

Discovering User Information Goals with Semantic Website Media Modeling

Bibek Bhattacharai^{*}, Mike Wong^{*}, and Rahul Singh^{*}

San Francisco State University, San Francisco, CA 94132
{bdb,mikewong}@sfsu.edu, rsingh@cs.sfsu.edu

Abstract. In this work we present an approach to capture the total semantics in multimedia-multimodal web pages. Our research improves upon the state-of-the-art with two key features: (1) capturing the semantics of text and image-based media for static and dynamic web content; and (2) recognizing that information goals are defined by emergent user behavior and not statically declared by web design alone. Given a user session, the proposed method accurately predicts user information goals and presents them as a list of most relevant words and images. Conversely, given a set of information goals, the technique predicts possible user navigation patterns as network flow with a semantically-derived flow distribution. In the latter case, differences between predicted optimal and observed user navigation patterns highlight points of suboptimal website design. We compare this approach to other content-based techniques for modeling web-usage and demonstrate its effectiveness.

Keywords: web usability, multimedia semantics, information foraging.

1 Introduction

The mission of many educational and business websites is to deliver on-demand information. The usability of a website depends on how easily a user finds what they want—the user’s *information goals*. Stated another way, a website is highly usable if its *content pages*—pages which are likely to contain information goals—are readily accessible. Improving website usability thus depends on accurately identifying prominent information goals. By enumerating information goals for frequent user sessions, the web designer can refactor the website design, making content pages easily accessible to users. Furthermore, given a set of information goals and the website structure, one can simulate the *flow* (traffic) pattern of users attempting to satisfy the information goals. The user flow can then be visualized and/or compared with the shortest path that satisfies the goals, revealing potential points of sub-optimal website design.

Prior research [11] in *information foraging* theory asserts that users typically navigate towards their information goal by following *link cues*, which are fragments of information within or near hyperlinks and relevant to the user’s information goal. A cognitive theory based web-page usability inspection tool based on [11] is

^{*} All authors contributed equally.

presented in [1]. Another work [3] presents two algorithms: one discovers information goals from user sessions and the other predicts probable user flow, given a set of information goals. Both algorithms are based on *spreading activation* theory [2] which stipulates that cognitive processes, such as memory recall, are imprecise and non-isolated, triggering a cascading recollection of semantically related subjects. This theory is modeled using network flow methods [13], where the nodes of a network are primed by an activation weight, and neighboring nodes are recursively activated with iteratively decreasing weights. In the context of web usage, user flow is directly analogous to activation weight. Nodes which have a large number of incoming edges (fan-in) receive activation weight contributions from many sources, and therefore indicate important nodes. In this context, it should be noted that works such as [4] demonstrate that a better understanding of web usage can be obtained by combining web usage mining with web content and structure.

The state-of-the-art is unsatisfactory in two ways. *First*, most approaches rely heavily on textual information as the only significant source of semantics and hyperlinks as the only means of semantic relationships. Because of this limitation both non-textual links and dynamic content are disregarded in analyzing user flow and the contribution of image-based semantics is ignored. *Second*, important pages (content pages) are assumed to be predefined as a consequence of web design. This assumption causes the following two significant problems: (1) the contribution of user context and emergent or exploratory behavior on information goals is missed and (2) pages with a large fan-in unduly influence information goal discovery and user flow prediction. The latter problem which severely hampers current techniques, such as [13], essentially occurs due to the inherent assumption that web site design drives user behavior more than content or user context. It is interesting to note this contradiction from the fundamental assertions of information foraging theory.

This paper presents an approach to capture the total semantic contributions of multimedia web pages and for predicting user flow over text and non-textual links as well as other interface modalities, such as HTML form queries and script-driven interfaces. The proposed method incorporates a semantic representation which allows for a dynamic automated discovery of content pages. This differs from the network flow model for information goal discovery, and thereby avoids static website design dependencies. For user flow prediction, this method improves upon the network flow model by including a semantic cue factor which re-aligns network flow towards content-driven behavior. Visualization of predicted user navigation patterns can then indicate problems in web usability, by showing information goals in the context of the current web structure. The efficacy of the proposed technique in accurately identifying information goals and predicting user navigation habits illustrates that accounting for semantics across multiple media leads to a demonstrably better understanding of user behavior.

