DETERMINING USER INFORMATION GOALS IN MEDIA-RICH WEB SITES BY RETHINKING INFORMATION SCENT THEORY

Rahul Singh and Bibek D. Bhattarai
Department of Computer Science, San Francisco State University, San Francisco, CA 94132
rsingh@cs.sfsu.edu[†], bdb@sfsu.edu

ABSTRACT

The "information goal" underlying user activity on the web is an important latent parameter. Its determination can help in a wide variety of tasks ranging from improving the quality of a search engine's results to design and evaluation of web sites. The problem of determining user information goals is exacerbated in media-rich websites. The complexity lies in the fact that the semantics as well as the user-content interactions associated with media-rich information are context-based, emergent, and hard to capture. Furthermore, simple summary data such as traversal history or links clicked are, by themselves, insufficient to obtain insights about the information needs of a user. Doing so requires an explanatory model of user action. In this paper we propose techniques for automatically estimating user information goals from usage patterns in media-rich sites. Our research builds on ideas from information foraging theory; specifically, we propose methods to capture the influence of media-rich information and its structure on information goals and re-think the classical notion of information scent to capture the subjective perception of the value/cost/access-path of media-based informational entities. Putative information goals can subsequently be estimated from browsing patterns. Experimental investigations study the efficacy of the proposed techniques for estimating user goals and user-content interaction patterns in complex media rich sites.

1. INTRODUCTION

The Web (WWW) represents the largest and arguably the most complex repository of content at the current state of technological development. Information on the web is represented using a variety of media, with a (current) predominance of text- and images-based data and increasing presence of other modalities such as video and audio. Being such a large and important repository of information, the development of theories and methods to understand human-content interactions on the media-rich web constitutes an important topic of research. The typical user behavior in a web site is not random; rather, it is driven by the user information goal. That is, the user makes traversal decisions looking for information that would satisfy his or her need. Given this fact, one of the key challenges is to understand the relationship between content/content-organization and user information need on one hand with the user behavior on the other. Specifically, this relationship can be explored in light of the following two fundamental and practically relevant questions:

- Given the user behavior, how can we deduce the underlying (unknown) user information need/goal?
- Given the user information goals what could be the (expected) user behavior patterns?

Current studies identify two predominant modes through which a user can access information on the web to satisfy their information needs [10]: by searching, also called "search by-query" or by browsing, called "search by-navigation". Addressing the above questions is crucial to improve the efficacy of techniques and information organization in both these settings. For instance, deducing the information goal can be used for improving page ranking [3,5] and result presentation. It can also be used in prompting users during site navigation so as to lead them to their information goal [2,6]. Finally, estimating the browsing pattern from the information goal can help web site design by providing the designers an understanding of the putative traffic patterns of the site and in identifying parts of the site where the design can be improved.

1.1. Problem Formulation

The research in this paper addresses the first of the aforementioned questions in a "search by-navigation" setting. That is, given a sequence of pages visited by a user, the problem is to identify the content which putatively satisfies the information goal(s) of the user. Formally, this can be stated as follows: consider a web site $W = (P, L)$, consisting of P pages which are connected through the set L of hyperlinks. Let the content of W be described through textual terms $T=\{t_1, t_2,...t_n\}$ and media elements $M=\{m_1, m_2,\ldots m_n\}$, which can be images, video, or audio. For the purposes of this paper, we shall assume the media to be image-based since this is the most commonly encountered case. Other forms of media can easily be incorporated in our approach, as needed, by using appropriate descriptors. Let now the sequence of pages in W visited by a user during a session be denoted by $\Pi = \{P_1, P_2, ...P_k\}$. Our problem then is to identify the respective elements of T and M that constitute

[†] Corresponding author

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

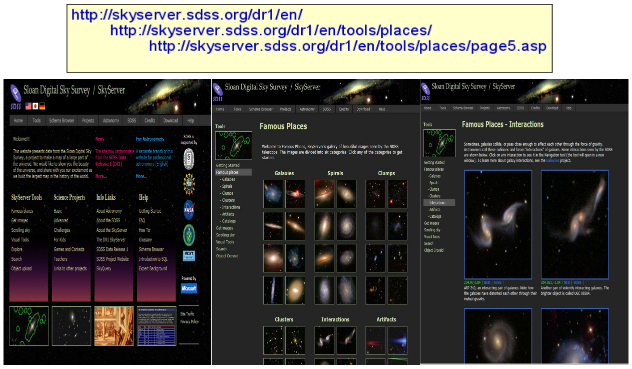
the putative information goals of the user. *It may be noted that in this formulation, the information goals are implicitly user-specific in that they are defined for each browsing patterns.* Consequently, the formulation is sensitive to both user context and data context. The importance of user specificity is crucial for media-rich site content since the semantics associated with media-based information is emergent, i.e. media is endowed with meaning by placing it in context of other similar media and through user interactions [11].

