MODELING BROWSING BEHAVIOR AND SAMPLING WEB EVOLUTION FEATURES THROUGH XML INSTANCES

Ioannis Anagnostopoulos[1], Christos-Nikolaos Anagnostopoulos[2] and Dimitrios D. Vergados[1]
[1]Department of Information and Communication Systems Engineering,University of the Aegean,
Karlovassi 83200, Samos – GREECE
[2]Department of Cultural Technology and Communication, University of the Aegean,
Mytiline 81100, Lesvos – GREECE

Keywords: Web search, web evolution, user modeling, XML services.

Abstract: Nowadays, many web services have the essential role of accessing and processing the disseminated information on the web, while in parallel, distribute and exchange metadata among them in order to deliver relevant information to the end user. However, their competence is hindered, due to the vast amount of information added, the poor organization, as well as the lack of effective crawling techniques that fail to follow the exponential growth of the web. In this paper we propose an algorithm, which uses five third party web search services and it is capable of self-adapting over the incessant changes that occur on the indexed web. The algorithm works in conjunction with a user browsing behavior model that monitors and records the users' interactions with the third-party services, using XML search session instances. From the assessment made, it was concluded that the proposed algorithm not only adapts to the users' web search profiles but also adapts to the evolutionary nature of the web.

1 INTRODUCTION

Due to the extremely rapid growth of the web, search engines cannot index all the new pages at the same time or with the same priority. Besides, the search engines update their catalogues with different algorithms, having as a result different response time in updating their directories (Oyama et al, 2004), (Pokorny, 2004). In this paper we propose a meta-search algorithm capable of self-adapting over the continuous changes that occur on the web, providing in parallel personalized information in respect to the diversity of the users' information needs. We prove in our approach, that when users' preferences are used jointly with dynamic survey mechanisms that adapts to web evolution events and changes, a more efficient search is provided. Transparency is achieved for both personalization and web evolution adaptation mechanisms, requiring virtually none effort from the user's part. We also deal with merging query results from five different search engines (meta-results) that use different ranking algorithms. For this problem several algorithms were proposed (Henzinger, 2001) and (Losee and Church 2004). The five web search services that we use in our work are publicly available and heavily-visited web search engines. However, we intentionally do not cite these web search services, since in our objectives was not their evaluation at the moment. Such kind of assessment, which requires repeatedly evaluations before its results are disseminated publicly is left for future work, where more web search services are about to be included in our research. Thus, in this work we will name these services as SE1, SE2, SE3, SE4 and SE5.

2 MODELING WEB SEARCH BEHAVIOR - RELATED WORK

There are several papers and approaches in the literature in respect to modeling user web search behavior as well as creating personalization patterns. A common approach is to monitor the users' browsing behavior and develop their profiles from data derived either from past actions (Discovering User Interests from Web Browsing Behavior: An Application to Internet News Services) or from momentary-changed actions (Adaptive Web Search Based on User Profile Constructed without Any Effort from Users). According to (Personalization

Anagnostopoulos I., Anagnostopoulos C. and D. Vergados D. (2007).
MODELING BROWSING BEHAVIOR AND SAMPLING WEB EVOLUTION FEATURES THROUGH XML INSTANCES.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 67-74
DOI: 10.5220/0001279500670074

of Web search) there are six approaches to web search personalization, namely relevance feedback and query modification, personalization by content analysis, recommender systems, personalization by link analysis, social search engines, and finally mobile and context-aware searching. Finally, there are some works that tackle the same problem characterizing and modeling the dynamic nature of internet itself as a medium that carries the disseminated information (Anagnostopoulos and Stavropoulos, 2006). Moreover, in (Predictive Modeling of First-Click Behavior in Web-Search) the authors predict user surfing patterns improving in this way web search.

2.1 Our Approach

In this paper the proposed personalization algorithm is a client-side agent that provides meta-results based on the users' web search interactions using five different web search services. The term meta-result defines a re-ranked result, which was acquired from one or more third party search services and it is presented to the user without labeling its source(s). In our approach, we assume that past search behavior is an indicator of the user's future behavior, as a basis for user modeling. The construction of the personalized preferences is performed in a totally transparent way, without interferences in the users' browsing behavior, while the merged meta-results are presented without labeling their source, ensuring in this way that the user is completely unbiased to his preferences. The only feedback that the user receives is a text paragraph regarding the URL as most of web search engines do. The personalized preferences are recorded client-side and they are updated continuously according to the meta-results acquired by the user, the time spent for their exploration as well as the search depth in terms of hyperlinks. Thus, the user's profile is also adjusted to any possible changes in respect to his information needs.