The remainder of the paper is organized as follows; Section 2 starts with an outline of the key problems involved in developing an understanding of website semantics. We then describe a model of multimedia websites used in our approach and discuss how the multimedia content of web-pages contributes to the relationships and attributes in this model. The section concludes with a discussion on how these relationships identify information goals for a user session and also how the model predicts user flow for a given set of information goals. Section 3 covers experimental

evaluations and is comprised of three different experiments: (1) a user study which evaluates the effectiveness of the proposed approach in capturing accurate information goals; (2) a goal relevancy comparison to the IUNIS algorithm proposed in [3] and another user study directly comparing the information goals predicted by both the proposed method and IUNIS, and (3) an evaluation of the benefits from multimedia/multimodal analysis over unimedia-unimodal analysis, which also demonstrates a typical use case for the system and the visualization of user flow. These evaluations are performed on the SkyServer [14] website, which is a large (~40TB) multimedia and multimodal site. The SkyServer site is designed to educate the public about astronomy and to serve the data needs of the astronomy community by offering direct access to all the data collected from the Apache Point Observatory in New Mexico.

2 The Proposed Approach

The essential tasks in finding user information goals for a particular user session are as follows: finding meaningful semantic relationships (how information goals are interconnected), representing total semantic information for all media (what to look for), and discovering content pages (where to look).

2.1 Modeling Hyperlinked Multimedia Semantics

We model web pages as containers for media objects, which are classified as *text*, *images*, and *links*. The model associates *semantic annotations*, which are modified term-frequency/inverse-document-frequency (TFIDF) vectors with media objects. Images are described by texture-color vectors, which are described in detail in section 2.4.

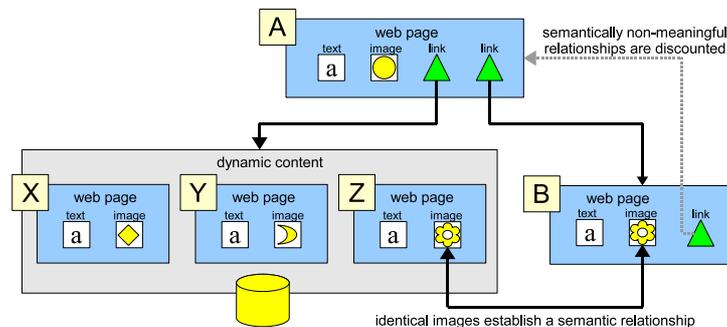


Fig. 1. A sample of the proposed semantic multimedia website model, showing meaningful semantic relationships and interoperability over dynamic and static content

As shown in Fig. 1, the links establish semantic relationships between the web pages. The figure also shows that parametric dynamic content may generate a number of views of the data, which manifest as web pages. For example, as shown in Fig. 1, web page *A* is the parent page for the web page *B* and any instance of the dynamic

content X , Y , or Z . If an image in B were to also appear in Z , then there would be an image semantics-based relationship between B and Z . Therefore the information from both pages B and Z would contribute to the semantics of this image.

2.2 Parent-Child Relationships as Strong Semantic Relationships

The first challenge in finding information goals is to determine how information is organized within a given website. Close relationships between pages in a user session give strong indications of what a user may be looking for. Many web sites have exceedingly interconnected pages; some links are semantically meaningful while others address usability issues. Common usability best practices include: offering a link to the home page on every page, providing a link so users can return to whence they last browsed, or showing a site map so a user can jump to any point in the overall organization of a subsection of the site. These links, in general, don't semantically relate the concepts of two web pages, and therefore need to be discounted when evaluating semantic relationships.

On the other hand, web sites are often organized from very broad and general introductions, down to specific information. For example, 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, which lists research interests, office hours, location, and contact information. These parent-child relationships are strong semantic relationships and therefore deserve special attention. Our approach identifies the website's parent-child relationships by traversing the links by breadth-first-search. If multiple parent pages link to the same child page, the child is clustered with the most semantically similar parent. This method preserves the semantic relationships between parent and child pages while discriminating links which exist for usability issues alone.