1.2. Overview of the Solution Framework

Our solution involves three key ideas. First, we note that in normal usage, user actions at a site are not random. Rather, they are purposive and a consequence of the user information need and the site content and structure. Unlike the user information need, which is unknown and needs to be determined, the site content can be captured using appropriate text/media descriptors. Furthermore, the user actions at a site, e.g. traversal history or links clicked, can be obtained from the web-log and provide cues to the underlying information needs. To motivate the second point, we note that summary data regarding user actions at a site is by itself insufficient to obtain insights about the information needs of a user. Doing so requires an additional component, namely, an explanatory model of user action. The Information Foraging Theory [8] is one such model that has received significant attention. This theory considers information seeking behavior to be adaptive within the constraints of the human-information environment in which the user interacts. An important component of this model is the notion of information scent [9], which is a psychological model of how cues, such as links in a web page, are used by users to make information seeking decisions. Our third and final idea involves rethinking the notion of information scent to account for media-rich web pages and user-behavior therein.

2. PRIOR RESEARCH AND CHALLENGES

In the search-by-query scenario, cumulative information related to click-behavior and anchor-link distribution available in search engine logs can be used to discern the correlation between user information goals and the query terms used by them [6]. However, in the search by-navigation scenario being considered by us, precise cues such as query-terms are unavailable. For this problem, in [8], a probabilistic model of information scent was developed using a network based spreading activation model from cognitive psychology. In [2], two algorithms called IUNIS (Inferring User Need by Information Scent) and WUFIS (Web User Flow by Information Scent) were proposed to respectively predict information goal(s) based on pages visited in a session and expected usage patterns based on information goals. The IUNIS algorithm is most directly relevant to our problem formulation. However, it was designed and tested on sites where text was the primary (though not exclusive) modality for conveying information. To understand the behavior of IUNIS in media-rich sites, consider the user session presented in Figure 1. In this session the user viewed three web pages containing images of interacting galaxies on the SkyServer, which is an internet portal to the multi-Terabyte Sloan Digital Sky Survey, the largest digital astronomy archive to date. The SkyServer contains many media-rich sections containing among others, images of galaxies, stars, and other astronomical entities. In this session, the images in the final page were of the interacting galaxies: *ugc1597*, *ngc799-800*, *ugc1077*0, *ugc08584*, *ngc428*, *arp240*, *ngc7603*. Given the specific nature of the session, it may be reasonably assumed that at least some of these galaxies were related to the user information need. On applying the IUNIS algorithm to this data, however, the top 15 putative information goals for this session were identified to be: *famous*, *place*, *tooltile*, *quasar*, *dr1*, *active*, *collision*, *pair*, *core*, *seyfert*, *navigate*, *chart*, *ugc*, *two*, and *call*. Remarkably, none of the interacting galaxies figured in this list, even though the corresponding images occupied significant portions of the final page visited. The results highlight the need to develop techniques for determining information goals in media-rich settings.



http://skyserver.sdss.org/dr1/en/
http://skyserver.sdss.org/dr1/en/tools/places/
http://skyserver.sdss.org/dr1/en/tools/places/page5.asp

Sloan Digital Sky Survey / SkyServer
Famous Places
Galaxies
Spirals
Clumps
Clusters
Interactions
Artifacts
Famous Places - Interactions

Fig. 1. Thumbnails of three image dominated pages from the Skyserver which were traversed during a user session. The URLs of these pages are shown on top of the thumbnails.

