Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.5, No.12, 2014

A User Behavior Based Study on Search Engine Ranking

Yu Huang [1] Mu Zhang [2*] Xuekui Ge [3]

1. School of Management, Jinan University, 601 Huangpu Avenue West, Guangzhou 510632, China
2. *Shenzhen Tourism College, Jinan University, 6 Qiaochengdong Street, Shenzhen 518053, China
3. Shenzhen Tourism College, Jinan University, 6 Qiaochengdong Street, Shenzhen 518053, China

* E-mail of the corresponding author: zhangmu@163.com

Abstract

In this era of information explosion, finding convenient ways to get the desired information is becoming ever mo re vital today. With a review of the existing information retrieval and feedback technology, this paper puts forward a method to establish and update user profile model through obtaining user's implicit feedbacks. The user's explicit information is not a must. Instead, this method, with the implicit information acquired by observing the behaviors of the users when browsing web pages, establishes and updates the user profile model and thus reduces the workload.

Keywords: Information retrieval；Implicit feedback；Relevance feedback; User profile model

1. Research Background

In this information age, everybody can search a great capacity of information with a computer and internet access. Therefore, it has become increasingly important to help people find their needed information without having to deal with the unwanted one. You may find an answer to a certain question from the web. However, this can not guarantee that the questioner finds the answer instantly. There are several reasons for this. On the one hand, there is no detailed catalogues or indexes. On the other, the way people describe questions vary. Only when the way the information provider describes the target information coincides with how the information searcher describes it can he find the wanted information immediately. Besides, it also involves commercial interests. That is to say, the search engine owners have their propensities in the search results for business benefit considerations, which means that when you are searching a certain subject, the search engine will put the agents who have sponsored at the noticeable places, the so-called "biding rank".

In 1998, researchers like Craig Silverstein analyzed a massive amount of English query log and drew the conclusion that 85% of the internet users view only the first page of the search results[1], which the research conducted by researchers like Yu Jiahui about the Chinese query log also corroborated[2]. Consequently, the collation rules of the search engines (especially on the first page) largely influence the assimilation of the useful information and thus determine the internet users' satisfaction about the search engines. Like other modern technologies, search engines should also be "human-oriented and serving the people". Facing the ever-growing dissatisfaction of the searching results, the search engines will have to improve its service so as to keep their market share and gain a foothold in the fierce commercial competition.

From the early Weighted Term Frequency Ranking Algorithms, Link Analysis Ranking Algorithms to the prevailing ranking bid Ranking Algorithms, all the search engine corporations have endeavored to improve the service quality of the search engines. Nevertheless, according to one survey of ComScore, even the strong performer, Google, can only attain an accuracy rate of 79.6%; others being unsatisfactory.

2. The Shortcomings of Traditional Search Engines

Although the search engines do help us find the useful information, with the rapid expansion of the information resources, people find that the traditional search engines do have their limitation which hinder the effective acquisition of the information.

(1) Low Coverage and Recall Ratio in Information Retrieval

The Internet information resources are largely dynamic, with quantity growing exponentially and content renewing constantly. On average one search engine can only cover about 5%-20% of the total information amount [3]. With this poor coverage, the recall ratio also remains very low. The net users often have to switch between various search engines, wasting much of the time and energy.

(2) Large amount of Noise and Redundant Information, Low Precision Ratio

The traditional search engines did not comprehensively process the search results. As a result, almost every page is teemed with repetitive and unhelpful contents. The users have to pick out the their own needed information from the large capacity of search results.

(3) Great Variance of Information Resources in Different Search Engines

The traditional search engines adopt their own indexing techniques, information gathering techniques and keyword searching techniques. Expectedly, they show divergences with regard to content, scope as well as validity, which bring inconveniences to the users.

(4) Massive Index Database, Slow Update and Long Response time

The index database has expanded continuously with its rapid development. It is a relatively difficult job to manage and maintain a massive index database. Besides, the rapid updates of the web page information often lead to failures in indexing. As the users click on the URL of the search results, they often come across invalid links and can not open the wanted web pages.

(5) Indistinctive Search Model and No Consideration of the Users' Individual Needs

Most of the existing search engines adopt the "one-fits-all" model, which fails to recognize the users' personal interest and likings. Accordingly users with different interests and needs get the same results when they type in the same keywords, using the same search engine, while hundreds of thousands of results pop out, good and bad intermingled, leaving the users puzzled about where to find the wanted information. In addition, the web page information is basically dynamic and users are concerned about it.

