Information-Theoretic Identification of Content Pages for Analyzing User Information Needs and Actions on the Multimedia Web

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

ABSTRACT

Increasing use of media in web pages is fueling the metamorphosis of the WWW into a multimedia web. The development of information organization, analysis, and search technologies that enable people to efficiently find the specific information they require becomes especially important in this context. Determining the *information goal* of a user is a critical step in this direction. However, the information goal is typically subjective and latent. Its determination is further complicated in multimedia websites since the semantics of media-based information is context-based and emergent. Furthermore, interaction modalities with media are, unlike static link-based browsing, complex to analyze. In this paper we address the problem of automatically determining the *content pages* given the browsing behavior of a user. The content pages contain the information the user came to the site to find. Thus, their identification is a critical step in reasoning about user information goals. We propose an information theoretic approach that takes into account the organization of the web site, the multimedia information content, as well as the influence of a specific browsing pattern to identify one or more pages that putatively contain the information goal(s). This method can be used irrespective of whether the user has a single information goal or is looking to satisfy multiple information needs. Experimental investigations on media rich sites illustrate the efficacy of the technique and its potential in modeling user information needs and actions in a multimedia web.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]; H.5.2 [User Interfaces]; H.5.4 [Hypertext/Hypermedia]; H.1.1 [Systems and Information Theory]

Keywords

Information goal, WWW, content pages, information theory.

1. INTRODUCTION

The success of a website depends critically on users finding the information they seek (i.e. satisfying their *information goal*). At the state-of-the-art, research in determining user information goals has occurred along two directions, corresponding to how users locate information [12]; either by searching (so called “search by query”) or by browsing (“search by navigation”). Estimating the user information goal is important for both these settings. For instance, knowledge of the information goal can be used for improving page ranking [4, 8] and result presentation [9]. It can also be used to evaluate/improve site usability and predict usage patterns [2, 3].

In context of search by querying, techniques have been developed that use features such as the occurrence patterns of query terms in the web-pages [8] or click-behavior and anchor-link distributions to classify queries (and the underlying information goals) as *navigational* or *informational*. The former class contains queries where the user intent is to visit a specific page while the in the latter, the user seeks to learn about a topic and can visit multiple pages. The history of user click-behavior has also been employed in conjunction with the concept of topic-sensitive page rank to estimate user sensitive information goals for personalized search (see [11] and references therein). *Our research relates to the “search by navigation” scenario where users locate information by browsing.* The intuition, derived from information foraging theory [10], is that in a web site, the user makes traversal decisions looking for information that would satisfy his or her information goal. The traversal history (which can be obtained through the web-log) is therefore representative of the user information need. Intuitively, certain pages visited during the traversal will have greater relevance towards the ultimate information goal than others. Thus, given a sequence of pages visited by a user, the problem is to identify the putative *content pages* – pages which contained information satisfying the information goal of the user. A website has a high usability factor if its content pages are readily accessible. Improving website usability thus depends on accurately identifying content pages. This information can be used, for instance, to re-factor the site design and make the content pages easily accessible. It can also be used to compare expected user paths with actual usage patterns and identify potentially problematic aspects of the site design.

Determining the content pages is fundamentally an under-constrained problem since in real-world setting the user intent is not known *a priori*. At the state-of-the-art, various ideas have been proposed to determine content pages. These include: (1) predefining the content pages for a site, (2) treating the last page(s) of a session as the content page(s), and (3) page access-based weighting. These methods have significant limitations. For instance, the simple and often used idea of predefining the content pages [3, 9, 13] assumes that such a distinction is possible as a consequence of web-design alone. However, recent research [15]

SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03…$5.00.

shows that the semantics associated with complex information, such as multimedia, is non-unique, user-dependent, and emergent. Therefore, in information and media rich web-sites, both data context and user context play a critical role in user behavior and need accounting. Consequently, content pages can not always be predefined. Similarly, treating the last page(s) of a session as the content page(s) is problematic since the user may have multiple information goals or may simply be lost. The idea of page access-based weighting [3] seeks to use access frequency in a manner mirroring the idea of TFIDF weighting strategy from text analysis [13] 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.