The data stored for the personalized algorithm, define some search preference features regarding the information explored by the user through the visited meta-results (query, involved search service, ranking position, timestamp, link depth). Every time the user browses a URL from the merged meta-results, these features are kept into an XML file and update in parallel the weights that define the user's confidence (or priority) in respect to each of the employed web search services. In the timestamp field the time where the user spends in order to explore the specific result is recorded. An instance of the XML files that stores the above information is depicted in Figure 1. Summarizing, the data regarding the personalization, are implicitly gathered through the user's interaction with the system, so the user is not biased nor encumbered with submitting information to the meta-search engine. Personalized data are stored client-side on the user's machine, providing privacy and security.

As far as the adaptation is concerned, the algorithm is dynamically adjusted in order to reflect the user's current interests and preferences, while the profile is updated on-line (during the search on the returned meta-results) so it instantaneously adapt to the user's behavior during his search. Using such adaptation mechanisms the proposed meta-search engine process and re-ranks third party search results.

In order to personalize the similarity of a meta-result in respect to the user's preferences in our approach, we use two probabilistic functions. These functions assign a probability value according to the time where the user spends for information exploration as well as according to the depth of the investigated link (web page). As depth we define the number of hyperlinks used from the initiation of the search where the starting point is the meta-results, until the URL reached by the user.

We consider time as an important factor in personalization since the more time the user spends exploring a specific result, the more this result is possible to be relevant and vice versa. The period of time consumed during a search session was modeled according to a standard lognormal distribution since this distribution fits with the results made in respect to web search behavior (Mondosoft Development Team - White paper, 2004).

The methodology and solutions presented in (Mondosoft Development Team - White paper, 2004), were based on an extensive set of real-world data gathered from 400 widely varying web sites that use a hosted-based search solution. According to the above study it was concluded that web users want to obtain search results as fast as they can and with the minimum possible effort. The typical behavior of a web user is similar to the aforementioned pattern as stated in (Mondosoft Development Team - White paper, 2004). Consequently, based on this pattern we once more modeled the web user search depth link behavior according to a lognormal survival function.

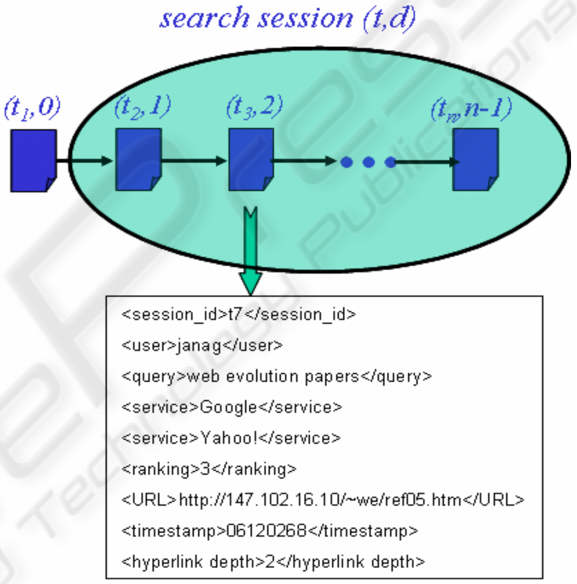
As far as the time spent in information exploration by the user is concerned, we assumed that if the investigation of a proposed result (or a result that derived from further link search) during a search session is reaching the threshold of five minutes, then this result is highly possible to be relevant. This threshold was derived from the survey described in (Mondosoft Development Team - White paper, 2004), where the authors experimentally concluded that on average, users made 2.4 searches per session. In addition, the average search session duration is 1 minute and 50 seconds. Based on these metrics, we assume that beyond this level, the user is rather confused on whether this result gratifies his information needs. According to (Mondosoft Development Team - White paper, 2004) web visitors use search services and portals that are oriented towards their information needs, to speed up the exploration of the results. Moreover, it was concluded that when a user uses the correct search service (or performs a local search in a relevant portal), his search duration is nearly two minutes at average (Mondosoft Development Team - White paper, 2004). This threshold is also used to avoid misjudges in the scoring of the result, due to idle activity periods in the user's work.