To obtain the constant-changing information, the users have to constantly search the same content on the web, wasting much time.

(6) Mechanical and Impersonal Algorithm, Neglecting Users' Feedbacks

Most of the existing search engines adopt the Matching Algorithm based on keyword searching. The aligning of the search results are mostly based on the analysis of the word frequency and links inside the web pages, largely neglecting the users' feedback of the search results.

The problems listed above are mainly brought by the limitations the development of the traditional search engines, which in return inhibit its own development. Aiming at these problems, the author suggests that a Search Engine Ranking Algorithms with a basis on user's behavior analysis make up for the shortcomings of the traditional search engines, resulting in the fulfillment of "the search results perfectly display the user's ideas."

At present, the implicit relevance feedback techniques which aim at improving ranking and realizing personalized searching has become a remarkably active filed abroad. The implicit feedbacks from the research conducted in a controlled environment by Joachims' and others has clearly shown the great value of integrating implicit feedbacks into the ranking process[4]. The author attempts to solve the following problem: how to apply the implicit feedback to improve its search quality in a large-scale operating environment? Although with implicit feedback and ranking algorithms based on web page content, anchor text or link, we can instinctively predicate that through the interaction of users and search engines, it at least will let on some useful information for ranking, however, it has always been an extremely big challenge to estimate user's preferences in real search engine setting because the real users tend to be much "noisier" than they are in the controlled environment.

3. Literature Review

Search result ranking is a very important issue in information retrieval. The most commonly used method is to employ both the comparing the similarity of the searched content and web page content and referring to the general quality of the web page[5,6]. An obsolete search engine may deploy hundreds of features to describe a web page and then use a complicated algorithm to rank the web pages measured by these features. The current search engine, nonetheless, becomes more humane. It more concerns about the searchers' evaluation of the results and tries its best to make the search results consistent with the searcher's aims. The specific operation method is as follows, first the web user makes relevance annotation to a series of web pages where a certain search request is made, which leads to a "golden criteria" to evaluate the effectiveness of different ranking algorithms. This program is implemented by the reduction of over-reliance on the explicit judgment with implicit feedbacks. Implicit feedback has currently become a hot spot in search engine research.

There are several international research groups which evaluated the relationships between implicit feedback and internet users' interests. Among those researches, the reading time and explicit interest rankings are included as the subjects. Morita and Shinoda studied the time spent by Usenet Internet users reading the news, and found that reading time can be used to infer Internet users' interest ratings while reading an article[7]. Konstan also proved that in the GroupLens system they designed reading time is a powerful tool used to speculate Internet users' interest[8]. Oard and Kim studied whether implicit feedback can replace the explicit rating in net users' recommendation system[9]. Then they also suggested a framework to describe Internet users' explicit behavior, which uses two-dimensional portray criteria: First, the purposes hidden behind the explicit behavior and second, the scope of the above behaviors[10].

By gathering and studying the characteristics of the internet users' browsing activities, Goeck and Shavlik distinguish the users' behaviors, as well as classify them in terms of their similarities. They believe that the browsing activity characteristics have correlations with the users' interests in a certain degree[11]. Convincible as the research results are, the sample size is relatively small. Moreover, it does not conduct a contrast test between the implicit feedback and the explicit judgment, which undermines its credibility.

Claypool and others studied how the implicit feedbacks are related to users' interests. They developed so called "Curious Browser", collecting data to study how to explore the explicit judgments of the web pages browsed through implicit interest indicators. They found that the user's explicit interest has a remarkably apparent positive correlation with the time spent on reading one certain web page, the time spent on scrolling the mouse, as well as the total sum of the two while the user, as an individual, his scrolling and clicking the mouse has nothing to do with the explicit interest[12]. Fox and the others' research proved that there is a certain relationship between the implicit feedback and the explicit behavior. They developed a toolkited browser to collect data and set up a Bayesian model to evaluate the implicit feedback and explicit relevance judgment in searching[13].