3. PROPOSED METHOD

The proposed method begins with a preprocessing step which consists of two phases. In the first phase, the HTML parser (http://htmlparser.sourceforge.net) is used to extract the content of the pages and separate the informational content from the navigational and ornamental motifs. Next, if frames are present, then the contents of each frame are extracted and combined consistently; that is, corresponding sections (menus, main content, etc.) are put together. This allows the overall structure across the frames to be retained. If a page does not have well defined subsections then the entire body is used as the page content. Subsequently, three matrices are constructed; the site connectivity matrix $C(page \times page)$ describing the interlinking of the pages in the site, the term matrix $T(page \times term)$ which captures the term frequency at each page, and the media occurrence matrix $M(page\times media)$ which captures occurrences of any media in each of the pages. In the second phase, the usage log is analyzed for user and session identification and identification of valid page access. User identification is done through unique combination of IP address and browser type. Session segmentation is based on a simple heuristic: a new session is designated when the time difference between consecutive page views from a specific IP address exceeded 30 minutes from the current access. Finally, human users are distinguished from programs such as spiders or administrator-scripts that commonly access site content using the method described in [13]. In the following subsections, we describe the key steps that occur after the preprocessing stage.

3.1. Text-Based Content Modeling and Description

Textual content in a page is analyzed using a grammarless statistical method, which includes stemming and stop word filtration. A variation of the TFIDF method is used to describe the text-based content. This version, which we call DTFIDF (dynamic-TFIDF), uses a dynamic background document set in determining the weights associated with terms. If a web-page P_i is represented by a normalized term frequency vector and P is the set of all pages in the web site, then the DTFIDF value for each term in P_i is calculated as shown in Eq. (1).

$$DTFIDF = \left(\frac{tf}{t_{total}}\right)\times \log\left(\frac{|N|}{|\{e \mid e\in N, t\subset e\}|}\right) \qquad (1)$$

In Eq. (1) tf denotes the frequency of term t in page P_i, containing t_{total} terms. Further, N denotes the set of pages in P such that each page e in N is linked to the page P_i, with the links in either direction counted. Further, pages in N are required, in terms of their content, to be similar to P_i. To enforce this, a similarity score r_{de} is computed between P_i and a page e linked to P_i. The page e is included in N only if the score r_{de} exceeds a threshold k. The similarity score between two pages is determined as the Pearson correlation between their normalized term frequency vectors. In our research $k = 0.85$. Thus, the background set is dynamic in that it uses pages within the site that are both linked to the page being analyzed and contain similar content. In contrast to the standard formulation of inverse document frequency, Eq. (1) is more selective as it is designed to consider only those pages that are expected to be related (both content-wise and in terms of the link connectivity) to the page being analyzed. At the end of this step, each page in the web site is represented by a term vector T_p containing terms having high DTFIDF values.

3.2. Image-Based Content Modeling and Description

Media (image)-based content is another important mode for expressing information in web pages. Unfortunately, determining the semantics associated with images, even when used as part of well structured web sites is highly complicated and at the state of the art, an open problem. In our work, this challenge is ameliorated by associating with an image its proximal text and thereby estimating the semantics associated with the image. Such an approach requires solving three sub-problems: (1) assigning meaningful text annotations to images, (2) dealing with images that are used in multiple contexts with possibly related yet different semantics associated with them, and (3) identifying images that serve only layout or navigational purposes and are consequently unrelated to the information content of the page.

For the first problem, the text associated with an image is drawn from the image URL, the ALT text attribute, page title, anchor text, and text surrounding the image. Solving the last two problems requires determining the signal-level similarity of images in a website. For images determined to be perceptually identical (or highly similar), we capture the variability in the associated semantics by combining the key-terms assigned to the corresponding images in different pages. Since images serving navigational or layout purposes tend to be re-used often and for unrelated topics, the key terms associated with them can be expected to be highly diverse. Our solution for identifying such images uses the information-entropy of the annotation associated with an image as a numeric measure of its heterogeneity. Images with annotation having high entropy are considered to be navigational/ornamental, and are excluded from subsequent analysis. This strategy requires efficiently and accurately computing the similarity of images. Color and texture are two key components of visual appearance and pre-attentive similarity. Consequently, they are used by us to compare images. Specifically, we use the JSEG [4] color/texture analysis system to segment and identify textures within the image. To characterize texture, Grey-Level Co-occurrence Matrices (GLCM) are used along with the four statistical properties: energy, entropy, contrast, and homogeneity. Additionally, a low-resolution color histogram is generated. Finally, relative size, energy, entropy, contrast, homogeneity, and the color histogram are combined to create a feature vector to describe every image. The similarity score between two images is computed as the Pearson's distance between their respective feature vectors. The maximal correlation score of 1.0 indicates identical images. Scores close to this value indicate visually similar images. The correlation threshold is a parameter that has to be established case-by-case depending on the nature of the content. For example, for the Skyserver website, we empirically established that images having correlation scores greater than 0.94 invariably captured the same or highly similar astronomical entities. For such images, corresponding key terms were combined.