2.2 Scoring Web Browsing Behavior

As search session, we define the time the user needs to accomplish his search from a starting point (which in our case is a meta-result). A search session includes a set of links that contain a chain of web pages (or documents) only in case that these were accessed within a short time interval (five minutes) or a web page was attained trough a link that belong to the session (even after the threshold of five minutes). Based on the above, the personalization score is provided according to Equation 1.

$$Pers\_weight(t,d) = \frac{max[P_t^{c_1} \cdot P_d^{c_2}]}{\sum_{d=1}^{l} \int_{t=0}^{T} (P_t^{c_1} \cdot P_d^{c_2})dt} \quad (1)$$

where $P_t(x) = \frac{2e^{-((ln x)^2/2\pi^2)}}{\sqrt{2\pi x}}$ stands for the standard lognormal distribution, while $P_d(x) = 1 - \Phi(2ln x)$ is the lognormal survival function, where Φ is the cumulative distribution function of the normal distribution. Parameters c_1, c_2 are used for fine-tuning purposes and their values are positive. This allows to weight the personalized score Pers_weight (t,d), according to the time t needed to explore the results as well as to the depth d of the investigated web sources. Fine-tuning of c_1 and c_2 is quite important and gives the opportunity for further adjustments over individual user behavior. However, in this paper we consider that both behavior metrics contribute equally to the personalized score and thus the fine-tuning parameters were set equal to one. Figure 1 presents the stored information in the XML file regarding the time spent and the hyperlink depth of the web resource of a specific session (t7) for the user "janag", who submitted the query "web evolution papers" and started his search from the provided result (http://147.102.16.10/~we/ref05.htm), which was ranked in the third position of the provided third-party results by Google and Yahoo!



```
<session_id>t7</session_id>
<user>janag</user>
<query>web evolution papers</query>
<service>Google</service>
<service>Yahoo!</service>
<ranking>3</ranking>
<URL>http://147.102.16.10/~we/ref05.htm</URL>
<timestamp>06120268</timestamp>
<hyperlink depth>2</hyperlink depth>
```

Figure 1: XML web search session instance.

3 WEB EVOLUTION MONITORING MECHANISM

This section describes the basic concepts and the main features of the proposed web evolution adaptation mechanism. This mechanism is put into the context of capture-recapture experiments used in wildlife biological studies. In such experiments animals are captured, marked and finally released on several trapping occasions. If a marked animal is captured on a subsequent trapping occasion, it is said to be recaptured. Based on the number of marked animals that are recaptured, one can estimate the total population size using statistical models and their estimators.

There any many capture-recapture sampling protocols in the literature. The sampling scheme chosen for capturing, marking and recapturing the meta-results is the robust design (Schwarz and Stobo, 1997), which extends the Jolly-Seber Model (Jolly, 1965). This model was chosen among other capture-recapture models since in wild-life experiments it is applied to open populations, in which there is possibly death, birth, immigration, and permanent emigration. In our approach, death corresponds to a meta-result that is no longer exists (dead links, errors 404), birth match to a new meta-result (new or updated information), while incidents of immigration and/or emigration correspond to active but temporary unavailable meta-results due to (errors of type 50* such as web server internal errors, bad gateway, service/host unavailable, etc). This incident indicate that web is a very volatile universe, where its information units (web pages, documents) do not live forever, or they are moved to another location (servers), or are renamed, while in parallel new information is added. The basic steps of the web evolution adaptation mechanism are summarized in the following sub-section.

3.1 Description

In order to adopt the real-life experiments to our algorithm, we considered the following assumptions described in (Eguchi, 2000). Note that in our paradigm, the third-party results are considered as wild animals, while the population under study consists of the collected third-party results.

1. Each result, which is present in the population (either marked or unmarked) during the time of the i^{th} sample (i = 1, 2,…, k) has the same probability p_i of being captured.
2. Every marked result present in the population immediately after the i^{th} sample has the same probability of survival φ_i until the $(i+1)^{th}$ sampling time (i = 1, 2, …, k-1).
3. Marks are not lost and not ignored.
4. All samples are instantaneous and each release is made immediately after the sample.
5. Immigration is permanent and cannot be separated from birth and death measurements.

