Multimed Tools Appl (2010) 50:491–507
DOI 10.1007/s11042-010-0487-1

Dynamic content-page identification for media-rich websites

Rahul Singh · Bibek D. Bhattarai

Published online: 26 February 2010

Abstract Knowledge of the *information goal* of users is critical in website design, analyzing the efficacy of such designs, and in ensuring effective user-access to desired information. Determining the information goal is complex due to the subjective and latent nature of user information needs. This challenge is further exacerbated in media-rich websites since the semantics of media-based information is context-based and emergent. A critical step in determining information goals lies in the identification of *content pages*. These are the pages which contain the information the user seeks. We propose a method to automatically determine the content pages by taking into account the organization of the web site, the media-based information content, as well as the influence of a specific user browsing pattern. Given a specific browsing pattern, in our method, putative content pages are identified as the pages corresponding to the local minima of page-content entropy values. For an (unknown) user information goal this intuitively corresponds to modeling the progressive transition of the user from pages with generic information to those with specific information. Experimental investigations on media rich sites demonstrate the effectiveness of the technique and underline its potential in modeling user information needs and actions in a media-rich web.

Keywords Multimedia information systems · Information goal · WWW · Content pages · Entropy · Surfing behavior · User-media interaction

1 Introduction

The success of a website depends critically on users finding the information they seek (i.e. satisfying their *information goal*). 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. Intuitively, the information content of certain pages visited during the traversal has greater relevance towards satisfying the user information goal than the information in other pages. These pages are called

R. Singh (✉) · B. D. Bhattarai
Department of Computer Science, San Francisco State University, San Francisco, CA 94132, USA
e-mail: rsingh@cs.sfsu.edu

content pages. Additionally, a website contains a variety of other pages, such as pages providing introduction to the site content, pages with general information, and navigational pages. Content pages in a web-site may be identified *statically* or *dynamically*. In the former case, the content pages are assumed to only be the consequence of the site design/content and therefore, are predefined. In contrast, dynamic content page determination requires factoring in not only the site content and structure but also the variability of specific user information need and behavior.

The ability to identify content pages is valuable in many contexts. For instance, such information can be used for improving page ranking [5, 9] and result presentation [11]. It can also be used to evaluate site usability, for example, by comparing expected paths to the content pages with the actual paths taken by users; divergences, if any, may point to potentially problematic aspects of the site design [4, 11]. Other applications include prediction of usage patterns [2, 4] and making pages containing information goals shared by a large number of users easily accessible [4, 11]. Knowledge of content pages can also be helpful in adaptive hypermedia systems [3] which try to adapt aspects of the system to user characteristics, including user information goals.

Current studies identify two predominant modes through which a user can access information on the web to satisfy their information needs [15]: by searching, also called "search by-query" or by browsing, called "search by-navigation". The complexity of determining content pages significantly varies across these modes. 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 information goals and queries (keywords). In the search by-navigation scenario however, the under-constrained nature of the task coupled with lack of precise cues (such as query terms), significantly increases the complexity of the problem. In this paper, we extend our work in [20], to investigate the problem of dynamic content page determination in a search by-browsing context. We propose a method for content page identification that takes into account both the specificities of a user's behavior as well as the content and structure of the website.

1.1 Problem formulation and characteristics

Given a sequence of pages visited by a user, our problem is to identify the pages which putatively contain information satisfying the information goal of the user. Formally, this problem can be stated as follows: given a web-site *W*, and a sequence of pages $\Pi = \{P_1, P_2, \ldots P_k\}$ in *W* visited by a user in a session, identify Δ, the proper subset of pages from Π (Δ⊂Π) which constitutes the set of content page(s).

An important facet of the problem is implicit in the formulation, namely the interplay between the site-content and specific user behavior in determining the content pages. This follows from the fact that the content pages are explicitly defined in terms of user-specific browsing patterns and are consequently sensitive to both user context and data context. The importance of user specificity is crucial for media-rich site content, since it is now well known that 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 [18, 19].

2 Prior research

In the search by-query scenario, utilizing the history of user click-behavior and anchor-link distribution has been proposed for content page and user goal determination [10, 14]. The basic idea behind using click behavior lies in correlating the common page(s) corresponding to specific query terms as determined by the links clicked by users after issuing a query through a search engine. In the search by-navigation scenario, however, precise cues (such as query terms), are unavailable. Consequently, a different set of ideas has been employed. This includes: (1) predefining the content pages for a site [4, 11, 16], (2) treating the last page(s) of a session as the content page(s), and (3) page access-based weighting. However, all these methods have important limitations: the simple and often used idea of predefining the content pages assumes that such a distinction is possible as a consequence of web-design alone. This strategy discounts the variability in user information needs. The limitations of such a simplification become acute 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. The idea of page access-based weighting [4] seeks to use access frequency in a manner mirroring the idea of TFIDF weighting strategy from text analysis [17] to identify pages that are common to different sessions and down-weight them. However, this approach 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.

3 Proposed method

In seeking to solve this problem, we begin by postulating that purposive user actions at a site are a consequence of the user information need and the site content and structure. At a syntactic level, both the user actions and the site content and structure can be assayed. The specific interactions of the user at the site, such as the traversal history, links clicked, or dynamic queries issued, can be obtained from the web-log. The content of the site described through text and/or other media as well as the site connectivity can similarly be quantitatively described. These observations underlie the first of the three facets of the proposed solution.

To motivate the second facet of the proposed method, we note that summary data of the above type is by itself insufficient to obtain insights about the information needs of a user. Doing so requires an additional factor, namely, an explanatory model of user action. Such a model needs to perform (at least) three important functions:

- Provide cues to the perceived value of specific information to a user, given their (unknown) information goal.
- Capture the tradeoff between the perceived value attained by the user through information gain and the perceived cost of collecting the information (e.g. due to time spent, cognitive load assumed, or interaction effort).
- Provide a mechanism which supports obtaining quantitative estimation of the value of specific information.