3.3. Accounting for the Media-Based Information

Terms that co-occur in the term-frequency matrix and image annotations are re-weighted to ensure that the effect of image size and complexity is reflected in the term weight. Specifically, if a term t with frequency f appeared in a page P and in the annotation related to the image I of size I_x

pixels and texture count T_c, then the frequency of the term is re-calculated as shown in Eq. (2) below, where f_{new} denotes the updated term frequency:

$$f_{new} = f + (\log(T_c) \times \log(I_x)) \quad (2)$$

The purpose of term re-weighting is to increase the importance of terms that are associated with large and visually noticeable images. Essentially, this step emphasizes the contribution of media (image)-based content to the overall information content of a page. Also, the term-image association can be used to retrieve the images that constitute the information goal(s).

3.4. Rethinking the Computation of Information Scent

Information scent is the subjective perception of the value and cost of information sources obtained from proximal cues (snippets and graphics associated with the link) representing the distal content (content at the other end of a link). Using this notion, a framework for predicting user behavior can be developed by postulating that the assessment of distal content and the decision to traverse a hyperlink is taken by users based on the "scent" of the distal content expressed through proximal cues. The basic formulation for computing information scent was proposed in [7] using an associative retrieval technique, called spreading activation [1] for identifying pages related and relevant to currently viewed pages. In this formulation the content and hyperlink networks of the site are represented as graphs (called activation graphs). Given an activation graph $\boldsymbol{G}$, the off-diagonal elements $\boldsymbol{G_{ij}}$ represent the strength of association between the pages $\boldsymbol{i}$ and $\boldsymbol{j}$, and the diagonal elements are set to zero. The association strengths determine how much activation flows from one node (page) to another. Next, the dynamics of activation over discrete time steps t is modeled using a recursive relationship as follows:

$$A(t) = ((1-\gamma)I + \alpha G)A(t-1) + E \quad (3)$$

In Eq. (3), the vector $E = \{E_i\}$ denotes the set of source activations being pumped into the network, I is the identity matrix, α the decay parameter controlling the amount of activation that can spread from one node to its neighbor, and $\gamma < 1$ determines the relaxation of node activity back to zero when it receives no additional activation input. Information goals (terms) are determined by pumping activation through the topology matrix and multiplying the resultant activity vector by the TFIDF weights of terms in the pages. Three issues merit further analysis here. *First*, the content of a site in the above formulation is explicitly text-based. *Second*, implicit in the above formulation is the notion of *content pages*, which are pages that have greater relevance towards satisfying the user information goal than other pages in the session. Content pages are given higher weight in computing the scent [2] and identified using one of the following strategies: predefinition, treating the last pages of a session as content pages, or using page access frequency in a manner mirroring the idea of TFIDF weighting to identify pages that are common to different sessions [2]. However, these methods all have important limitations: predefining the content pages assumes that such a distinction is possible as a consequence of web-design alone. Such a simplification becomes acutely limiting in the presence of multimedia based information owing to the emergent nature of its semantics. Similarly, treating the last page(s) of a session as the content page(s) is limiting in cases where the user has multiple information goals or is lost in the website. Finally, page access-based weighting can not capture the influence of specific user context and runs into limitations if a content page is common to a large number of sessions. The *third issue* relates to the fact that information scent propagation critically depends on the parameters α and γ in Eq. (3), however, there no data driven way to determine them. Moreover, [7] indicates that in highly interconnected sites, for $\alpha > \gamma$, the total activation does not converge and inputs of activation at any node tend to create the same pattern of activation. Unfortunately, high-interconnectivity is inherent to many modern websites.