This research not only studied the common clicking behavior, but also the users' behavior in a broader sense, such as time-on-page, scrolling time and overtype mode. They found that in the retrieval process click-through is the most important individual variable, but the predictive accuracy can be improved through study of more variables. Joachims provides some pretty valuable insights on how to collect implicit feedback, including the introduction of a click-through data based technique to interpret the ranking mechanism[14]. Later Joachims and his colleagues also designed an empirical evaluation mechanism to interpret the user's click-through evidence. Combined with an Eye-tracking study, they can also be applied to correlate the user behavior prediction technique and the explicit ranking technique[15].

4. Design and Construction of the User Behavior Model

In a flood of information, the real search engine user behaviors are probably very "noisy". In a sense, only when the users' behaviors are associated with the explicit correlation judgment and preferences can they be accurately evaluated and predicted. Consequently, we should not so much think of every user as a reliable "expert" as attempt to collect and generalize useful information from the numerous ordinary users' searching behaviors and traces. The author's main idea is to construct and integrate a user behavior model for internet search engines. All user behaviors that meet the following criteria can be studied as the research factor of the model. One is "relevance". A special search behavior obviously is influenced obviously by the relevance of the search results. The other is "noise", that is, the user's irrelevant search results from the random and casual clickings. This method intends to set up a model based on the study of the user's behaviors known with possible deviations. As a result, except some fundamental ones, this model will calculate the deviation characteristics, which are mainly used to measure the deviation between the observed value and the expected value of the search results.

4.1 Clicking Distribution

The clicking of the users upon a certain search result may imply that they think it is very likely to be the target web page. On the other hand, this behavior may also down to a causal click. Generally speaking, the user's individual behavior is "noisy", and accordingly can not be judged with an accurate correlation.

Through an analysis of the users' log from March to May, AOL made a statistic about the top 18 users' clickings. Number of logs collected: 36389567 Number of search items: 19434540[16]

The statistical data of the clickings of AOL's search engine rankings indicate (Figure1 and 2):

(1) The first ranking item accounts for the 42% click-through share while the second place item only accounts for about 12%, 72% lower than the first one. The third ranking item only accounts for 8.44%, 30% lower than

the second one and over 80% lower than the first.

(2) The lower place the item is in the ranking, the smaller the click-through share it has. The 10th place is a little bit higher than the 9th one, the same level with 8th. This is very likely due to the fact that the users scroll to the web navigation part and see that is the last item and then click in.

(3) Nearly 90% of the total click-through is accounted for by the first page of the top ten search results.

(4) From the first page of the search results to the second, there is a drastic decrease in terms of share of clickthroughs.