In this paper, we base ourselves on a cognitive model of web navigation called Information Foraging Theory [13]. 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 theory of information scent [12], which is a psychological theory of how cues, such as links in a web page, are used by users to make information seeking decisions. Information theory and specifically the concept of information scent have been used to predict information goals and simulate usage of a site [4, 11] using a network flow algorithm called spreading activation. As part of these efforts, mechanisms were proposed to quantify the value or saliency of specific information as weights associated with terms describing the information (in a manner akin to TFIDF weighting with modifications).

The third and the final facet of our approach involves site content/structure modeling and the idea of information scent to identify content pages. To explain it, we note that the transition of a user from one page in a site to another (including leaving the site) can occur due to the following reasons: (1) the user anticipates that the information in the subsequent page will better satisfy the information goal (i.e. the likelihood of the current page being a content page is low), (2) user has satisfied the current information goal and browses to the subsequent page to satisfy a different information need (i.e. the current page is a content page and user has additional information needs), (3) the user leaves the site with the information need satisfied (current page is a content page and user has no further information need), or (4) the user leaves the site with the information need unsatisfied (current page is not a content page). Obviously, were the information goals of the user known, the likelihood of each of the above cases could be directly calculated by determining the fit of the page content to the user information need. In absence of this information, however, we can only *estimate* if a page represents a putative content page for a specific user. A critical observation in this context is 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. Similarly, a university website may start with general information of the campus, and then link to a page on academic departments, which in turn links to a list of faculty, which finally links to the page about a particular professor, wherein are listed research interests, office hours, location, and contact information. The key intuition lies in noting that 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.* This idea lies at the core of the proposed method. In the following subsections, we describe in detail each of its steps.

3.1 Usage pattern extraction and data preprocessing

In the first step of our method, the usage log is analyzed during the pre-processing stage 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 min from the current access. Since we are interested in analyzing user behavior at a site, it is important to distinguish human users from programs such as spiders or administrator-scripts that access site content. Following [22], if a client IP address is known to be a spider (based on the *AgentString* variable) or if the client's IP address visited *robots.text*, then all sessions from that IP address are designated as spiders. Finally, a reverse lookup of the IP address contained in the SQL log entries is performed to obtain the institution name corresponding to the IP address. It should be noted, that not all IP addresses could be resolved in terms of their precise location/institution since some addresses mapped to IP-address blocks leased to multiple users/institutions. For further details on how the origin of requests can be used to reason about usage patterns, we refer the reader to [21].

We use the Java-based HTML parser [8] to extract the content of the pages and separate the informational content from the navigational and ornamental motifs. The presence of frames or the absence of well defined subsections in the page can complicate the content extraction process. In the former case, 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. Next, based on the main page content three matrices are constructed; the site connectivity matrix C (*page*×*page*) describing the interlinking of the pages in the site, the term matrix T (*page*×*term*) which captures the term frequency at each page, and the media occurrence matrix M (*page*×*media*) which captures occurrences of media (images, audio, or video) in each page.

3.2 Modeling the page content

The content in a website is described through text and media. For the purposes of this research, we 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, if needed, by using appropriate descriptors. For example, video-based information can be described and matched using MPEG-7 descriptors [1]. The characterization of media-based content is preceded by extraction and pre-processing of the web-page contents as described in the previous section.

Textual content within 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 d is represented by a normalized term frequency vector and D is the set of all pages in the web site, then the DTFIDF value for each term in d is calculated as shown in Eq. 1.

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

In Eq. 1 tf denotes the frequency of term t in page d, containing t_{total} terms. Further, N denotes the set of pages in D such that each page e in N is linked to the page d, with the links in either direction counted. One further condition is applied to pages in N, namely, that their content must be similar to the content of page d. To enforce this condition, a similarity score r_{de} is computed between the page d and a page e which is linked to d. 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 and the value of k=0.85 is used in our research. 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 in that, 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.

Alongside text, 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 approach, 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), our strategy is to capture the variability in the associated semantics by coming up with an annotation consisting of 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 or ornamental, and are excluded from subsequent analysis.

The above strategy requires efficiently and accurately computing the similarity of images. Color and texture are two key components of visual appearance and preattentive similarity. Consequently, they are used by us to compare images. Specifically, we use the JSEG [6] color/texture analysis system to segment and identify textures within the image. To characterize texture, Grey-Level Co-occurrence Matrices (GLCM) [7] are used along with the four statistical properties: *energy, entropy, contrast, and homogeneity.* For a normalized co-occurrence intensity matrix, *M(i, j),* the definition of these properties are as follows (Eq. 2–5):

$$Energy = \sum_{i=0}^{n}\sum_{j=0}^{n}(M(i,j)M(i,j)) \tag{2}$$

$$Entropy = \sum_{i=0}^{n}\sum_{j=0}^{n}M(i,j)\log(M(i,j)) \tag{3}$$

$$Contrast = \sum_{i=0}^{n}\sum_{j=0}^{n}(i-j)(i-j)M(i,j) \tag{4}$$

$$Homogenity = \sum_{i=0}^{n}\sum_{j=0}^{n}\frac{M(i,j)}{1+|i-j|} \tag{5}$$

In addition to the above texture-based descriptors, 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.