We propose rethinking the propagation of information scent to address the above issues. Incorporating media-based semantics is addressed by us through the approach described in Section 3.3. To determine the content pages in a manner that takes into account user specificity, we note that web sites are typically organized from broad and general topics to specific information. For example, in a website such as Amazon, users proceed from a general page displaying various product categories to pages related to highly specific products. Therefore, if the browsing pattern of a specific user is considered in combination with some measure of the information specificity of pages, then, the general-to-specific nature of web site organization taken in conjunction with nature of user information seeking behavior will imply that in a purposive (i.e. non-random) browsing pattern, pages with highly specific information will typically be content pages. Given a web page $\boldsymbol{P}$, represented by its unified term vector, $\boldsymbol{T} = [t_1, \ldots t_n]$, obtained as described in sections 3.1-3.3, we define its information specificity as the Shannon entropy of the semantic annotation of a given page:

$$H(P) = -\sum_{i=1}^{n} p(t_i)\log_2 p(t_i) \quad (4)$$

The probabilistic importance of term t_i, $p(t_i)$ is calculated as show in Eq. (5), where $w(t_i)$ represents the DTFIDF weight of the i^{th} term in $\boldsymbol{P}$.

$$p(t) = w(t) / \sum_{i=1}^{n} w(t_i) \quad (5)$$

The page entropy incorporates the contribution of both image-based and text-based page content and is inversely related to the information specificity of a page. Next, the content pages are estimated as follows: given $[P_1, \ldots P_m]$, with the corresponding page entropy values $[H_1, \ldots H_m]$, the putative content page(s) are defined as the page(s) corresponding to the local minima of the sequence of page entropy values given the constraint that two local minima must be at least k+1 steps apart unless they have similar page entropy values. The similarity threshold and k are predefined (we use 5% and 2 respectively in all our experiments). The sole purpose of this criterion is to avoid cluttering of content pages. If two minima occur within *k*-steps of each other, then the page with the lowest entropy value is selected as the content page. Finally, terms which occur in content pages are preferentially weighted. In our experiments we increment the weight of these terms by a factor of five. For further details of this method, see [12].

To address the third issue, we begin by noting that scent propagation is manifested by re-weighting terms (information entities) based on some form of relationship between pages (e.g. the site connectivity) and the functional form driving the propagation. As opposed to Eq. (7), our approach to scent propagation is based on the actual sequence of pages $\Pi = \{P_1, P_2, ... P_k\}$ visited in a session. The sequence Π clearly constitute the most unbiased estimation of the user perception-action as (partially) driven by the information scent. However, users may also follow a link because it serves some usability function (e.g. returning to the home page). Clearly, usability-links do not semantically relate the content of the connected pages and therefore need to be discounted during scent propagation. This is done as part of the scent propagation as follows: given a sequence of pages visited by the user in a session, the information scent is back-propagated from the child page P_{i+1} to the parent page(s) P_i. The strength of the propagation is based on how similar the content of these pages are, as determined by the Pearson correlation (d_{CP}) between the contents of each parent page (P) and child page (C). Since Π may include cycles, we need to ensure that a page does not back-propagate annotations to itself. Cycles can take one of two forms: a session may revisit a page and stop or it may contain multiple revisits to a page, e.g. $\{P_1, P_2, P_3, P_2, P_4, P_5, P_2, P_6\}$. In the former case, the cycle is broken, and the scent is propagated from child page to parent page. In the latter case, pages having multiple children (e.g. the page P_2 in the above example) are detected and the maximum correlation value (d_{CPmax}) across the child pages is stored. Subsequently, if a term t has importance value of t_c in the child page and t_p in the parent page and the Pearson correlation between the content in the child page and the parent page is d_{cp}, then the weight w_p of the term t in parent page is re-calculated as shown in Eq. (6).

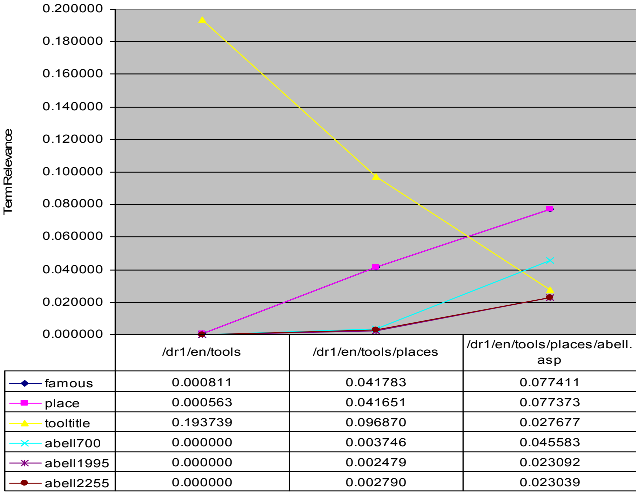
$$w_p = t_p + d_{cp} \times \lambda \times t_c \qquad (6)\text{, where:}$$

$$\lambda = \frac{d_{CP}}{d_{CP\max}} \; if \; d_{CP} > 0 \forall P \text{ and } \lambda = 0 \text{ otherwise.}$$