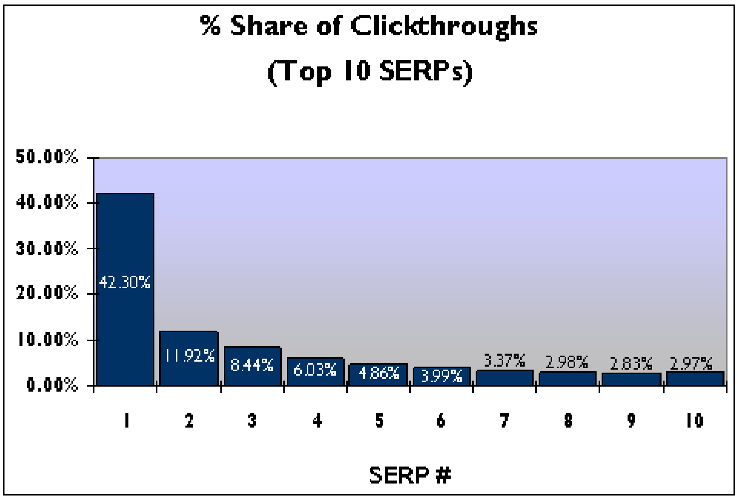


Figure 1. Distribution of Top10 SERPs' Clickthrough Shares

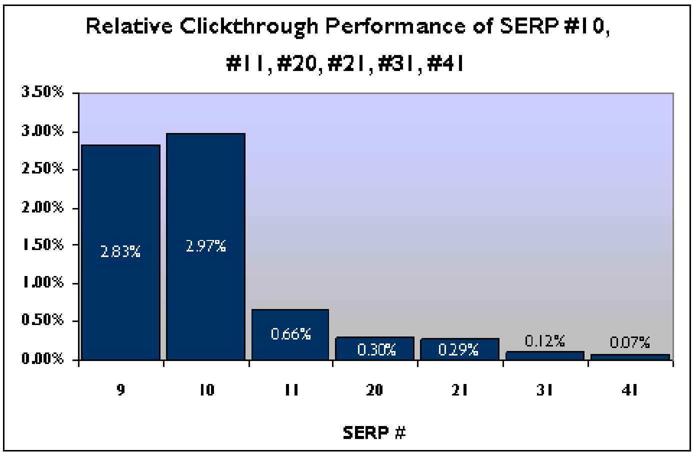
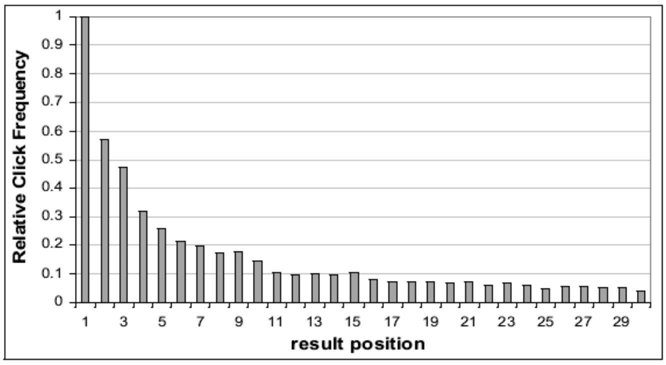


Figure 2. Relative Clickthrough Performace of SERP #9 #10 #11 #20 #21 #31 #41

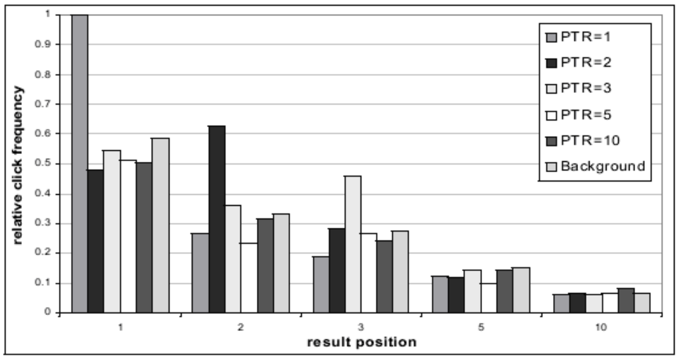
Microsoft Corporation recently selected randomly 3'500 queries from 12'000 search logs over the last 3 weeks and studied the explicit correlation of the top 10 result of each search[17].



Relative click frequency for top 30 result positions over 3,500 queries and 120,000 searches.

Figure 3. Relative Click Frequency for Top 30 Result Positions

Figure 3 shows the relative click frequency for the top 30 result position, i.e. the possibility for the top 30 results of being clicked. To get the aggregate clicks of the P position, it needs to follow steps: first calculate the frequency of clicking P for each search, i.e. to estimate the possibility of clicking P position after any clicks). Then work out the average of these frequencies. The one which comes at the top of the search result positions, its frequency is indicated as 1.



Relative click frequency for queries with varying PTR (Position of Top Relevant Document)

Figure 4. Relative Click Frequency for Queries with Varying PTR

Figure 4 shows the distribution of relative click frequency for queries with varying PTR(Position of Top Relevant Document), i.e., the first search result that user clicks is the user's most recognized and relevant one. If PTR=n, it signifies this most relevant search result ranks No. n place in the search result list.

From Figure 4, we can see that although each time the most relevant search result has the highest click frequency, obviously the No.1 search result set by the search engine gets the universal "peaks". For example, (see result position 2), the most relevant search result ranks the second place (PTR=2), in terms of this search result, the click frequency is higher than any other positions. However, there are still many users the irrelevant results which ranks the first place in the query list, which means that the users have a strong preferences towards the results who have the top positions in the query lists. Even if they are not relevant, they will also be clicked for their top positions.

4.2 A Summary of Internet Users' Behavior Characteristics

The purpose of this study is to design a search engine model which sequences the search results on the basis of users' real needs reflected by massive amount of behavioral characteristics. Once the user submits the search request, they respond and take such actions as reading the reading the abstract, clicking the results, navigating and redefine the query text. The model will capture, analyze and classify these actions. All these information will be obtained through users' internet browser.