Based on the aforementioned analysis, we can identify the key characteristics which a method for content page identification needs to have. These include:

- *Data and user-context sensitivity*: Content pages should be identified by analyzing the interplay between site content *and* user behavior rather than being driven by site-design and content alone.
- *Ability to account for non-textual information*: Web-sites are increasingly employing non-textual media to convey information. As discussed earlier, media introduces important challenges including polysemy (e.g. same image used in different settings), use in settings where subtle differences in appearance have significance, and most importantly due to the emergent nature of its semantics. Careful modeling of the overall semantics of web-pages including both text-based and media-based information therefore needs to underlie content page determination.
- *Generality*: The method should be applicable to a large class of web-sites and usage patterns and use fundamental criteria to identify content pages rather than using simplistic criteria (such as using the last page in the session as content page) that work only for certain sites.

In this paper we present an algorithmic approach for content page determination that fulfils all the above criteria. The intuition behind the approach lies in co-analyzing usage patterns (sequence of pages visited) and site content to identify pages that contain the most specific information given the browsing context of the user. To account for site content and structure, the method models the contribution of both text-based and media-based information towards the overall semantics of the site. The effect of user behaviour and context is taken into account by considering the browsing history of the user within the site. The central idea of the approach lies in analyzing the sequence of pages visited by the user and determining how the specificity of the information provided across the pages varies. The information specificity of a page is determined by computing the Shannon entropy of its content. Consequently pages corresponding to local information-theoretic minima contain the most specific information given the user browsing context. Intuitively, they are therefore most likely to satisfy the user information goals. This intuition, as our investigations indicate, is strongly supported in case studies and experiments.

The rest of the paper is organized as follows: in Section 2 we begin by identifying the primary factors that influence whether a page can become a 'content page'. Subsequently the modelling of multimedia web-pages is discussed to account for these factors. Based on these models, the strategy for automatic identification of content pages is presented in Section 3. The effectiveness of the method is analyzed through case studies and experiments in Section 4. Finally in Section 5 discuss the applicability of the method in multimedia web design and web structure evaluation.

2. MODELING THE SEMANTICS OF MULTIMEDIA WEB PAGES

Determining the content pages requires studying the interplay between three factors: (1) the web-site content, (2) the site structure, and (3) the (latent) user context and information need. In the following we discuss the first two factors in detail. The third factor, namely the user context and information goal, drives the choice of visiting specific pages and is hence reflected (possibly partially) by the sequence of pages visited in the session.

2.1 Modeling Site Content

The website content is described through text and media. We assume the media to be image-based as this is the most commonly encountered case. Other forms of media can 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 part of this step, web pages are analyzed with a Java-based HTML parser [7] and navigational and ornamental motifs separated from the primary content. In case of the web pages with unclear HTML structure, the entire body section is used as page content.

Textual content is analyzed using a grammarless statistical method, which includes stemming and stop word filtration. Subsequently, a variation of the TFIDF method is used to approximate the text-based semantics. This version, which we call DTFIDF, uses a dynamic background document set. If a web-page *d* is represented by a normalized term frequency vector and *D* is the set of all documents in the web site, then the DTFIDF value for each term 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 *d*, containing t_{total} terms and *N* denotes the set of pages in *D* such that each page *e* in *N* has a similarity measure $r_{de} \geq k$ with respects to *d* and there exists a link between *d* and *e*. Next, each page in the web site is represented by a term vector T_p containing terms having high DTFIDF values.

In the case of web-pages the problem of determining the semantics of images can be ameliorated by associating with an image, the proximal text, and thereby estimating the semantics associated with the image. This presents three technical challenges: (1) Assigning meaningful text to images (2) Dealing with images that are used in multiple contexts with possibly 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. Our approach to solving the last two problems is based on determining the similarity of images in a website. For images determined to be perceptually identical, key-terms assigned to the images are combined, thus capturing the possibly variable semantics associated with these images. Images serving navigational or layout purposes tend to be re-used often and for unrelated topics. Consequently, the variability (defined as the entropy) of the associated key-terms of such images is large. This criterion is used to exclude such images from consideration.