The rationale for finding strong semantic relationships is to capture total media semantics as the user would perceive it. To capture total media semantics given parent-child relationships between pages, the semantics associated with a parent page should include a fraction of each of the child page semantics. In our approach, the fraction of semantic back-annotation is proportional to the semantic similarity of the two pages. If the web page organization is such that there is no true parent-child relationship between two linked topics, then there should be no appreciable semantic similarity between the pages, and no back-annotation would occur.

2.3 Evaluating Multimedia Semantics

Once the strong semantic relationships are identified, the evaluation of multimedia semantics can begin. First we retrieve the web pages from the website. Dynamic content requests which culminate in an *HTTP GET request* are recorded in the usage logs. Therefore the results can be reconstructed and analyzed as if it were static content. Next, we decompose a webpage into its media components. Webpage structure is directly analyzed to separate navigational and ornamental motifs from the most conspicuous or "main" region of the web page. When applicable, the content of the "main" region is used for all semantic evaluations. The two most frequent media for most websites are text and images and therefore we focus on these two media in

our analysis. Textual content is analyzed using a grammarless statistical method, which includes stemming and stop word filtration. This simple method enables the analysis of very large websites, using limited computing resources, in a reasonable amount of time. In the course of this research, several analysis techniques have been empirically evaluated: term-frequency/inverse-document-frequency (TFIDF) [12] and latent semantic analysis (LSA) [8], as well as combinations of TFIDF and LSA. Of these, one variation of TFIDF is satisfactorily found to approximate the semantic content with the least computational expense. This version of TFIDF (which we call DTFIDF) uses a dynamic background document set. This background set is comprised of pages that meet three criteria: they are part of the website, they are semantically similar to the page of interest, and they *link to*, or *are linked from* the page of interest. This helps avoid the unwanted contribution of terms with overloaded semantics by using a smaller, relevant document set as a background set. The semantic annotation of a media object is represented by a DTFIDF-weighted term frequency vector.

To calculate the DTFIDF-weighted term frequency vector for a document (web page), let the document d be represented by a normalized term frequency vector (tf) which has a non-zero value for each term t in d . Let D be the set of all documents in the document collection (web site). Let N be the set of documents in D such that each document e in N has a similarity measure $r_{de} \geq k$ with respects to d and there exists a link between d and e . The value k is an empirically-determined constant threshold. The dynamically-calculated inverse-document frequency (idf) is as follows:

$$idf = \log \left(\frac{|N|}{|\{e \in N, t \subset e\}|} \right) \quad (1)$$

Our next goal is to characterize the information represented through images. Image semantics are challenging to ascertain because of the complexities of visual analysis and interpretation. One leading concept, in this context, is the use of image ontology [7], [10] to map semantics to images. The greatest challenge in developing an image ontology lies in finding ways to automatically annotate images. Fortunately, many websites include images for purposes of illustration, and augment the image with some text information. By taking advantage of webpage substructure, we can isolate proximal text which can then be used to annotate images.

There are two challenges associated with extracting image annotations from websites. The first is related to the emergent nature of image semantics; the same image may be used in multiple contexts and have multiple meanings. Second, the image may serve only layout or navigational purposes and not have any relevant semantic contribution to the pages on which it is located. To resolve the first challenge, the semantic annotations of identical images are summed, regardless of where they occur in the website. Images which aren't meaningful tend to be re-used often and for unrelated topic. In these cases, the entropy of the associated semantics of each image is measured. Images whose semantic annotation entropies exceed an empirically-determined threshold are cleared of any semantic annotation; these images are meaningless with regards to identifying user information goals.

2.4 Image Content Analysis and Semantic Associations in Image Clusters

Image content provides semantic value, and similar images may have semantic similarities as well. Therefore it is necessary to analyze the image content. We analyze images by first segmenting the images and then performing color and texture analysis on each segment, which results in a feature vector for each image segment. We use the JSEG [5] color/texture analysis system to identify textures within the image. An example showing the heart of a celestial phenomenon and its coronas is shown in Figure 3. Texture characterization is done with Grey-Level Co-occurrence Matrices (GLCM) [6]. We use eight vectors as the offset parameter for GLCM, and measure four statistical analyses for each co-occurrence matrix: energy, entropy, contrast, and homogeneity. In addition, we generate a low-resolution color histogram for each texture. Relative size, energy, entropy, contrast, homogeneity, and the color histogram are combined to create a feature vector to describe an image segment.