Two important observations should be noted here. *First*, the approach for representing images and comparing their perceptual similarity is computationally simple and empirically effective in identifying both identical and highly similar images. However, this method is only meant to efficiently compare similarity at a relatively coarse visual level and not intended to be a refined measure (such as those covered in the review [23]) for content-based retrieval. *Second*, the specific image similarity threshold needs to be calibrated based on data. In Fig. 1, we present five pairs of images from the Skyserver website along with their similarity scores and URLs to illustrate the application of this method in practice. The score for the first pair indicates that the images, from two different locations in the site, are identical. Consequently, the key terms in the annotation of these images are combined to capture the (possible) semantic variability associated with it. The second pair of images is

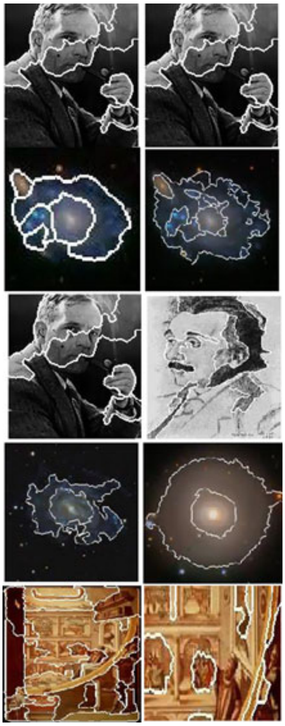


Image-1	Image-2	Similarity Score	Image Source URL Prefix: http://skyserver.sdss.org/dr1/en
		1.0	/astro/universe/images/hubble.jpg /proj/advanced/galaxies/images/edwinhubble.jpg
		0.947	/tools/places/thumb/ngc450.jpg /tools/places/images/ngc450.jpg
		0.715	/astro/universe/images/hubble.jpg /astro/images/einstein.jpg
		0.742	/astro/images/752-1-432.jpg /tools/places/images/ngc2681.jpg
		0.830	/astro/images/mquadrant.jpg /images/new_astronomy_1.jpg

Fig. 1 Examples of matching images using the Pearson distance between the color-texture descriptors of the corresponding images. All images are from the Skyserver website. The segmentation of the images using JSEG [6] is also displayed by dimming the images to emphasize the boundaries. For the first two cases, the image URLs illustrate that perceptually identical images can be used in related semantic contexts within a website. The other cases present example of images that are similar in terms of form, texture, and/or color and yet perceptually distinct

not identical. Yet it describes the same astronomical phenomenon, namely, the galaxy pair NGC450/UGC 807. The similarity of these images is 0.947 and exceeds the empirically established threshold used by us for this site. Consequently, the annotations of these images too are combined. The other examples illustrate the performance of the measure on distinct images which had somewhat similar appearances. Based on the similarity score, all these cases are treated to be distinct.

At this point in the method, the textual content and image content in a web site are represented through a weighted term vector. Next, the terms that co-occur in the term-frequency matrix and image annotations are reweighted 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. 6 below, where f_{new} denotes the updated term frequency:

$$f_{new} = f + (\log(T_c) \times \log(I_x)) \tag{6}$$

The purpose of term re-weighting, as shown in Eq. 6, 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.

3.3 Explanatory model of user behavior and its implementation

Information scent can be thought of as the subjective perception of the value and cost of information sources obtained from proximal cues representing the content. In [4], this notion was used to model and predict user behavior on the web. The basic idea was to postulate that assessment of distal content (page at the other end of a link) was done by users based on snippets and graphics associated with the link. However, the central problem of our research was unaddressed in [4]: in it, the identification of content pages was not based on user-data interactions as determined through information scent. Rather, the content pages were independently identified (using one or more of the strategies covered in Section 2) and information in these pages was preferentially weighted. Our method, described below, is partially based on ideas introduced in [4] and modifies an earlier approach proposed by us in [2] to model the information scent of a site.

Our core idea seeks to model the information scent of content in a page by addressing two issues. *First*, we seek to model the relationship between content in hyperlinked pages while simultaneously excluding links that are not semantically meaningful. This is because most websites provide links that support usability functions. Examples include, among others, link to the home page from every page and links allowing users to jump to any point in the overall organization of a subsection of the site. Clearly, such links, do not semantically relate the concepts of two web pages, and therefore need to be discounted when analyzing information foraging behavior. *Second*, given the fact that most websites present information using a general-to-specific approach, we seek to better reflect in the numeric characterization of the information scent, the possible holonymous relationships that may exist between the content of two pages.

Both these issues are addressed by traversing the links in the site using breadth-first-search. During this process the Pearson correlation, denoted d_{CP} in the following, between the content of each parent page (P) and child page (C) is evaluated. If multiple parent pages link to a child page, then additionally, the maximum correlation value (d_{CPmax}) across the parent pages is also stored. Next, the terms weights in the parent page (which capture the information scent) are updated by back annotation from the child page. Specifically, 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. 7.

$$w_p = t_p + d_{CP} \times \lambda \times t_c \tag{7}$$

Where $\lambda = \frac{d_{CP}}{d_{CP\max}}$ *if* $d_{CP} > 0 \, \forall P$ and λ=0 otherwise.

In Eq. 7, the strength of the back annotation is maximal for the parent-child pair that shares the greatest semantic similarity. If the web page organization is such that there is no semantic similarity between two linked pages, then no back-annotation occurs.

3.4 Content page determination

Given a web page x, represented by its unified term vector, T = [t_1, . . .t_n], obtained as a consequence of the previous steps, we define its page entropy as:

$$H(x) = -\sum_{i=1}^{n} p(t_i)\log_2 p(t_i) \tag{8}$$

Where $p(t_i)$ is probabilistic importance of term t_i in the page x and is calculated as show in Eq. 9, where $w(t_i)$ represents the DTFIDF weight of the i_{th} term.

$$p(t) = w(t) / \sum_{i=0}^{n} w(t_i) \tag{9}$$

The reader may note that the page entropy (which we define as the Shannon entropy of the semantic annotation of a given page) gives the inverse of the informational specificity of a page. Furthermore, owing to our modeling of the multimedia content, the semantic entropy incorporates the contribution of both image-based and text-based page content as well as the influence of web-site linkages.