The author mainly refers to Fox's research results to describe the users' behavioral characteristics[14]. The difference lies in the fact that the features that the author studies are more concrete, Fox's being more general. Apart from that, as chart 1 shows, for more clarity, the author generalizes and classifies the users' searching behavioral features. It is classified as the following: clickthrough features, browsing features and query text features.

Table 1. User's Behavioral Features in Query

Clickthrough features	
Position	Position of the search result in the existing ranking
ClickFrequency	The frequency of the clicks in terms of the this search result (link)
ClickProbability	The expected click rate of this search result
ClickDeviation	The deviation from the expected click rate
IsNextClicked	Click the next link 1: Yes; 0: No.
IsPreviousClicked	Click the previous link 1:Yes; 0:No
IsClickAbove	Click the link above 1:Yes; 0:No
IsClickBelow	Click the link below 1:Yes; 0:No
Browsing features	
TimeOnPage	The dwelling time on page
CumulativeTimeOnPage	The aggregate time spent on all pages
TimeOnDomain	The aggregate time spent on all domains
TimeOnShortUrl	Time spent on filling the short url, without the parameters
IsFollowedLink	Whether it is followed by links 1:Yes; 0:No
IsExactUrlMatch	Use the apparent normalization 1:Yes; 0:No
IsRedirected	Whether the startup page is the same as the end page. 1:Yes; 0:No
IsPathFromSearch	Only click the links behind the searches 1:Yes; 0:No
ClicksFromSearch	Number of clicks before getting the target page
AverageDwellTime	The average dwelling time on each page in this search
DwellTimeDeviation	Deviation from the average dwelling time
CumulativeDeviation	The aggregate deviating time from the average dwelling time
DomainDeviation	Deviation between domain average dwelling time and the other average dwelling time
Query-text features	
TitleOverlap	The number of common words shared by the query text and web page title
SummaryOverlap	The number of common words shared by the query text and abstract
QueryURLOverlap	The number of common words shared by the query text and web address
QueryDomainOverlap	The number of common words shared by the query text and domain address
QueryLength	The length of the query in terms of words
QueryNextOverlap	Common words shared by this query and the previous one.

(1) Clickthrough features: Clicking is the a special interactive behavior between the users and search engines. The model includes all necessary features to describe users' clicking behaviors. For example, for a "query request---query result" information pair, this model not only describes the click frequency of a result(ClickFrequency), but also the question of whether there is click above or below(IsClickAbove，IsClickBelow).

(2) Browsing features: These features are for describing the interactive behavior after users click on the query lists and start to browse new pages. For example, this model will study the dwelling time on a certain page (TimeOnPage) and domain (TimeOnDomain), as well as the deviations from the expected dwelling time.

(3) Query-text features: The users decide on which search results to click with the information of search result names, url, and abstracts. Sometimes whether this file is authentic or original is not important. This model defines the relationship between query text and the abstract. These features include the The number of common words shared by the query text and web page title (TitleOverlap) and The number of common words shared by the query text and abstract (SummaryOverlap), etc.

5. User Interest Modeling based on Implicit Feedback

User interest modeling is one the fundamental issues in personalized search engine and information service studies. An implicit feedback based modeling is the one of the most frequently used methods, as well the one of the focal points in personalized search engine domain. Through automatically gathering and analyzing users' feedback information and modeling, an implicit feedback based search engine submit the matching search results to the users. At the same time, the system employs a certain algorithm to find out the query history related to the current query and automatically make the extended query.

5.1 Update User Interest Vector

Gerard Salton put forward several related feedback techniques in Smart system[18], combined with re-weighting the features and query extensions, and defined query correction method based on VSM (Vector Space Model) [19].

$$Q' = Q + \frac{1}{n_1}\alpha\sum_{i=1}^{n_1} R_i - \frac{1}{n_2}\beta\sum_{i=1}^{n_2} S_i \quad \text{(Formula 1)}$$