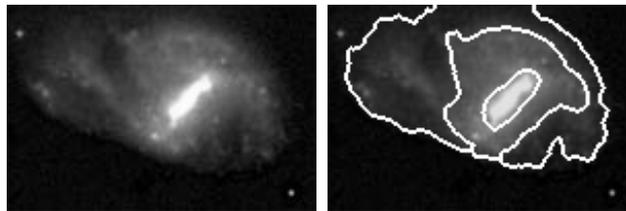


Fig. 2. Texture Segmentation on an image of a galaxy is the first step towards texture-color image feature extraction. Images are clustered by feature similarity. Commonly shared semantics in a cluster contribute more to information goal prediction.

Image similarity is measured through the normalized sum over segment similarity. Images are then clustered by similarity and each cluster is analyzed for frequently co-occurring terms in the semantic annotation of each image. Then, these terms are proportionally weighted. For example, the SkyServer website offers images of galaxies which share a high degree of similarity. These images often share the term *galaxy* in their semantic annotation, which is then given greater importance. Images of galaxy clusters are also similar to one-another, and are annotated with the terms *cluster* and *galaxy*, with *cluster* being more frequent than *galaxy*. Therefore, the term *cluster* is weighted more heavily than *galaxy* for the annotations which contain occurrences of *cluster*.

2.5 Dynamic Content Page Discovery and Information Goal Extraction

The current state-of-the-art takes the web designer's perspective that content pages are static as consequence of web design. We take a fundamentally different approach and assert that content pages are dynamic and relative to the user's information goals and the pages that the user had visited (in the session). Websites are generally organized so that users can drill down from pages with less specific information to pages with more specific information. A page with precise information (*e.g.* a page about a specific professor) is most likely to contain information goals. However, users

are free to traverse this hierarchy in many ways, foiling a designer's efforts to guide them. Measuring the semantic entropy of a given page gives the inverse of the specificity of the page. If a user follows a series of pages with monotonically decreasing entropy, then the web page at the local minima of entropy is likely to provide one or more information goals. These pages with low semantic entropy are identified as content pages. Every page in the user session may contribute towards a user's information goal prediction, but content pages are weighted to have a stronger contribution. Equation 3 shows the formula for the Shannon entropy of a semantic annotation x , being a term-frequency vector consisting of a set of term (t) and normalized frequency-count (c_t) pairs using DTFIDF.

$$H(x) = -\sum_{t \in x} c_t \log(c_t) \quad (2)$$

Information goals are extracted as a subset of the semantic information of web pages visited in a user session using both text and image information. The semantic contribution of each page is considered, with a greater weight placed on the semantic contribution of content pages. The term list is sorted first by page navigation order, and then by weight. Twenty most important terms are then used to form the basis of the predicted information goals. Images that are seen in the user session and have a semantic contribution towards the top 20 terms are also included, completing the predicted multimedia information goals.

2.6 User Flow Prediction and Implications on Usability

User navigation can be modeled as a network flow where the user may either leave the site or visit a different page of the website. Given a set of information goals, our model uses the semantic annotation to generate a probability distribution of user flow. The distribution asserts that links which are likely to lead to a page which satisfies a information goal are more probable than links which do not appear to satisfy a information goal. The criteria for a link which is likely to result in goal satisfaction are: (1) the link is associated with text or image semantics which is relevant to a goal; or (2) the linked page is semantically relevant to a goal.

$$p(l_0) = \frac{\sum_{t \in (S_{l_0} \cap G)} c_t}{2 \sum_{l \in L} \sum_{t \in (S_l \cap G)} c_t} + \frac{r_{GP_0}}{2 \sum_{l \in L} r_{GP_l}} \quad (3)$$

Users forage for their information goals by following information cues, which are bits of semantic information within (or near) a hyperlink. For a given set of information goals (G) and a given page with a set of links (L), each link (l) associated with semantic information (S_l), we can create a probability distribution which predicts the user flow to the linked page (P_l). The *information cue probability* ($p(l_0)$) for a particular link (l_0) is the probability a linked page will contain or lead to an information goal, is calculated as average of the link cue and the semantic cue. The link cue is the normalized sum of DTFIDF frequency counts of the terms that are

present in both the link semantics and the information goal. The semantic cue is given as the distance of the goal semantics and the linked page.