Let us now introduce the parameters and the metrics involved in our case. M_i is the number of marked results in the population at the time where the i^{th} sample is collected (i = 1, 2, …, k; M_1 = 0), where k is the number of primary sampling periods, N_i is the total number of results in the population at the time where the i^{th} sample is collected and B_i stands for the total number of new activated results entering the population between the i^{th} and $(i+1)^{th}$ samples and still remain in the population at the time $(i+1)^{th}$ sample is collected (i = 1, 2, …, k-1). In addition, φ_i defines the survival probability for all results between the i^{th} and $(i+1)^{th}$ samples, while p_i corresponds to the capture probability for all results during the i^{th} sample. Moreover, m_i and u_i correspond to the number of the marked and unmarked results captured in the i^{th} sample, while their sum defines the total number of results captured in this sample (n_i). Finally, R_i is the number of the n_i that were released after the i^{th} sample, r_i is the number of the R_i results released at the i^{th} sample that are captured again and z_i is the number of results captured before the i^{th} sample, not captured at the i^{th} sample and captured again later.

Taking the above parameters into consideration, the population size in sample i is given according to Equation 2, where m_i is the number of marked results in the i^{th} sample and n_i is the number of total results captured in the i^{th} sample (i=2,3, …, k-1).

$$\hat{N}_i = \frac{n_i \hat{M}_i}{m_i} \qquad (2)$$

Thus, the survival rate estimator is obtained by first considering the number of marked results in the population, immediately after i^{th} sample was collected and it is defined by Equation 3, where M_i– m_i is the marked results not captured in the i^{th} sample, whereas R_i is the number of results captured, marked in the i^{th} sample and then released. Thus, an intuitive survival rate estimator is given by Equation 4, where i=1,…, k-2.

$$w_i = M_i - m_i + R_i \qquad (3)$$

$$\hat{\phi}_i = \frac{\hat{M}_{i+1}}{\hat{M}_i - m_i + R_i} \qquad (4)$$

An estimator of the birth between the i^{th} and the $(i+1)^{th}$ samples is provided from Equation 5, where i=2,…, k–2, highlighting the difference between the estimated population size at the $(i+1)^{th}$ sample and the expected survived results from the i^{th} to $(i+1)^{th}$, which is actually defined by $\phi_i \cdot (N_i - n_i + R_i)$.

$$\hat{B}_i = \hat{N}_{i+1} - \hat{\phi}_i(\hat{N}_i - n_i + R_i) \qquad (5)$$

Furthermore, Equation 6 gives the capture probability p_i, which can be estimated as the proportion of marked or total marked and unmarked active results that are captured in the i^{th} sample, where i=2, …, k-1.

$$\hat{p}_i = \frac{m_i}{\hat{M}_i} = \frac{n_i}{\hat{N}_i} \qquad (6)$$

Finally, in order to estimate M_i we use Equation 7, which provides the future recovery rates of the two distinct groups of marked results in the population at sampling period i. Moreover, $(M_i - m_i)$ defines the marked results not captured during the i^{th} sampling period, while R_i are the results captured at the i^{th} period, marked, and then released for possible recapture in future samplings. As a result, the estimator of M_i is defined from Equation 8, where i=2, …, k-1.

$$\frac{z_i}{M_i - m_i} \approx \frac{r_i}{R_i} \tag{7}$$

$$\hat{M}_i = m_i + \frac{R_i z_i}{r_i} \tag{8}$$

However, although these estimators are intuitively reasonable, yet they are biased. Thus, Seber and Jolly based on the above assumptions suggested in (Jolly, 1982) and (Seber, 1982) the unbiased estimators, which are defined in Equations 9 up to 13. Especially, Seber and Jolly recommended that m_i and r_i should be greater than 10 for satisfactory performance of these bias-adjusted estimators as stated in (Eguchi, 2000). Hence, this composes the sixth assumption in respect to the five stated in the beginning of this section, which does not affect the adaptation of the real-life experiments to the web (m_i and r_i are easily satisfy this condition). Equation 14 defines the birth rate between i^{th} and $(i+1)^{th}$ primary sampling periods.