As you see from the above formula, Q is the eigenvector of the original query; R_i is the eigenvector of related articles; S_i is the eigenvector of the unrelated articles i. n_1 is the quantity of related articles and n_2 that of the unrelated articles. α and β are the contribution factors of related and unrelated articles towards query vectors. So Q' is the sum of feature vectors like original query, related and unrelated articles. Because the methods involve all of the feature items of the articles, they can not recognize which features are to user's taste, resulting in more noises and computational complexity. It was shown in an experiment that the sum weightings of top 30 feature items (sequenced in accordance with weighting) account for 83% of the total weightings of all feature items. The feature items after the 60^{th} weigh insignificantly towards the total vector, the sum of which account for about 4-7% of the total weightings[20]. So the author makes some meliorations to it and select the feature items whose weightings are relatively high ($30 \leqslant m \leqslant 60$), and revise the corresponding weighting of the feature items used for querying Q like what formula 2 indicates:

$$\mathbf{W_{qk} \leftarrow Wqk + \beta \cdot f(i) \cdot W_{ik}} \quad \text{(Formula 2)}$$

f (i) is the user's feedback value on article i, W_{ik} is the weighting value of article i's k-th feature item. W_{qk} is the weighting value of querying Q's k-th feature item. $\boldsymbol{\beta}$ is the learning factor.

At present most of the relevant feedback techniques require the users to participate in the sample training, pointing out which information is relevant and which is irrelevant, which brings inconveniences to the users. The author proposes a method to obtain the latent views of the users (i.e. implicit feedback) through observing user behaviors during the process of information interaction. The implicit feedback can be obtained through clickthrough features, browsing features and query-text features I mentioned above. A user's implicit feedback value on article i is the sum of these features, which can be formulated like below,

$$f(i) = \sum_{b \in B} c_b f_b(i) \quad \text{(Formula 3)}$$

$0 \leqslant f_b(i) \leqslant 1$ B is the set of all users' behavioral features. C_b is the weighting factors of the feedback behavior.

Normally the user's reading time is related to the length of the article. The longer the article, the more time user spends on reading. The author set the reading time feedback value $f_{rt}(i)$ as proportional to reading time and inversely proportional to article length. To reduce the noise disturbance, the author also set a minimum valid time t_0, only reading time that exceeds t_0 is the valid feedback.

The character string users search in the article is part of the user's interest, so we should extract features from them and use the following formula (2) to update Q.

5.2 Model Design of IFA System

Based on VSM model and user interest model, IFA (Implicit Feedback Agent) is a personal information retrieval system, which combines meta-search engine and Agent development. Through implicit feedback, Its aim is, with the use of the user interest model, to reflect the user's need to the maximum and thus promote the query accuracy. It is composed of five parts---user's Agent, meta search Agent, robot Agent, learning Agent and user interest model base(See in Fig.5).

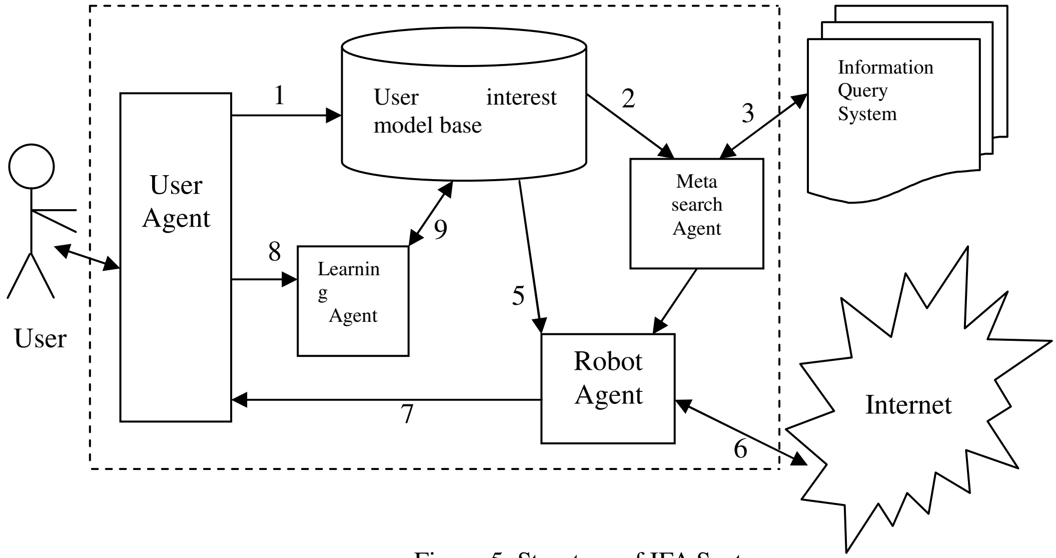


Figure 5. Structure of IFA System

The workflow of IFA:

(1) User initiates a user interest model q, and saves it to the user interest model base.