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

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

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

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

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

In Addition to the above descriptors, a low-resolution color histogram is generated. Relative size, energy, entropy, contrast, homogeneity, and the color histogram are combined to create a feature vector to describe an image. The similarity of two images is computed as the Pearson's distance between their respective feature vectors and key terms are combined for identical images. A score of 1.0 indicates identical images (for which corresponding key terms are combined) and low scores indicate highly dissimilar images. This approach for comparing image similarity is computationally simple and empirical results indicate it to be effective in identifying highly similar images and differentiating between non-similar images (see Figure 1 for representative examples).

Next, a unified semantics representation of the entire information in the website is obtained by re-weighting terms that co-occur in the term-frequency matrix and image annotations to ensure that the effect of image size and complexity is reflected in the term weight. Specifically, if a term *t* with frequency value of *f* appears in a page *P* and in the annotation related to the image *I* of size I_x pixels and texture count T_c, then its new frequency is calculated as shown in Eq. (6):

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

2.2 Modeling Site Structure

Web-site structure also influences the semantics of the information being conveyed as related pages are typically linked and provide users with *information scent* [10] through proximal or distal information cues [3, 9]. Browsing patterns can be explained using these cues as demonstrated by the framework of information foraging theory [10]. Moreover, the linkage structure often represents an organization, for instance, a transition from general to specific information. We identify these 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 and the semantics of the parent page is updated by back-annotation from the child page semantics. Specifically, if a term *t* has importance value of t_c in child page and t_p in parent page and semantic distance between the child page and the parent page is d_{cp}, then weighted importance value w_p of the term *t* in parent page is re-calculated as shown in Eq. (7). The rationale for adjusting the importance of terms within a page based on the connectivity of the page 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, as described in Eq. (7), 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 would be no appreciable semantic similarity between the pages, and no back-annotation would occur.

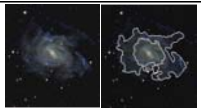
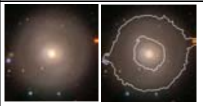
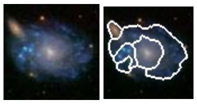
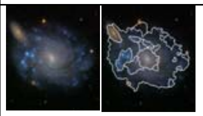
Image-1	Image-2	Similarity scores
		1.0
		0.74
		0.95

Figure 1. Examples of image matching. The similarity scores are shown in the last column.

$$w_p = t_p + d_{cp} \times t_c \qquad (7)$$

3. DETERMINING THE CONTENT PAGES

Given a web page *x*, represented by its semantically unified representation vector, T= [t_1, . . .t_n], we define its page entropy as:

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

Where *p(t_i)* is probability of term t_i occurring 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. .

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

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 website structure.

The content pages are subsequently determined as follows. Let the browsing pattern of a user in a particular session by 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 range and k are predefined (we use 5% and 2 respectively in all our experiments). This avoids cluttering of content pages. If two minima are k-neighbors, the page with the lowest entropy value is selected as the content page.

