Future work focus:

(1) Update the classification of reference database, collecting the authority of the training data to improve the reference model.

(2) The time complexity of the design is the bottleneck of Web information retrieval, but also need to improve the algorithm, improve the efficiency.

(3) Consider the synonyms and the implication of words when user query, do optimization on the query and display the results.

(4) Considering the mass usage, for optimization of large-scale concurrent query.

(5) Mobile platform development like a raging fire could also be considered to be ported to mobile platform system.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (61173026), the National High Technology Research and Development 863 Program of China (2012AA02A603), the Fundamental Research Funds for the Central Universities of China, and the Defense Pre-Research Project of the 'Twelfth Five-Year-Plan' of China (513***301).

References

[1] C. Zeng, C. X. Xing, L. Z. Zhou. A survey of personalization technology: Journal of Software, 13 (10), (2002), 1952 – 1961.

[2] J. P. Xu, Y. Q. Zhai. Personalized information service based on Agent technology research: Computer engineering and Science, 24 (3), (2002), 73 – 76.

[3] E. J. Glover, S. Lawrence, W. P. Birmingham. Architecture of a metasearch engine that supports user information needs, in: Proceedings of the eighth intemational conference on Information and knowledge management, 1999, pp. 210 – 216.

[4] L. Kerschberg, W. Kim, A. Scime. A Semantic Taxonomy-Based Personalizable Meta-Search Agent, in: Proceedings of the Second International Conference on Web Information Systems Engineering (WISE'01), 2001, pp. 41 – 50.

[5] W. F. Zhang, B. W. Xu, L. Xu. The use of Agent personalized search results: Micro computer system, 22 (6), (2001), 724 – 727.

[6] L. Kerschberg, W. Kim, A. Scime. Intelligent Web Search via Personalizable Meta-search Agents, in: CoopIS, DOA, ODBASE, 2002, pp. 1345 – 1358.

[7] K. Xu, Z. M. Cui. Search based on the history of the research on the user interest model: Computer technology and development, 16 (5), (2006), 18 – 20.

[8] C. Kim, J. Y. Jung, H. C. Zin. An Application of Meta Search Agent System Based on Semantized Tags for Enhanced Web Searching: Journal of Universal Computer Science, 14 (14), (2008), 2400 – 2415.

[9] H. M. Li, Z. G. Ding, L. H. Zhou. Agent based intelligent meta search engine technology research: Computer science, 35 (10), (2008), 90 – 93.

[10] R. W. White, A. Kapoor, S. T. Dumais. Modeling Long-Term Search Engine Usage, in: UMAP2010, 2010, pp. 28 – 39.