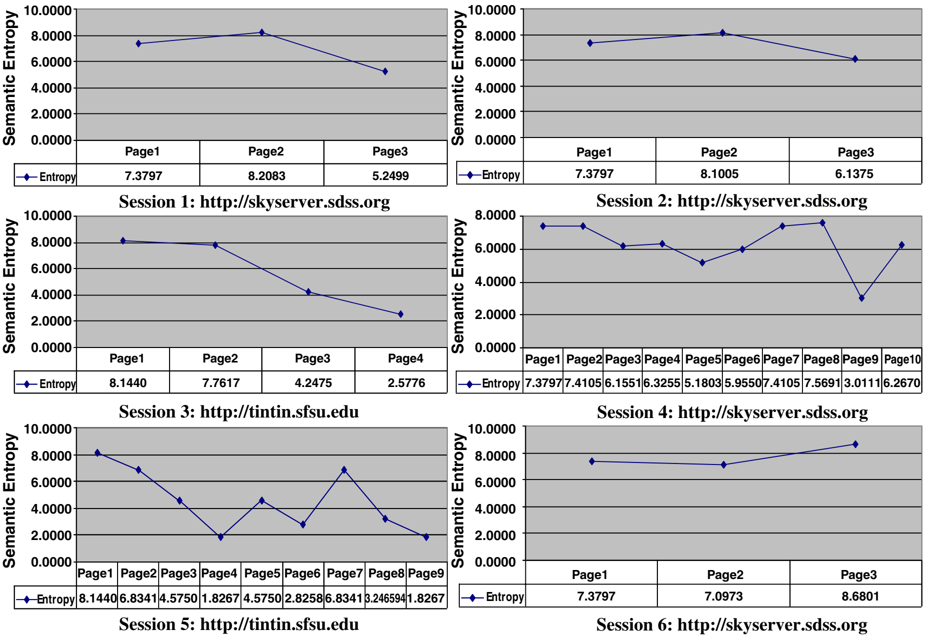
Finally, the content pages are estimated as follows. Let the browsing pattern of a user in a particular session be given by the pages [x_1, . . .x_m], with the corresponding page entropy values [$H(x_1)$, . . .$H(x_m)$]. We call pages x_i and x_j that are k steps apart in the traversal order during a session, as k-neighbors of each other. 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 no two local minima can be k-neighbors of each other 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 are k-neighbors, the page with the lowest entropy value is selected as the content page.

4 Experiments

4.1 Case study

We begin with a case study that investigates six different user sessions. For each of these sessions, the user information goals were known a priori and the sequence of the pages that were visited, were directly observed and noted. The design of the experiment therefore allowed us to compare the actual content pages (as determined by the users) with those predicted using the proposed method. The changes in the page entropy values, as the users navigated from page-to-page, are shown in Fig. 2 for each of the sessions.

In the first session the user was seeking information related to the Sloan Digital Sky Survey (SDSS) telescopes which is available at the Skyserver website. In this case the content page (the last page of the session) had the lowest entropy as identified in the graph for session 1. In the second session, the user information goal was to find images of famous galaxies. The user started from the index page of the Skyserver website followed it by visiting the "Famous Places" page which contains thumbnails of astronomical entities and finally a page with images of different galaxies. The page entropy values plotted in the graph for Session 2, show the final page of this session to have the entropy minima, thus correctly identifying the content page. The user in session 3 chose to search for the biographical information of the first author of this paper by browsing the web-site of our research group. Starting from the index page, the user successively visited the directory page of all the research group members, the personal page of the first author, and finally his biography page. The final page of this session had the lowest entropy as shown in the plot for session 3 and correctly corresponded to the content page. Sessions 4 and 5, unlike the previous cases contained multiple information goals. Specifically, session 4 corresponded to a user who was exploring the educational projects and games sections of the Skyserver. In this case Page-5 (constellation game) and Page-9 (challenges/difficult questions) represented local entropy minima and were correctly identified as content pages by the method.



Session 1: http://skyserver.sdss.org

Session 2: http://skyserver.sdss.org

Session 3: http://tintin.sfsu.edu

Session 4: http://skyserver.sdss.org

Session 5: http://tintin.sfsu.edu

Session 6: http://skyserver.sdss.org

Fig. 2 Case study investigating the proposed method in six different user sessions where the content pages were known a priori. The graphs plot the variation in page entropy for each session. The sessions are numbered and referred to in the text in a row major order. Sessions 1, 2, 4, and 6 were on the Skyserver (http://skyserver.sdss.org/)—a multimedia astronomy website. Sessions 3 and 5 involved the research group website of the authors (http://tintin.sfsu.edu)

Session 5 also had multiple information goals and used the web page of our research group. The goals in this case were to find student(s) who worked on research involving the Skyserver website and those working on the "cross modal information retrieval" project. The session started at the index page of the research website then visited the project directory page of the research group. Next the session visited the project page involving the Skyserver and then the home pages of the two students working on it (page 4 and page 6) with one intervening backtrack to the project page (page 5). Then the user backtracked to group project list page followed by a visit to the Cross-modal information retrieval project page and finally the home page of the student involved in this research (page 9). Based on the changes in the page entropy values, the proposed approach correctly identified Page-4, Page-6, and Page-9 as the content pages. The final session exemplified a case where the proposed content discovery technique failed. In this session, the information goal was the "Glossary" section of the SkyServer, which is the third and final page visited in the session. However, the "Glossary" page contained information about various topics and therefore, intrinsically had high variability of content and correspondingly high semantic entropy. Therefore, the proposed technique incorrectly identified the second page as the content page. We note however, that based on our experience, such special cases occur rarely.