The user flow is computed by simulating users through an activation function $A(t)$ as shown in equation 5. The total percentage of users at a given time in a page depends on total information correlation value for all the links pointing to the page. The dampening factor α represents the probability of the user leaving the website from any given page. The matrix I represents the information cue matrix where the rows of the matrix represent a set of links from a web page. Each element of the row is calculated as the information cue probability described in equation 4. E simulates users flowing through the links from the entry (or start) page of the usage pattern. The initial activation vector $A(1) = E$. The final activation vector, $A(n)$, gives the percentage of users in each node of the website after n iterations.

$$A(t) = \alpha IA(t-1) + E \quad (4)$$

For each user session in a usage pattern, the algorithm computes the shortest path which covers all information goals in the predicted order. Our underlying assumption is that the shortest path represents the most optimal (direct) path to the desired information goal. User sessions which fit the predicted user flow are considered to have similar information goals, and are therefore fit for comparison. Comparing these sessions with the optimal shortest path provides an analysis of website design. If the user navigation patterns diverge from the optimal path, then there may be design problems at the point(s) of divergence.

3 Experiments and Results

The evaluation consists of three different experiments: (1) a user study which evaluates the effectiveness of the model in capturing accurate information goals, (2) a goal relevancy comparison to the IUNIS algorithm [3] and another study directly comparing the information goals predicted by both the proposed system and IUNIS, and (3) an evaluation of the benefits from multimedia-multimodal analysis over unimedia-unimodal analysis. Included here is a typical use case for the system and the visualization of user flow. These evaluations are performed on the SkyServer website [14] website, which is a large (~40TB) multimedia and multimodal website. The SkyServer website is designed to educate the public about astronomy and to serve the data needs of the astronomy community by offering direct access to all the data collected from the Apache Point Observatory in New Mexico.

3.1 Evaluation of User Information Goal Discovery

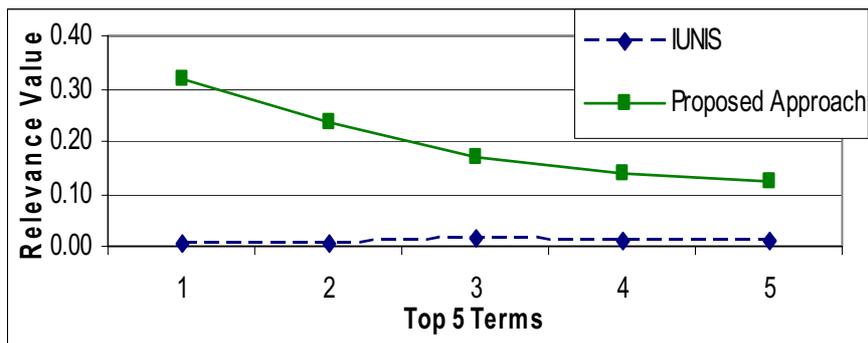
To evaluate the correlation between predicted information goals and users' information goal, a user study was conducted with 8 participants, none of whom were familiar with our research. Ten frequently recurring user sessions were selected from the usage logs and analyzed using the proposed approach, generating ten sets of predicted multimedia information goals. Four sessions were selected from the ten, and each of the web pages for these sessions were printed and bound in chronological order, producing four booklets. Each user was asked to read three of these web

session booklets, one-at-a-time, and for each booklet the user was asked to compare the information they just read to the 10 sets of predicted information goals. Each user was then asked to 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. Therefore, we conclude that it is likely that there is a strong correlation between the predicted goals and the users' information goals. 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. We tested for systemic bias in the study by measuring mean pair-wise Pearson's correlation of answers for each booklet and found a low correlation between answers ($\bar{r}^2 = 0.233, \bar{r} = 0.483, s = 0.280$), indicating no significant systemic bias. Users scored the proposed approach's set of goals high, ($\bar{x} = 4.38, s = 0.824$, out of 5) compared to all other sets of goals ($\bar{x} = 2.56, s = 0.160$), indicating that the proposed approach predicts distinguishably relevant information goals for a given user session.