Next, the highest weighted terms are identified as information goals. Unlike Eq. (3), the proposed formulation is neither limited by web-site topology nor does it require hand-crafted parameters. Most importantly, it captures the emergent behavior characteristic of user-media interactions

4. EXPERIMENTAL INVESTIGATIONS

We begin by analyzing a user session that focused on the "*Famous Places*" section of the SkyServer website (http://cas.sdss.org/dr7/en/) and contained pages with a significant amount of media-based (graphics and image) content. The session started at the "*Tools*" page that has only textual information followed by the "*Famous Places*" page that has images as dominant content. Next, the user went to the *"Abell"* galaxy cluster page which consists of various images and information about the Abell galaxy clusters. The top ten information terms determined using IUNIS for this session were: *famous, place, tooltitle, dr1, 509, navigate, data, file, csv, program*. In comparison, the top ten information terms determined using the proposed approach were: *famous, place, abell700, tooltitle, dr1, abell1995, abell2255, abell1218, abell2197, navigate.* The reader may note that the proposed approach captured information associated with image-based data such as the galaxy names/images *abell700, abell1995, abell2255*, *abell1218,* and *abell2197.* Moreover, it also captured information from text-based content as reflected by the presence of terms such as *famous, place, tooltitle, navigate, dr1* amongst the top information goals. In the next experiment, we track and study the change in the putative information goals as given by the corresponding term-weights for this session as the user moved from the primarily textual content of the first page to the image-dominated content of the second and third pages. This information is presented in Figure 2 below.



	/dr1/en/tools	/dr1/en/tools/places	/dr1/en/tools/places/abell.asp
famous	0.000811	0.041783	0.077411
place	0.000563	0.041651	0.077373
tooltitle	0.193739	0.096870	0.027677
abell700	0.000000	0.003746	0.045583
abell1995	0.000000	0.002479	0.023092
abell2255	0.000000	0.002790	0.023039

Fig. 2. Term relevance scores across pages. The weight of terms associated with images of the Abell clusters is low in the first (text oriented) page and progressively increase thereafter as the user browses to the image-heavy pages. In contrast, the weights of most (but not all) text terms decrease in the latter pages of the session.

In the third experiment the proposed approach and the IUNIS algorithm were applied for identifying user information goals from browsing patterns from six sessions extracted from the Skyserver logs. The relevance score of a term in a session was defined as its maximum weight across pages visited in that session. For each user session, the top five information goals were determined using each of the two methods. These are shown in Table 1 along with the corresponding relevance scores. We observe that the relevancy scores obtained with the proposed method were higher than those obtained using IUNIS. Further, the proposed algorithm was sensitive to variations in browsing patterns. In contrast, INUIS often predicted information goals that rarely varied between sessions. For instance, terms such as *SDSS*, *Project*, *Query*, *Tool*, *Schema* were reported as goals across most of the sessions. This is due to the fact that unlike the proposed approach, IUNIS is not sensitive to user variations.

Table 1. Top five information goals and their relevancy scores

	Term Relevancy Score (Proposed Method)	**Term Relevancy Score (IUNIS)**
1	Cosmic (0.2776), Map (0.2393), Sky (0.0703), Structure (0.0329), Cluster (0.0284)	Universe (0.0194), SDSS (0.0030), Download (0.0013), Tool (0.0010), Project (0.0007)
2	Service (0.3551), Web (0.3201), Map (0.1368), Site (0.1154), Dr1 (0.1001)	Dr1 (0.0257), SDSS (0.0194), Query (0.0181), Project (0.0136), Schema (0.0115),
3	Command (0.3013), Tutorial (0.2953), Order (0.2557), How (0.2461), Help (0.2342)	Query (0.0712), Tool (0.0247), Browser (0.0159), Schema (0.0144), SDSS (0.0031)
4	Group & Win (0.3309), Science (0.2769), Hunt (0.2025), Scavenger (0.2008)	Project (0.0545), Tool (0.0133), SDSS (0.0061)
5	Science (0.2978), Distance (0.1046), Dr1 (0.1001), Tooltitle (0.0952), Simple (0.0890)	Project (0.0572), Dr1 (0.0257), Query (0.0181), SDSS (0.0159), Schema (0.0115)
6	Tooltitle (0.1937), Dr1 (0.1639), Famous (0.0827), Place (0.0827), Ngc450 & Ngc60 (0.0488)	Famous (0.1962), Place (0.1608), Ned (0.049), Tool (0.0231), Query (0.0144)