4.2 Accuracy of content page identification

To determine the accuracy of content page identification under variable conditions a user study was designed involving 20 users (all were students of the San Francisco State University and unfamiliar with this research) and five public websites. Each user was given three distinct information goals for each of the five websites and asked to find the relevant information. This setup yielded 300 different sessions (20 users×5 websites×3 information goals). Each of the sessions started from the index page of the websites and the sequence of pages visited by the users was recorded. At the end of each session, users reported the page(s) they found most relevant to the information goal. These pages were considered to be the content pages and constituted the ground truth for this study. Subsequently, for each session, the sequence of traversed pages was analyzed using the proposed method and the content pages identified. These pages were then compared with the ground truth in terms of precision and recall. The websites used in the study were: (1) Skyserver (http://skyserver.ssd.org) (2) News and Sports sections of BBC News (www.bbc.co.uk) (3) The San Francisco State University (SFSU) website (www.sfsu.edu) (4) The Computer Science department website at SFSU (www.cs.sfsu.edu), and (5) the research web-site of the authors (http://tintin.sfsu.edu). Examples of information goals assigned to the users as part of this study are presented in Table 1. The second column in the table indicates if the information goal was media or text oriented.

Depending on the information goal and the user response, the sessions could be grouped into the following three categories: (1) User identified a single page as the content page (henceforth referred to as Type-I sessions), (2) users identified two or more pages as informative (referred to as Type-II sessions), and (3) users found all traversed pages as informative (referred to as Type-III sessions). Of the 300 total sessions, 160 sessions were identified as Type-I, 121 sessions had multiple content pages (Type-II), and in 6 sessions, the corresponding users failed to clearly identify a single content page and considered all the pages to be informative (Type-III sessions). Finally, in 13 sessions, the user got lost or failed to find the relevant information and left the site. These 13 sessions were excluded from analysis, since the ground truth could not be established for them.

Table 1 Examples of information goals given to the users in the study

Website	Information Goal Focus	Example Information Goals
Skyserver	Text and Media Oriented	Find (1) introduction for skyserver tools, (2) information about the galactic census, (3) images of spiral galaxies.
BBC News	Text and Media Oriented	(1) Find images from the 2007 T20 Cricket championship final game (2) find report on Barak Obama's victory in the democratic primaries (3) what type of tomatoes have anti-cancer properties?
San Francisco State University (SFSU)	Text and Media Oriented	(1) How many departments constitute the college of science and engineering (COSE) (2) Find information about the student center (3) What types of scholarships are available for students in COSE.
Computer Science at SFSU	Text Oriented	(1) Find faculty involved in the multimedia and visualization laboratory (2) what are the protocols for human subject research, (3) Where did Dr. Chung-Sheng Li, member of the department advisory board get his doctorate from?
Author's Research Website	Text Oriented	(1) List the collaborators in the XMAS project, (2) which students are involved in the molecular visualization project (3) How many papers have been published in the area of experiential information management

The performance of the method across these sessions was quantified using precision and recall values computed in terms of the number of content pages correctly retrieved. The results are presented in Table 2. The method performed best in cases of a single content page. In a predominant number of cases (both in terms of precision and recall), the content page corresponded to the entropy minima and was correctly captured by the proposed method. For sessions which had multiple pages, the performance was observed to degrade somewhat. Among others, this was caused due to the fact that in certain sessions, users considered successive pages to be important. Due to the *k*-neighbor constraint, content page(s) immediately following the first content page were often excluded. The lowest recall values were observed for Type-III sessions, where users considered all pages to be informative. It may be noted that not only were such cases relatively few, but also that they did not fit the problem formulation (since in such cases the set of content pages was not a proper subset of the set of all pages) as well as the intuitive notion of content pages.

4.3 Identification of user information goals

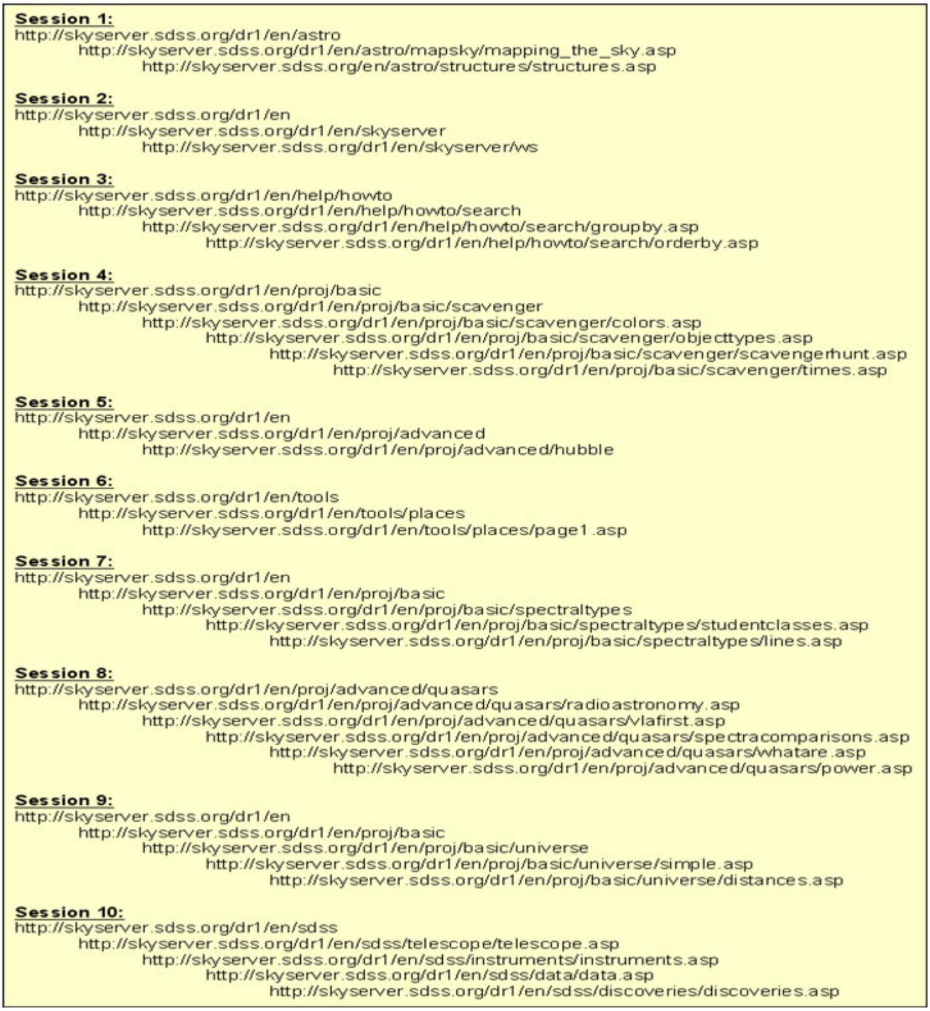
In this experiment the proposed approach was applied for identifying user information goals from browsing patterns. The results obtained with the proposed approach were then