(2) Submit all feature items whose weight value is non-zero as the query keywords to meta-search Agent.

(3) Meta-search Agent sends query requests to diverse information query systems and retrieve URL lists that match the conditions.

(4) Gathering articles for every URL, preprocess them and obtain feature items, generating the eigenvector.

(5) Pattern matching and calculate the correlation between the current article and user interest model q.

(6) If the correlation is above the prescribed lowest $R_{min,}$ starting from the current URL, work the robot Agent to conduct a heuristic search, pattern match every article encountered in query.

(7) Submit the articles that fit the user interest model to the user.

(8) Observe the user's behavior, calculate user's relevant feedback value f (i) with formula (3).

(9) Use formula (2) to adjust user interest model q and switch to formula (3) to go on with the query until it meets the user's needs.

5.3 Structure and Adjustment of User Interest Model

The user interest model base is the core of IFA, deciding the meta-search Agent and robot Agent's working route and optimizing it through learning Agent. The user interest model base includes 1-N interest models, each one reflecting one type of user's interest, which can be expressed with the formula used to query q's vector as, $W_q = (W_{q1},W_{q2},...,W_{qk},...,W_{qn})$. W_{qk} is the weighting value of the k-th feature item Tk; n is the number of the feature items in the user interest model data. Through a Web browser imbedded on user's interface, IFA observes the behaviors in the relevant articles and sends the feedbacks to learning Agent which analyzes the implicit feedback behavior and adjust the user interest model q. Implicit feedback value f(i) can be obtained with formula (2). The weighting value W_{qk} of feature item T_k in the user interest model q can be modified with formula (1). To reduce the noise disturbance, the user interest model q should be normalized after each feedback; the feature item's weighting value which is lower than the threshold value should be set to zero.

5.4 Article Feature Acquisition and Pattern Matching

In VSM model, article i can be expressed in vector with the following formula:

$W_i = (W_{i1}, W_{i2}, \ldots, W_{ik}, \ldots, W_{in})$

$$w_{ik} = \frac{tf_{ik} \cdot \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{k=1}^{n} tf_{ik}^2 \cdot \log^2(\frac{N}{n_k} + 0.01)}}$$

(Formula 4)

In formula 4, tf_{ik} is the frequency of feature item T_k 's occurrence, n_k is the occurrence frequencies of feature item T_k in all the N queried articles. n_k and N are obtained with a dynamic statistical method during the user's query process.

Since the subjects that the information retrieval system processes are basically HTML with many markers, which have a high generality for the article content. So these markers can be used to improve the accuracy in the acquisition of the feature items. When obtaining the features, we may set the γ_{tag} weight factor to adjust the weight value of the feature item which appears in the HTML article marker domains.

Tag $\in$ {Clickthrough features ,Browsing features,Query-text features }

The correlation between article, i and user interest model q can be calculated with formula 5.

$$R(i) = \frac{W_i \cdot W_q}{|W_i| \cdot |W_q|} = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{qk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} w_{qk}^2}}$$

(Formula 5)

5.5 Heuristic Robot Based on Reinforcement Learning Algorithm

In pattern matching, if the correlation is lower than the minimum correlation R_{min}, then starting from the current URL, the heuristic robot, on the basis of reinforcement learning algorithm, will collect the information from the article in depth[21]. The web information space is composed of HTML articles and the hyperlinks in it. If we view HTML as node, the hyperlinks as directed edge, the web information space will become a directed graph. Starting from a certain article, we will definitely find the next most relevant article. However the common Robot 's roaming along the hyperlink on the web is largely blind, which lead to a waste of cyber resources and the system efficiency. As a result, to make the information retrieval more purposive, some improvements have to be made to the Robot. The author introduces reinforcement learning algorithm in Robot's path choosing, choosing the path that best reflects user's interest to retrieve from. Features are also obtained from every article on the retrieval path and match it with the user interest model. If the correlation is higher than R_{min}, then add it into the result list.