The final experiment, involved a user study to test the validity of information goals determined by our method. The study involved eight participants. Four frequently recurring user sessions were selected from the usage logs and analyzed using the proposed approach. The sequence of pages for these sessions were printed and bound in chronological order, producing four booklets. Each participant was asked to study three of these booklets and determine which predicted information goal best describes the information of the given booklet. Overall, the users agreed with the information goals predicted by the proposed approach 15 out of 24 times. The *t*-value for 15 agreements for a binomial distribution where n=24 and p=0.1 is 42.0, which strongly rejects the null hypothesis of no correlation. It should be noted that of the 9 incorrect answers, 5 were due to confusion with a very similar set of information goals which shared 60% of the terms, and 25% of the images with the predicted set of information goals. To cross-evaluate the previous question, users were also asked to rank all 10 sets of goals on a Likert scale from 1 (not relevant) to 5 (highly relevant) with 3 (neutral) as the mid-point. Users scored the proposed approach's set of goals high, (with mean score μ = 4.38, out of 5) compared to all other sets of goals (μ =2.56), indicating that the proposed approach predicted distinguishably relevant information goals for a given user session.

5. CONCLUSIONS

In this paper we have investigated the problem of estimating information goals from user browsing patterns by building on the information foraging theory and fundamentally rethinking information scent propagation. Experimental results indicate the efficacy of the proposed method. The proposed approach can be expected to have significant impact on development of techniques for user-goal identification, new paradigms for search in media-rich settings, modeling of website usability, and in applied problems such as user-context driven advertising.

6. ACKNOWLEDGEMENTS

The authors thank Mike Wong for his participation in parts of this research. Jim Gray from Microsoft Research showed great enthusiasm for this research, participated in numerous discussions, and was instrumental in providing access to the Skyserver logs. This work was funded, in part, by a Microsoft research grant to RS.

7. REFERENCES

[1] J. R. Anderson, "A Spreading Activation Theory of Memory" Verbal Learning and Verbal Behavior, vol. 22, pp. 261-295, Aug. 1983.

[2] E. Chi, P. L. Pirolli, K. Chen, J. Pitkow, "Using Information Scent to Model User Information Needs and Actions on the Web" ACM CHI 2001, pp 490 – 497

[3] N. Craswell, D. Hawking and S. Robertson "Effective site finding using link anchor information", ACM SIGIR 2001

[4] Y. Deng and B. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. on PAMI, vol. 23, no. 8, pp. 800-810, 2001

[5] I Kang and G. Kim, "Query Type Classification for Web-Document Retrieval", ACM SIGIR 2003

[6] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search", WWW 2005

[7] Pirolli P., Pitkow J., and Rao R., "Silk from a Sow's Ear: Extracting Usable Structures from the Web", ACM Conference on Computer Human Interactions, pp. 118-125, 1996

[8] P. Pirolli and S. Card, "Information Foraging", Psychological Review, 106 (4), pp 643-675, 1999

[9] P. Pirolli, "A Theory of Information Scent", In J. Jacko and C. Stephanidis (Eds), Human-Computer Interaction, Vol. 1, pp. 213-217, Mahwah, NJ: Lawrence Erlbaum

[10] D. E. Rose and D. Levinson, "Understanding User Goals in Web search", WWW, pp. 13-19, 2004

[11] S. Santini, A. Gupta, and R. Jain, "Emergent Semantics Through Interaction in Image Databases", IEEE Trans. on Knowledge and Data Engineering, Vol. 13, No. 3, pp. 337-351, 2001

[12] R. Singh, and B. D. Bhattarai, "Information-theoretic identification of content pages for analyzing user information needs and actions on the multimedia web", ACM Symposium on Applied Computing, pp. 1806-1810, 2009

[13] V. Singh, J. Grey, A. Thakar, A. S. Szalay, J. Raddick, B. Boroski, S. Lebedeva, and B. Yanny, "SkyServer Traffic Report – The First Five Years", Microsoft Technical Report, MSR TR-2006-190, December 2006