Table 2 Performance of the method in terms of precision and recall

Session Type	Total Count	Precision	Recall
Type-I	160	83.44%	81.87%
Type-II	121	71.91%	73.94%
Type-III	6	100%	30.77%

compared with those obtained using the INUIS (Inferring User Need by Information Scent) algorithm [4], which is one of the established methods in this area.

Ten user sessions were analyzed from the Skyserver logs. These sessions are shown in Fig. 3. 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 ten information goals were determined using each of the two methods. For the purpose of comparison, the top five information goals and their corresponding relevance scores are shown in Table 3. Two important observations can be made based on this table. First, the relevancy scores for information goals obtained using the proposed method were higher than those obtained using IUNIS. Second, the proposed algorithm was more sensitive to variations in browsing patterns. In contrast, INUIS often predicted information goals that, while reflecting the overall content of this section of the website, varied little between sessions (for example terms such as *SDSS*, *Project*, *Query*, *Tool*, *Schema* were common to most of the sessions). The inherent differences between the proposed method and IUNIS provide a clear mechanistic explanation of these observations. In the



```
Session 1:
http://skyserver.sdss.org/dr1/en/astro
        http://skyserver.sdss.org/dr1/en/astro/mapsky/mapping_the_sky.asp
                http://skyserver.sdss.org/en/astro/structures/structures.asp

Session 2:
http://skyserver.sdss.org/dr1/en
        http://skyserver.sdss.org/dr1/en/skyserver
                http://skyserver.sdss.org/dr1/en/skyserver/ws

Session 3:
http://skyserver.sdss.org/dr1/en/help/howto
        http://skyserver.sdss.org/dr1/en/help/howto/search
                http://skyserver.sdss.org/dr1/en/help/howto/search/groupby.asp
                        http://skyserver.sdss.org/dr1/en/help/howto/search/orderby.asp

Session 4:
http://skyserver.sdss.org/dr1/en/proj/basic
        http://skyserver.sdss.org/dr1/en/proj/basic/scavenger
                http://skyserver.sdss.org/dr1/en/proj/basic/scavenger/colors.asp
                        http://skyserver.sdss.org/dr1/en/proj/basic/scavenger/objecttypes.asp
                                http://skyserver.sdss.org/dr1/en/proj/basic/scavenger/scavengerhunt.asp
                                        http://skyserver.sdss.org/dr1/en/proj/basic/scavenger/times.asp

Session 5:
http://skyserver.sdss.org/dr1/en
        http://skyserver.sdss.org/dr1/en/proj/advanced
                http://skyserver.sdss.org/dr1/en/proj/advanced/hubble

Session 6:
http://skyserver.sdss.org/dr1/en/tools
        http://skyserver.sdss.org/dr1/en/tools/places
                http://skyserver.sdss.org/dr1/en/tools/places/page1.asp

Session 7:
http://skyserver.sdss.org/dr1/en
        http://skyserver.sdss.org/dr1/en/proj/basic
                http://skyserver.sdss.org/dr1/en/proj/basic/spectraltypes
                        http://skyserver.sdss.org/dr1/en/proj/basic/spectraltypes/studentclasses.asp
                                http://skyserver.sdss.org/dr1/en/proj/basic/spectraltypes/lines.asp

Session 8:
http://skyserver.sdss.org/dr1/en/proj/advanced/quasars
        http://skyserver.sdss.org/dr1/en/proj/advanced/quasars/radioastronomy.asp
                http://skyserver.sdss.org/dr1/en/proj/advanced/quasars/vlafirst.asp
                        http://skyserver.sdss.org/dr1/en/proj/advanced/quasars/spectracomparisons.asp
                                http://skyserver.sdss.org/dr1/en/proj/advanced/quasars/whatare.asp
                                        http://skyserver.sdss.org/dr1/en/proj/advanced/quasars/power.asp

Session 9:
http://skyserver.sdss.org/dr1/en
        http://skyserver.sdss.org/dr1/en/proj/basic
                http://skyserver.sdss.org/dr1/en/proj/basic/universe
                        http://skyserver.sdss.org/dr1/en/proj/basic/universe/simple.asp
                                http://skyserver.sdss.org/dr1/en/proj/basic/universe/distances.asp

Session 10:
http://skyserver.sdss.org/dr1/en/sdss
        http://skyserver.sdss.org/dr1/en/sdss/telescope/telescope.asp
                http://skyserver.sdss.org/dr1/en/sdss/instruments/instruments.asp
                        http://skyserver.sdss.org/dr1/en/sdss/data/data.asp
                                http://skyserver.sdss.org/dr1/en/sdss/discoveries/discoveries.asp
```