We next compared goal analysis from IUNIS [3] with the proposed approach over the same ten user sessions from the user study. For each user session, the top five terms of the information goals predicted by IUNIS were compared with the top five terms of the information goals predicted by the proposed approach. As seen in table 1, the top 5 terms score up to 47.6 times more strongly by the proposed approach than by IUNIS; the mean increase in term relevancy is 24.6 times ($\bar{x} = 24.58, s = 18.23$).

Table 1. Comparison of the Relevance Scores for the Top 5 Goal Terms for IUNIS vs. the Proposed Approach Reveals a Mean 24.6 Times Improvement



3.2 Evaluation of Multimedia and Multimodal Analysis

This experiment evaluated the effectiveness of text-and-image analysis versus text-only analysis. The evaluation focused on nine user sessions where graphics provide a

significant contribution to web content. Figure 3 shows the first user session; the other user sessions were identical except for the last page, which correspondingly were *page1.asp* to *page6.asp*, *ngc.asp*, *abell.asp*, and *messier.asp*, respectively. In these sessions, the user visited pages that have thumbnail images of galaxies, spirals, clusters, and so on with the last page displaying specific annotated image examples of these celestial objects.

```

Base URL: http://skyserver.sdss.org/dr1/en/tools
1: default.asp
2: places/default.asp
3: places/page1.asp

```

Fig. 3. User Session with Text and Image Media

The text-only model was found to undervalue image semantics leading to the names of the featured galaxies not appearing in the information goal, despite the prominent captions. In contrast, the multimedia model captured image relevance, ranking the galaxy names in the top 20 information goals, while preserving the content and rankings of the top 5 terms as discovered in text-only analysis. Moreover, due to the semantic association of similar images, the top terms showed an improved relevance score, skewing the relevance curve towards the most important terms. The mean relevance improvement was found to be 4.1 times ($\bar{x} = 4.13, s = 1.49$).

The final evaluation is meant to demonstrate the ability of the proposed approach to predict user flow through multiple modalities, specifically hyperlink browsing and script-enabled dynamic content (a simulation of a virtual telescope, allowing the user to view different areas of the night sky). The session shown in Figure 4 considers a typical use case: the user first browses through the static pages and then interacts with the script-enabled dynamic content in the third page of the session. The proposed approach captures the information goals for all pages in the user session, regardless of the interaction mode. As shown in Figure 4, the proposed approach predicts a non-zero probability for user flow towards the dynamic content page, which is a capability not available to the current state-of-the-art approaches.

```

Base URL: http://skyserver.sdss.org/dr1/en/
1: default.asp
2: /tools/places/default.asp
3: /tools/places/page2.asp
4: /tools/explore/obj.asp?ra=221.546&dec=-0.223

```

Fig. 4. User Session with dynamic content access

In Figure 5, the lines represent hypertext links and the nodes represent web pages. The figure shows the user flow (orange solid line), the user session path (green dotted line), the shortest path (blue broken line) and the red bar representing the user flow probability computed by the system. Selecting a node shows a thumbnail image of the page and the URL.