$$\tilde{M}_i = m_i + \frac{(R_i + 1)z_i}{r_i + 1} \tag{9}$$

$$\tilde{N}_i = \frac{(n_i + 1)\tilde{M}_i}{m_i + 1} \tag{10}$$

$$\tilde{\phi}_i = \frac{\tilde{M}_{i+1}}{\tilde{M}_i - m_i + R_i} \tag{11}$$

$$\tilde{B}_i = \tilde{N}_{i+1} - \tilde{\phi}_i(\tilde{N}_i - n_i + R_i) \tag{12}$$

$$\tilde{p}_i = \frac{m_i}{\tilde{M}_i} \tag{13}$$

$$\tilde{b}_i = \frac{\tilde{B}i}{\tilde{N}_i} \tag{14}$$

Finally, Equation 15 highlights the ability of each tested web service to adapt in the incessant evolution of the web, which requires high freshness rates, exclusion of dead links and continuous control over the validity of the provided results. Moreover, Equation 15 assigns to each involved third party web service a weight value (*Ad_weight*), calculated at the end of the capture-recapture experiments, by the product of the average values of the birth and survival rates between the following primary sampling periods, as these are calculated by Equations 11 and 14 respectively. In Equation 15 *s* stands for a scaling parameter, which in our case was arbitrary set equal to 100.

$$Ad\_weight = s \cdot avg(b) \cdot avg(\phi) \tag{15}$$

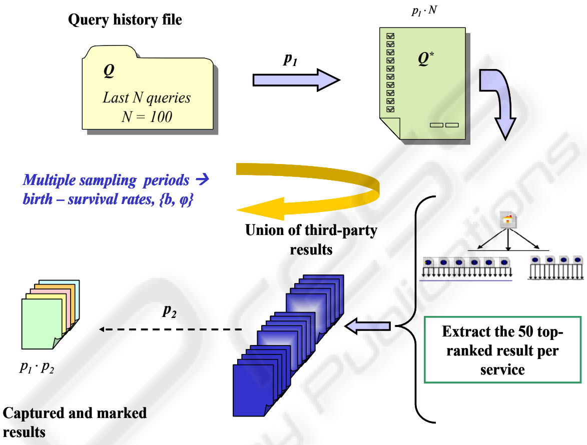

Query history file
Q
Last N queries
N = 100
p_1
$p_1 \cdot N$
Q^*
Multiple sampling periods →
birth – survival rates, {b, φ}
Union of third-party results
p_2
$p_1 \cdot p_2$
Extract the 50 top-ranked result per service
Captured and marked results

Figure 2: The sampling procedure.

3.2 Sampling Third Party Services

In this section we describe the proposed web capture-recapture sampling method, which is used in our mechanism. As mentioned above, this experiment aims to simulate the characteristics of animal capture in wildlife. In this simulation, different animal species under study that live in an area correspond to different meta-results derived from different queries, while the traps correspond to our sampling method. In the real life experiment the traps are set up for a specified amount of time and the theory used assumes that all animals have the same probability of being caught in a trap. Thus, in our algorithm we had to ensure that results of different queries have the same probability of being captured during the sampling procedure.

Taking into account these considerations the sampling method is as follows (Figure 2). Working in the background the algorithm keeps a file, which records of all the submitted queries of the user. This file is continuously updated, while its buffer size is *N*, working according to the leaky bucket model in the steady state mode recording the last N queries submitted by the user. This file, (called *Q* from this point forward) is traversed sequentially and each record is included to a poll Q^* under a probability value p_1. Then, these queries are simultaneous submitted to the selected web search services. This means that at each sampling period a sample that consists of about $p_1 \cdot N$ is drawn. In the sequel, the transmitter needs to understand and query each web search service in its own interface, integrating varied search interfaces. Furthermore, before the query submission, the transmitter module performs a check of resource accessibility such as network bandwidth and connection availability, before making any use of network resources. The first twenty results for each of the five web search services are collected, the duplicate fields are removed (keeping in parallel the search services that provided the result) and finally the merged results are stored (randomly ranked) in a local file. Then a portion of the merged-results is selected under a second probability value p_2. The probability values p_1 and p_2 are fine-tuned according to the amount of the previously submitted queries N in the buffer Q. However, with this process we allow each sampling instance to have equal probability values of being included in each sample, independently of the instances that have already been sampled. Thus, the probability of selection of an individual result derived for each of the five web search services used is defined by the product $p_1 \cdot p_2$, while the interval between primary are secondary sampling periods were fine-tuned during pilot executions. Furthermore, each sampled meta-result is labeled with four attributes. These are the primary and the secondary sampling period, the URL-URI (identifier) of the sampled result, and finally the identifier of the web search services that provide the respective meta-result.