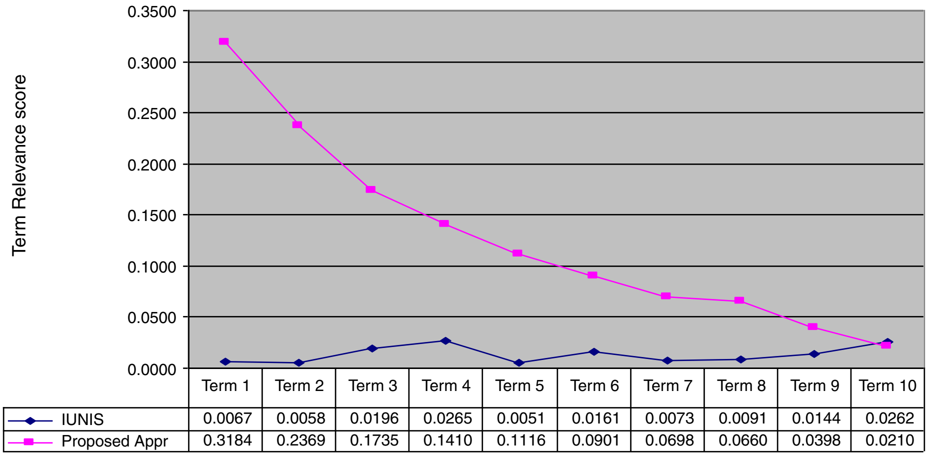
Fig. 3 Ten user sessions used for information goal prediction and comparison with IUNIS

Table 3 Top five terms along with their term relevancy scores for each session from Fig. 3, obtained using the proposed method and the IUNIS algorithm. Terms having the same score are separated by ampersand. Only terms with non-zero relevancy scores are shown

Session Number	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)
7	Science (0.2769), Classify (0.0944), Line (0.0850), Identify (0.0790), Star (0.0567)	Project (0.0545), Dr1 (0.0257), Query (0.0181), SDSS (0.0159), Schema (0.0115),
8	Typical & Third & Strange (0.0266), Spectrum (0.0241), Research (0.0214)	Project (0.0263), SDSS (0.0160), Tool (0.0054)
9	Distance (0.3745), Science (0.2769), Simple (0.2313), Universe (0.2098), Galaxy (0.0425)	Project (0.0545), Dr1 (0.0257), Query (0.0181), SDSS (0.0159), Schema (0.0115)
10	Instrument (0.4790), Discovery (0.4543), Telescope (0.3807), SDSS (0.0571), CCD (0.0309)	SDSS (0.0160), Schema (0.0021), Place (0.0017), Dr1 (0.0015), Project (0.0004)

proposed method, dynamic determination of content pages not only ensured that these pages were sensitive to variations in usage patterns but also allowed allocation of greater weights to the terms in the (dynamically determined) content pages. Because of this, the variations in usage patterns were found to correspond to the variability in information goals. In the case of IUNIS however, the content pages were determined statically. Consequently, both the structure and the content of site had a much greater influence on the predicted information goals than the variability in user information need. Therefore, the information goals determined by IUNIS showed little variability across the sessions.

To provide a synopsis of the difference between the relevance scores for the top ten terms using each of the methods, the average relevance scores were calculated for each of the top 10 information goals (terms) across all sessions. As shown in Fig. 4, the relevance of the top 10 goals determined using the proposed method, scored up to 47.6 times more strongly than the relevance of the information goals obtained using IUNIS. Further, the mean increase in term relevancy using the proposed method was 15.06 time that of IUNIS. Moreover, terms associated with images found greater relevance. Thus, by directly modeling the influence of media-based information and user variability, a different and arguably more complete understanding of information scent and user behavior was obtained.



	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10
IUNIS	0.0067	0.0058	0.0196	0.0265	0.0051	0.0161	0.0073	0.0091	0.0144	0.0262
Proposed Appr	0.3184	0.2369	0.1735	0.1410	0.1116	0.0901	0.0698	0.0660	0.0398	0.0210

Fig. 4 Comparison of the average relevancy of the top ten user goals determined using the proposed approach (top curve) and IUNIS (bottom curve) across ten user sessions on the SkyServer

As alluded above, the sessions described in Fig. 3, provide insight about another facet of the proposed method, namely, its ability to reflect the contribution of media-based information in determining information goals. To understand this, consider the pages visited in session 6. In this session the user had started from the "*tools*" page and subsequently visited the *"famous places"* page and the *"galaxy"* page. The *"famous place"* page consisted of thumbnails of the galaxies and was linked to the *"galaxy"* page which itself contained images and information about different galaxies (snapshots of these pages are shown in Fig. 5). The top twenty information goals for this session as determined by IUNIS were: *famous, place, tooltitle, dr1, edge, navigate, dust, arm, elliptical, neighbor, filament, 907, chart, ned, jpeg, distant, two, image, extractor,* and *find*. It is important to note that terms such as names of the galaxies which are self-evidently important, given the prominence accorded to the images of the galaxies in these pages, were not reflected even among the top twenty information goals as determined by IUNIS. In contrast the top twenty information goals determined using the proposed method included, among others, names of all the galaxies in these pages. These goals were: *famous, place, ngc450, ngc60, tooltitle, dr1, ngc5792, ngc1032, ngc4437, ngc4753, ngc936, ngc5496, navigate, edge, chart, jpeg, dust, elliptical, arm,* and *neighbor.*