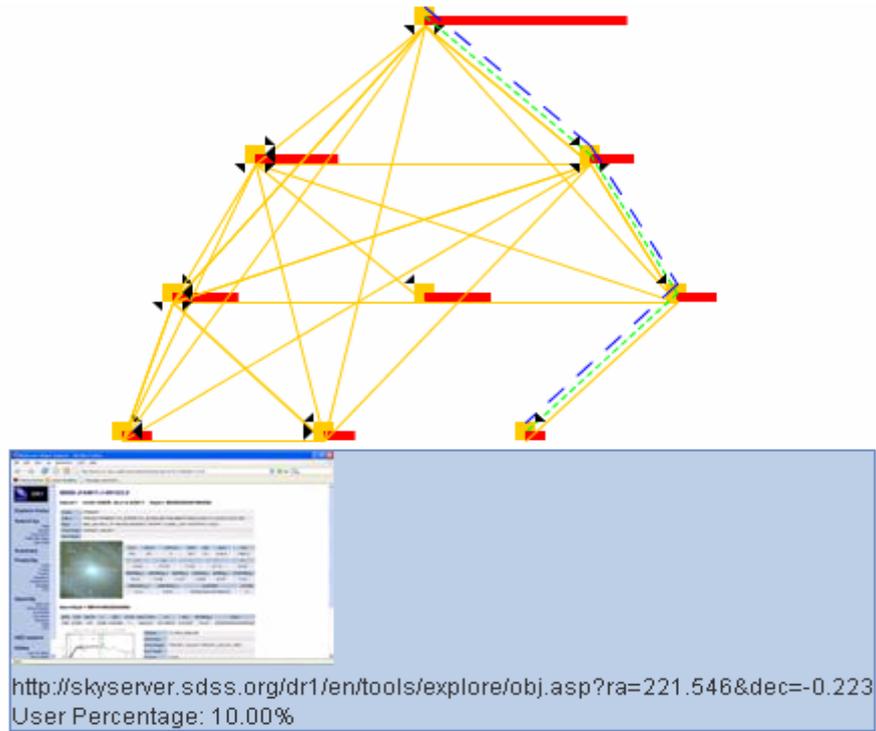


Fig. 5. User flow diagram shows the given user session with dynamic page content access

4 Conclusion

In this paper we presented a new approach for capturing total semantics in multimedia web pages and demonstrated how the method can be used to identify user information goals and predict user flow using a network flow methods. Two key contributions of our research are: (1) an algorithm for dynamic content page identification and assimilation for information goal discovery and (2) a semantically determined (over multimedia content) probability distribution which works over multimodal interfaces for user flow prediction. User study evaluation of this model shows that the model predicts accurate information goals for a given user session with a mean of 24.6 times greater relevance than the current state-of-the-art. This remarkable improvement is due to a fundamentally different approach to localizing semantics and dynamically aligning web page semantics and user navigation. Experimental studies underline the significant improvements brought about by capturing semantics over multimedia content. Finally, we presented a mechanism to predict user flow over rich multimodal interfaces for website dynamic content. In conclusion, this research provided compelling and exciting evidence towards the usefulness of characterizing multimedia-multimodal semantics in context of analysis of web usability.

References

- [1] M. H. Blackmon, P. G. Polson, M. Kitajima, C. Lewis, Cognitive Walkthrough for the Web. ACM Proceedings of Conference on Human Factors in Computing Systems (CHI '02), 2002.
- [2] J. R. Anderson, P. L. Pirolli, Spread of Activation, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 10, pp. 791-798, 1998
- [3] E. Chi, P. L. Pirolli, K. Chen, J. Pitkow, Using Information Scent to Model User Information Needs and Actions on the Web. ACM Proceedings of Conference on Human Factors in Computing Systems (CHI '01)
- [4] R. Cooley, The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns. 2003 ACM Transactions on Internet Technology 3(2), pp 93-116, 2003.
- [5] Y. Deng, and B. S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI '01)*, vol. 23, no. 8, pp. 800-810, 2001
- [6] P. Howarth, S. Rüger, Evaluation of Texture Features for Content-based Image Retrieval, *Lecture Notes in Computer Science*, Volume 3115, pp. 326 – 334, 2004
- [7] E. Hyvönen, A. Styrman, S. Saarela, Ontology-Based Image Retrieval. Towards the semantic web and web services, *Proceedings of XML Finland 2002 Conference*, pp. 15-27, Helsinki, Finland, October 21-22, 2002
- [8] T. K. Landauer, P. W. Foltz, and D. Laham, An Introduction to Latent Semantic Analysis, *Discourse Processes*, Vol. 25, pp. 259-284, 1998.
- [9] B. S. Manjunath, and W. Y. Ma, Texture Features for Browsing and Retrieval of Image Data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, 1996
- [10] V. Mezaris, I. Kompatsiaris, M.G. Strintzis, An ontology approach to object-based image retrieval, *International Conference on Image Processing* vol. 3, no. 2, pp. II- 511-14, 2003
- [11] P.L. Pirolli, and S. K. Card, Information foraging. *Psychological Review*. 106: p. 643-675, 1999
- [12] G. Salton, and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, *Technical Report: TR87-881*, 1987.
- [13] G. Salton, and C. Buckley, On the Use of Spreading Activation Methods in Automatic Information Retrieval, *ACM Conference on Research and Development in Information Retrieval*, pp. 147-160, 1988
- [14] Sloan Digital Sky Survey project's website SkyServer: <http://skyserver.sdss.org/dr1/en> (English version for Data Release 1)