Consequently, after multiple capture-recapture and taking into account the total amount of captured instances (result) in a current sample, the previously marked as well as the marked sampling instances in subsequent sampling periods, the proposed web evolution adaptation mechanism estimates the birth rate and the survival rate for each one of the five web search services according to Equations 11 and 14.

4 HANDLING THE THIRD-PARTY WEB SERVICES

As it is mentioned in section II the user behavior adaptation mechanism is responsible for re-ranking the third party search results according to the user activity in respect to the time spent for information exploration as well as according to the depth of the investigated web page or document. On the other hand, in section III we introduced an adaptation mechanism, which is capable of adjusting a scoring value that reflects to the ability of each search service to follow web evolution events (up-to-date results, validity verification, dead-links exclusion, etc).

Both mechanisms implicitly adapt over the user's preferences and search profile as well as to the dynamic nature of the web. Thus, we propose an isolated scoring mechanism that combines user personalization along with the search service's ability to keep in touch with the incessantly changes on the web. This scoring formula jointly employs both of the proposed weighting mechanisms (Equations 1 and 15) and the assigned weight is given by the product $\alpha \cdot Per\_weight(t,d) \cdot \beta \cdot Ad\_weight(b,\phi)$, where α, β are fine-tuning parameters. Further details can be found in (Anagnostopoulos and Stavropoulos, 2006).

5 PERFORMANCE ASSESSMENT - CONCLUSION

This section describes some experimental evaluations made in order to assess the influence of our proposed adaptive weighting schemes. Two kinds of assessments were made. During the first, we evaluated the impact of the proposed personalization algorithm*,* while in the mean time we run the capture-recapture experiments so as to further appraise the effect of the web evolution adaptation mechanism in the re-ranking procedure of the third-party acquired results.

In both cases, we quantified the average precision for the top-50 merged meta-results, over different recall levels using 50 different queries, which were submitted every month from January to Aug of 2006.

Figure 3 depicts eight Precision-Recall (PR) diagrams for each month respectively (PR(Jan), PR(Feb), PR(Mar), PR(Apr), PR(May), PR(Jun), PR(Jul) and PR(Aug)).

In Figure 4 a precision-recall curve is marked with an asterisk, when both proposed adaptation mechanisms are jointly employed (PR(May*), PR(Jun*), PR(Jul*) PR(Aug*)).

Table 1: PR diagram.
[*Pers_weight(t,d)*]

Recall	Precision							
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
0	0.82	0.84	0.87	0.88	0.89	0.86	0.90	0.88
0.1	0.75	0.78	0.77	0.79	0.84	0.83	0.86	0.87
0.2	0.74	0.76	0.75	0.77	0.80	0.82	0.83	0.85
0.3	0.64	0.68	0.73	0.75	0.77	0.76	0.79	0.78
0.4	0.60	0.59	0.61	0.63	0.66	0.72	0.72	0.77
0.5	0.46	0.56	0.58	0.60	0.65	0.65	0.65	0.69
0.6	0.38	0.43	0.45	0.49	0.51	0.59	0.58	0.61
0.7	0.32	0.33	0.33	0.34	0.36	0.43	0.49	0.53
0.8	0.14	0.16	0.18	0.17	0.22	0.27	0.31	0.36
0.9	0.12	0.08	0.10	0.09	0.17	0.18	0.20	0.20
1	0.03	0.04	0.05	0.06	0.08	0.05	0.04	0.06