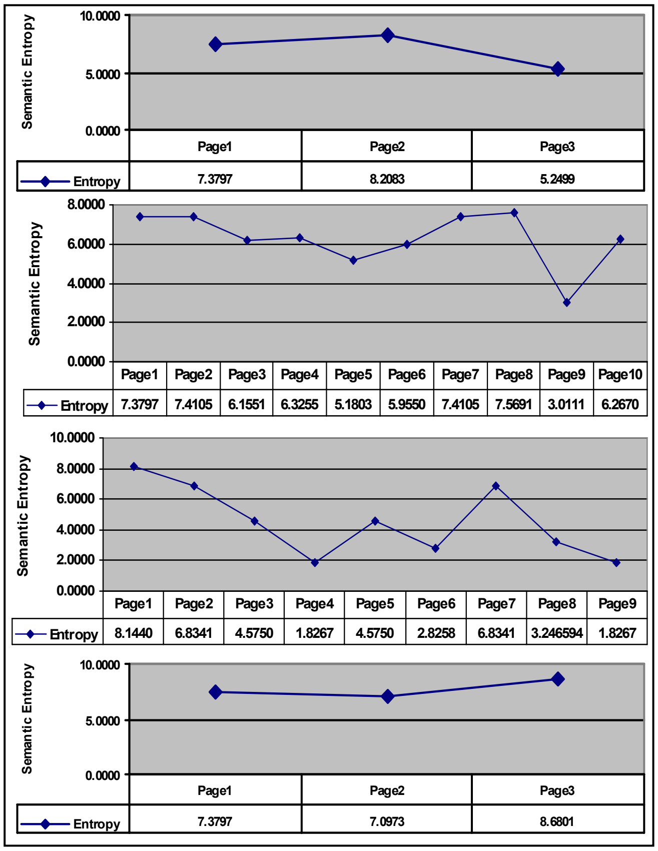


Figure 2. Case study investigating the proposed method on four different user sessions where the content pages were known *a priori*. The graphs plot the variation in page entropy for each session. Sessions 1, 2 and 4 were on the Skyserver (http://skyserver.sdss.org/) – a multimedia astronomy website. Session 3 was on the website of the author's research group.

4. EXPERIMENTS AND CASE STUDIES

We begin with a case study that investigates four different user sessions where the user information goals were known. Thus the actual content pages could be compared with those obtained using the proposed method (see Figure 2). 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 top graph. The second session captures the behavior of 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) are correctly identified as content pages by the method. Session 3 was also designed to have multiple information goals and uses 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 starts at the index page of the research website then visits the group project directory page. Next the session visits 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 session backtracks to group project list page followed by a visit to the Cross-modal information retrieval project page and finally the home page of the member student (page 9). The content page discovery technique correctly identifies Page-4, Page-6, and Page-9 as the content pages. The final session exemplifies a case where the proposed content discovery technique fails. 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 contains information about various topics and therefore has high semantic entropy. Therefore, our technique incorrectly identifies the second page as the content page. We note however, that based on our experience, such degenerate cases occur rarely.

In the first experiment the efficacy of the method was investigated (in terms of precision and recall) with multiple users and websites. Five websites were used in the study: (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). 20 users participated in the study and explored each of the five websites with three different information goals, yielding 300 sessions. All the users were students at San Francisco State University and drawn from different departments. However, none were familiar with our research. At the end of each session, users reported the page(s) they found most relevant (the content pages). In 160 sessions, the users identified a single content page (henceforth referred to as Type-I sessions), 121 sessions had multiple content pages (Type-II sessions), in 6 sessions users failed to clearly identify content pages and considered all the pages to be informative (Type-III sessions) and in 13 sessions users got lost and left the site. The precision (recall) values for the proposed method were: Type-I: 83.1% (81.87%), Type-II: 71.91% (78.51%), Type-III: 100% (30.77%).

In the second experiment we applied the proposed approach for identifying user information goals. We compared the results with the INUIS (Inferring User Need by Information Scent) algorithm [3], which is one of the established methods in this area. The fundamental distinction of our approach from INUIS lies in multimedia content modeling and in giving greater weight to information in content pages (which were defined using the proposed method). Ten user sessions were analyzed from the Skyserver logs. The relevance score of a term in a session was defined as its maximum DTFIDF value across pages visited in that session. To provide a synopsis, the average relevance scores were calculated and the top 10 information goals (terms) identified across all sessions. As shown in Figure 3, the top 10 goals scored up to 47.6 times more strongly by the proposed approach than by IUNIS and the mean increase in term relevancy was 15.06. Moreover, terms associated with images found greater relevance. Thus, by directly modeling the influence of media-based information, a different and arguably more complete understanding of information scent and user behavior was obtained.