5.6 Application Prospect

The personalized artificial intelligence vertical search platform which has emerged recently stands for the trend of the internet development. A domestic corporation, Hangzhou Joinvc co. ltd. has already developed personalized artificial intelligence vertical search platform. This is a personalized, intelligent vertical searching request; by and large it has realized the fundamental functions of the next generation search engine. Its query behavior can automatically adjust the query results in accordance with user's search habits. The longer the user uses this search engine, the better it can meet the user's personalized needs. This "vertical search and calling" platform (www.aiyah.cn) is the world's first personalized, intelligent vertical search engine. With an online calling function, it can automatically pick out the most accurate, up-to-date, and genuine information in accordance with the users' query keywords, search habits, interests. As a search engine specially designed to serve the enterprises, it can also be customized according to the customers' needs.

Once the "Aiyah vertical search calling" is promoted successfully, this will be a new wealth growth point for the lackluster internet marketing companies. For the small and medium-sized enterprise owners, it will be a convenient, time saving, and highly efficient. information acquisition tool which helps the internet enterprises with better market orientation.

6. Conclusion

Aiming at improve the query quality of search engine, through gathering users' implicit feedbacks, the author sets up a user interest model and design a personalized information query system with the use of techniques like metasearch and Robot. This system is able to make individualized information queries, feed information back intelligently and learn automatically the users' interests, helping the users obtain their wanted information conveniently, rapidly and accurately.

References

[1] Craig Silverstein, Hannes Marais, Monika Henzinger, Michael Moricz. Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, Fall 1999, 33(1): 6-12.

[2] Yu Huijia, Liu Yiqun, Zhang Min, Ru Liyun, Ma Shaoping. Research in Search Engine User Behavior Based on Log Analysis. *Journal of Chinese Information Processing*, 2007, 21(1): 109-114.

[3] Sander-Beuermann W, Schomburg M. Internet Information Retrieval:the Further Development of Meta-Search Engine Technology. *Proceedings of the Internet Summit. Internet Society,* 1998.

[4] Thorsten Joachims. Optimizing Search Engines Using Clickthrough Data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data-mining*, 2002: 133-142.

[5] Sergey Brin, Lawrence Page. The Anatomy of a Large-scale Hypertextual Web Search Engine, Proceedings of the Seventh International World Wide Web Conference. *Computer Networks and ISDN Systems*, 1998, 30(1–7): 107-117.

[6] G. Salton, M. McGill. Introduction to modern information retrieval. *McGraw-Hill*, New York. 1983.

[7] Masahiro Morita, Yoichi Shinoda. Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994: 272- 281.

[8] J.Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J Riedl. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 1997, 40(3):77-87.

[9] Douglas W. Oard, Jinmook Kim. Implicit feedback for recommender systems. *Proceedings of AAAI Workship on Recommender Systems*, 1998: 81-83.

[10] Douglas W. Oard, Jinmook Kim. Modeling information content using observable behavior. *Proceedings of the 64th Annual Conference of the American Society for Information Science and Technology*, 2001: 481-488.

[11] Jeremy Goecks, Jude Shavlik. Learning users' interests by unobtrusively observing their normal behavior. *Proceedings of the 5th international conference on Intelligent user interfaces*, 1999: 129-132.

[12] M. Claypool, D. Brown, P. Lee and M. Waseda.Inferring User Interest. I*EEE Internet Computing*, 2001, 5(6): 32-39.

[13] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais and T. White. Evaluating implicit measures to improve the search experience. *ACM Transactions on Information Systems*, 2005, 23(2):147-168.

[14] Thorsten Joachims. Optimizing Search Engines Using Click through Data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data-mining*, 2002: 133 -142.

[15] Thorsten Joachims, L. Granka, B. Pang, H. Hembrooke, G. Gay. Accurately Interpreting Click through Data as Implicit Feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005: 154-161.

[16] Click Through Rate of Google Search Results. AOL-data.tgz,2006.

[17] Eugene Agichtein, Eric Brill, Susan Dumais , Robert Ragno. Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006: 3-10.

[18] Gerard Salton. The SMART retrieval system: experiments in automatic document processing. *Prentice-Hall*,

1971: 7-12.

[19] Gudivada, V.N., Raghavan, V.V., Grosky, W.I. Kasanagottu, R. Information Retrieval on the World Wide Web. *Internet Computing, IEEE*. 1997, 1(5): 58-68.

[20] Zou Tao, Wang Jicheng, Huang Yuan, Zhang Fuyan. Design and realization of an automatic classification system for the Chinese Documents. *Journal of Chinese Information Processing*, 1999(3): 27-33.

[21] Mitchell T. Machine Learning. *McGraw-Hill*, 1997.