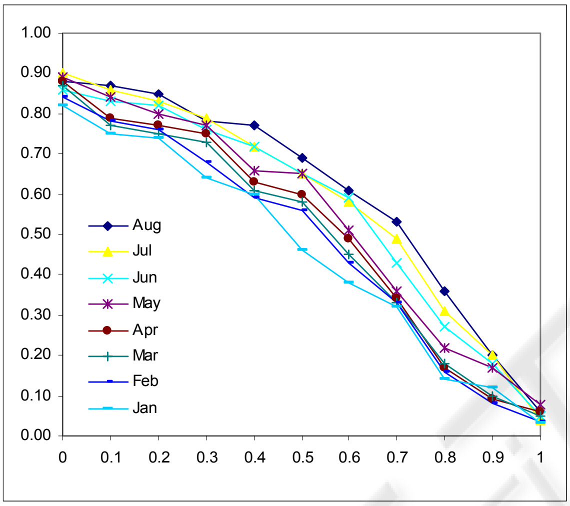


Figure 3: Results – Precision / Recall diagrams [*Pers_weight(t,d)*].

The lower average precision values correspond to PR(Jan) and this was in a way expected since there was no considerable feedback provided to the meta-results ranking mechanism, due to the fact that, at that time period the personalization algorithm had just started to adapt to the user's browsing behavior and preferences.

However, a significant improvement was measured after one month, since we measured nearly 2.2% increase (in average) for eleven recall levels from zero up to one (0, 0.1, …, 1.0), as depicted in Table 1. In the sequel, the average precision values were further increased between subsequent months ([Feb-Mar]:1.6%, [Mar-Apr]:1.4%, [Apr-May]:3.4%, [May-Jun]:1,9%, [Jun-Jul]:3.8%, and [Jul-Aug]:2.1%). A significant improvement in the relevancy of the returned merged meta-results was also noticed, when the web evolution adaptation mechanism contributed to the ranking of the third party results for both evaluation measurement through Equation 15 ([May*]:2.5%, [Jun*]:2.1%, [Jul*]:2.5%, [Aug*]:1.9%) as depicted in Table 2.

Table 2: PR diagram.
[*Pers_weight(t,d)*Ad_weight(b,φ)*]

Recall	Precision							
	May	May*	Jun	Jun*	Jul	Jul*	Aug	Aug*
0	0.89	0.89	0.86	0.9	0.9	0.91	0.88	0.9
0.1	0.84	0.86	0.83	0.87	0.86	0.88	0.87	0.88
0.2	0.8	0.8	0.82	0.84	0.83	0.83	0.85	0.85
0.3	0.77	0.76	0.76	0.77	0.79	0.81	0.78	0.8
0.4	0.66	0.69	0.72	0.74	0.72	0.75	0.77	0.81
0.5	0.65	0.68	0.65	0.68	0.65	0.69	0.69	0.73
0.6	0.51	0.56	0.59	0.6	0.58	0.6	0.61	0.63
0.7	0.36	0.41	0.43	0.44	0.49	0.53	0.53	0.57
0.8	0.22	0.26	0.27	0.29	0.31	0.32	0.36	0.33
0.9	0.17	0.2	0.18	0.17	0.2	0.24	0.2	0.21
1	0.08	0.11	0.05	0.09	0.04	0.08	0.06	0.1

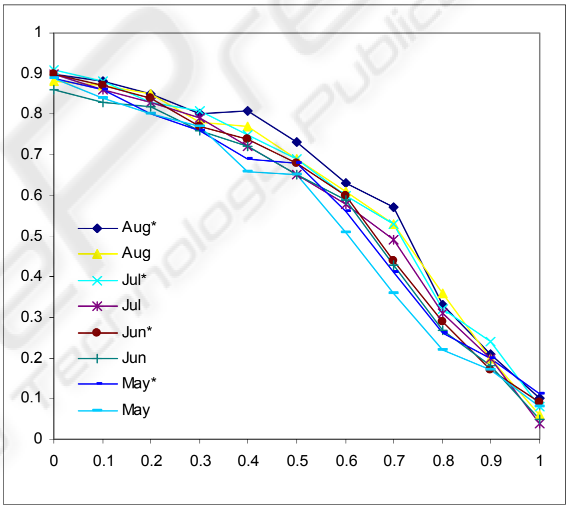


Figure 4: Results – Precision / Recall diagrams [*Pers_weight(t,d)*Ad_weight(b,φ)*].