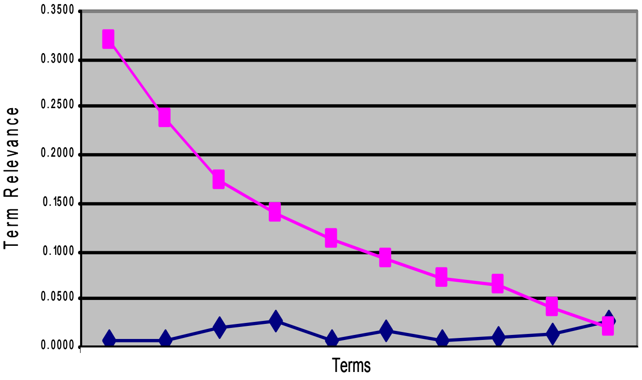


Figure 3. Comparison of the relevancy of the top ten user goals determined using the proposed approach (top curve) and IUNIS (bottom curve) across ten user sessions on the SkyServer

5. DISCUSSIONS AND CONCLUSIONS

This paper investigated the problem of determining content pages in multimedia web-sites. It proposes an information-theoretic approach, where the putative content pages are identified as the pages corresponding to the local minima of page-entropy values. This step is preceded by analysis and modeling of the interplay of user context, media-based website content and website structure. 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 web-site modeling step can be done offline.

Given the under-constrained nature of the problem and the complexity of indirectly assessing user intent, in our opinion, no single technique for determining content pages can be expected to perform well in all conditions. However, the proposed method, as our investigations indicate, provides a highly effective automated approach to content page estimation. We consequently believe that the proposed approach forms a powerful conjunct to existing methods for content page determination. Furthermore, as demonstrated in our experiments the method can be used to improve the performance of web-usage analysis techniques and has the potential to play an important role in the development of techniques for information organization, analysis, and search technologies for the multimedia web. Future directions of our research include investigating extensions to the proposed method by incorporating additional information such as temporal characteristics of the user path, including time spent at a page, to improve content page detection.

ACKNOWLEDGEMENTS

The authors thank Mike Wong for his participation in parts of this project. This work was funded in part by a Microsoft unrestricted research grant to RS.

REFERENCES

[1] Bertini M., Del Bimbo A., and Nunziati W., “Video Clip Matching Using MPEG-7 Descriptors and Edit Distance”, Conference on Image and Video Retrieval, LCNS 4071, pp. 133-142, 2006

[2] Bhattarai B., Wong M., and Singh R., "Discovering User Information Goals with Semantic Website Media Modeling", ACM International Conference on Multi-Media Modeling, Lecture Notes in Computer Science, Vol. 4351, pp. 364 – 375, 2007

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

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

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

[6] Howarth P, Rüger S., Evaluation of Texture Features for Content-based Image Retrieval, LCNS, Vol. 3115, pp. 326 – 334, 2004

[7] http://htmlparser.sourceforge.net

[8] Kang I and Kim G, “Query Type Classification for Web-Document Retrieval”, ACM SIGIR 2003

[9] Olston C and Chi E, “ScentTrails: Integrating Browsing and Searching on the Web”, ACM Trans on Human-Computer Interactions, pp. 177 – 197, Vol. 10, No. 3, 2003

[10] Pirolli P and Card S, “Information Foraging”, Psychological Review, 106 (4), pp 643-675, 1999

[11] Qiu F, Cho J "Automatic Identification of User Interest for Personalized Search." WWW, pp. 727-736, 2006.

[12] Rose DE and Levinson D, “Understanding User Goals in Web search”, WWW, 2004

[13] Salton G, and Buckley C, “On the Use of Spreading Activation Methods in Automatic Information Retrieval”, ACM Conference on Information Retrieval, pp. 147-160, 1988

[14] Salton, G., and Buckley, C., “Term Weighting Approaches in Automatic Text Retrieval,” Technical Report: TR87-881, 1987.

[15] Santini S., Gupta A, and Jain R, “Emergent Semantics Through Interaction in Image Databases”, IEEE Trans. On Knowledge and Data Engineering, Vol. 13, No. 3, 2001