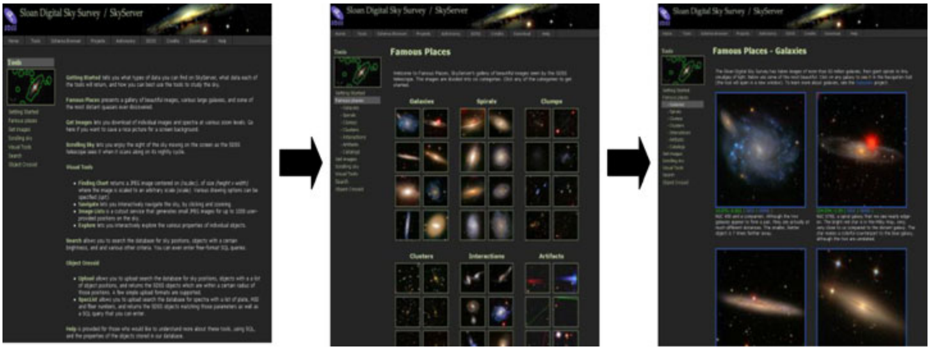


Fig. 5 Snapshots of the pages visited in session 6 from Fig. 3. The last two pages in this session are dominated by images of various galaxies

5 Conclusions and discussions

In this paper we investigated the problem of determining content pages in multimedia websites. The paper describes an approach which combines content modeling, with the use of information foraging theory and an information theoretic criteria to identify content pages. The method is generic and experiments indicate that it can provide insights on how users interact and assimilate information in multimedia web-pages. The complexity of determining the content pages is linear in the size of the session (number of pages visited). Moreover, the time consuming step of modeling the web site content and structure, can be done offline. Given the under-constrained nature of the problem and the complexity of indirectly assessing user intent as well as the semantics associated with media-based information, in our opinion, no single technique for determining content pages can be expected to perform well in all conditions. In the investigations presented in this paper, we have tried to address, one of the more complex and (relevant in real-world situations) formulations of this problem. We believe the proposed work will form a powerful conjunct to existing methods for content page determination and has the potential to constitute the basis for further advancements in this area.

Acknowledgments The authors thank the anonymous reviewers for comments that helped in improving the quality and presentation of the paper. The authors also thank Mike Wong for his participation in early parts of this research, results from which were published in [2]. 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.

References

1. Bertini M, Del Bimbo A, Nunziati W (2006) Video clip matching using MPEG-7 descriptors and edit distance. Conference on Image and Video Retrieval, LCNS 4071:133–142
2. Bhattarai B, Wong M, Singh R (2007) Discovering user information goals with semantic website media modeling, ACM International Conference on Multi-Media Modeling. Lect Notes Comput Sci 4351:364–375
3. Brusilovsky P (2001) Adaptive hypermedia, user modeling and user adapted interactions 11:87–110
4. Chi E, Pirolli PL, Chen K, Pitkow J (2001) Using information scent to model user information needs and actions on the web. ACM CHI 490–497
5. Craswell N, Hawking D, Robertson S (2001) Effective site finding using link anchor information. ACM SIGIR
6. Deng Y, Manjunath B (2001) Unsupervised segmentation of color-texture regions in images and video. IEEE Trans on Pattern Analysis and Machine Intelligence 23(8):800–810
7. Howarth P, Rüger S (2004) Evaluation of texture features for content-based image retrieval. LCNS 3115:326–334
8. http://htmlparser.sourceforge.net
9. Kang I, Kim G (2003) Query type classification for web-document retrieval. ACM SIGIR
10. Lee U, Liu Z, Cho J (2005) Automatic identification of user goals in web searclh. WWW
11. Olston C, Chi E (2003) Scenttrails: integrating browsing and searching on the web. ACM Trans Comput-Hum Interact 10(3):177–197
12. Pirolli P (2003) A theory of information scent. In: Jacko J, Stephanidis C (eds) Human-computer interaction, Vol. 1, (pp. 213–217), Lawrence Erlbaum, Mahwah
13. Pirolli P, Card S (1999) Information foraging. Psychol Rev 106(4):643–675
14. Qiu F, Cho J (2006) Automatic identification of user interest for personalized search. WWW 727–736
15. Rose DE, Levinson D (2004) Understanding user goals in web search. WWW
16. Salton G, Buckley C (1988) On the use of spreading activation methods in automatic information retrieval. ACM Conference on Information Retrieval, pp. 147–160
17. Salton G, Buckley C (1987) Term weighting approaches in automatic text retrieval. Technical Report: TR87-881
18. Santini S, Jain R (1997) Similarity is a geometer. Multimedia Tools and Applications 5:277–306
19. Santini S, Gupta A, Jain R (2001) Emergent semantics through interaction in image databases. IEEE Trans Knowl Data Eng 13(3)

20. Singh R, Bhattarai B (2009a) 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
21. Singh R, Bhattarai B (2009b) Analysis of usage patterns in large multimedia websites. In: Chbeir R, Hassanien A-E, Abraham A, Badr Y (eds) Emergent web intelligence. Springer Verlag (To Appear)
22. Singh V, Grey J, Thakar A, Szalay AS, Raddick J, Boroski B, Lebedeva S, Yanny B (2006) "SkyServer traffic report—the first five years", Microsoft Technical Report, MSR TR-2006-190, December
23. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380

Rahul Singh received the diploma in Computer Science, summa cum laude, from the Moscow Power Engineering Institute and his PhD from the University of Minnesota in 1999. He is currently associate professor in the department of Computer Science and the associate director of the Center for Computing in Life Sciences at San Francisco State University. His technical interests are in computational drug discovery, bioinformatics, and multimedia information modeling and management. Prior to joining academia, Dr. Singh was principal staff scientist at Scimagix Inc. Earlier, he founded and headed the computational drug discovery department at Exelixis Inc. At Scimagix, he was the co-designer of the ProteinMine^{TM} software which received the Frost & Sullivan Technology Innovation Award. Dr. Singh was a San Francisco State University Presidential Fellow (2006) and is a recipient of the CAREER award of the National Science Foundation.

Bibek D. Bhattarai received the MS degree in Computer Science from San Francisco State University specializing in multimedia information systems. He is currently at eBay Inc.