From the assessment results it was concluded that the personalization algorithm adapts to the user's web search profile quite fast. As illustrated in Figures 3 and 4, this is mainly noticeable for the lower recall levels where the top re-ranked results reside. Similarly, the web evolution adaptation mechanism seems to have a positive impact over the delivery of relevant results to higher ranking positions. The results clearly shown that if we trust search services, which are able to adapt to the continuous changes that occur on the web we will earn an important amount of information in the higher ranking positions. In addition, we will dispose of information that either is not up-to-date or is not suitably disseminated through the web (dead links, old records, validity errors etc).

As far as future considerations are concerned, studies over users with different information needs (e.g. researchers, students, doctors, etc) need to be made in order to fully evaluate the performance of the algorithm. This will allow us to proceed to a thorough investigation over the influence of the fine-tuning parameters c_1, c_2, α and β.

It is also possible to discover relations between these fine-tuning parameters and different behavior browsing patterns. Finally, in future we consider a server-side implementation, in order to be easier to perform multiple capture-recapture experiments on the web. This will be also helpful for studies over personalization and browsing behavior, sharing data among different topologies of web servers and services.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Photis Stavropoulos and Dr. Georgios Kouzas for their valuable help and comments, in related research done during last years.

REFERENCES

Anagnostopoulos I, Psoroulas I, Loumos V, and Kayafas E, 2002. Implementing a customised meta-search interface for user query personalisation, 24th International Conference on Information Technology Interfaces, ITI 2002, pp. 79-84, June 24-27, Cavtat/Dubrovnik, Croatia.

Anagnostopoulos I., Stavropoulos P, 2006. Adopting Wildlife Experiments for Web Evolution Estimations: The Role of an AI Web Page Classifier, IEEE/WIC/ACM International Conference on Web Intelligence (WI 06) In: Main Conference Proceedings pp. 897-901, 18-22 December 2006, Hong Kong, China.

Eguchi T, 2000. Statistic class notes (Topics in Ecological Statistics), Montana State University, based on book by Seber (1982; The estimation of animal abundance and related parameters. Second edition, Macmillan Publishing Co., New York, NY), downloaded from http://www.esg.montana.edu/eguchi/pdfFiles/markRecapSummary.pdf

Henzinger MR, 2001. Hyperlink analysis for the Web. IEEE Internet Computing. 5(1):45-50

Jolly G, 1965. Explicit estimates from capture-recapture data with both death and immigration stochastic model, Biometrika vol. 52, pp. 225-247.

Jolly GM, 1982. Mark-recapture models with parameters constant in time, Biometrics vol. 38, pp. 301-321.

Keenoy K, Levene M, 2005. Personalization of Web Search, B. Mobasher and S.S. Anand (Eds.): ITWP 2003, LNAI 3169, pp. 201–228, Springer-Verlag Berlin Heidelberg 2005

Liang TP, Lai HJ, 2002. Discovering User Interests from Web Browsing Behavior: An Application to Internet News Services, Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS-35/02), 0-7695-1435-9/02, 2002.

Losee RM, Church LJr, 2005. Information retrieval with distributed databases: analytic models of performance. IEEE Transactions on Parallel and Distributed Systems. 15(1):18-27

Mondosoft Development Team – White paper, 2004. Web Site Usability Metrics - Search Behavior and Search Trends, last updated May 2004, downloaded from http//www.mondosoft.com/SearchBehaviorWP.pdf.

O'Brien M, Keane MT, Smyth B, 2006. Predictive Modeling of First-Click Behavior in Web-Search, WWW 2006, ACM 1-59593-323-9/06/0005, May 23–26, 2006, Edinburgh, Scotland.

Oyama S, Kokubo T, Ishida T 2004, Domain-specific Web search with keyword spices. IEEE Transactions on Knowledge and Data Engineering. 16(1):17–27

Pokorny J, 2004. Web searching and information retrieval. Computing in Science & Engineering. 6(4):43-48

Schwarz C. and Stobo W, 1997. Estimating temporary migration using the robust design, Biometrics vol.53, pp. 178–194.

Seber GA, 1982. The estimation of animal abundance and related parameters, 2nd edition, Macmillan Publishing Co., Inc. New York.

Sugiyama K, Hatano K, Yoshikawa M, 2004. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users, WWW2004, May 17–22, 2004, New York, New York, USA, ACM 1-58113-844-X/04/0005.