Emergent Semantics and the Multimedia Semantic Web[i]

W.I. Grosky*, D.V. Sreenath**, and F. Fotouhi**
*Department of Computer and Information Science, University of Michigan-Dearborn
Dearborn, Michigan 48128
**Department of Computer Science, Wayne State University, Detroit, Michigan 48202
wgrosky@umich.edu, {sdv, fotouhi}@cs.wayne.edu

ABSTRACT

It is well known that context plays an important role in the meaning of a work of art. This paper addresses the dynamic context of a collection of linked multimedia documents, of which the web is a perfect example. Contextual document semantics emerge through identification of various users' browsing paths though this multimedia collection. In this paper, we present techniques that use multimedia information as part of this determination. Some implications of our approach are that the author of a webpage cannot completely define that document's semantics and that semantics emerge through use.

1 INTRODUCTION

Information is increasingly becoming ubiquitous and all pervasive, with the world-wide web as its primary repository. The rapid growth of information on the web creates new challenges for information retrieval. The goal of our research in the last few years has been to bridge the semantic gap between the ways in which users request web pages (or resources, in general) and those users' real needs, and ultimately, to improve the quality of web information retrieval.

The development of feature-based techniques for the retrieval of multimedia information has emphasized the notion of similarity with respect to low-level features. In our view, researchers in content-based retrieval should now concentrate on extracting semantics from multimedia documents so that retrievals using concept-based queries can be tailored to individual users. Following the semantic web paradigm, techniques for the semi-automatic annotation of multimedia information should be developed. Following Berners-Lee et. al. [1], a typical example would be a user request to plan a vacation where the painting shown in Figure 1 is being exhibited, where the user knows neither its title nor who painted it, or where paintings of the same style are being exhibited.

Existing management systems for multimedia document collections and their users typically are at cross-purposes. While these systems normally retrieve multimedia documents based on low-level features, users usually have a more abstract notion of what will satisfy them. Using low-level features to correspond to high-level abstractions is one aspect of the *semantic gap* Gudivada and Raghavan [3] between content-based system organization and the concept-based user. Sometimes, the user has in mind a concept so abstract that he himself doesn't know what he wants until he sees it. At that point, he may want to access multimedia documents similar to what he has just seen or can envision. Again, however, the notion of similarity is typically based on high-level abstractions, such as events taking place in the document or evoked emotions. Standard definitions of similarity using low-level features generally will not produce good results.

Figure 1 – A Particular Painting

In reality, the correspondence between user-based semantic concepts and system-based low-level features is many-to-many. That is, the same semantic concept will usually be associated with different sets of features. Also, there can be many different multimedia documents having similar features, which satisfy different needs for different users, such as when their relevance depends directly on an evoked emotion.

Multimedia annotations should be *semantically-rich,* and exhibit the multiple semantics mentioned above. It is our belief that these multiple semantics can be discovered by the way multimedia information is used. This can be accomplished by placing multimedia information in a natural context-rich environment, for it is by context that multiple semantics emerges.

There are two important kinds of context: *static context* and *dynamic context.* The author of the multimedia document, who places semantically similar information in physical proximity to each other, defines the static context, which we may also call *structural context*.

Dynamic context is the user's contribution. Semantics emerge through identification of various users' browsing paths through a linked multimedia document collection. This follows from the fact that over short browsing paths, an individual user's wants and needs are uniform. Thus, the various sub-documents visited over this path exhibit uniform semantics in congruence with these wants and needs. This illustrates the concept of *semantic coherence.*

The web happens to be a perfect example of such a context-rich environment.

Based on the above, we are convinced that an important problem is to semi-automatically develop web page annotations based on cross-modal techniques for text, images, video, and audio information.

The semantics of a web page is defined by its content and context. Understanding of textual documents is beyond the capability of today's artificial intelligence techniques, and the many multimedia features of a web page make the extraction and representation of its semantics even more difficult. Modern search engines rely on keyword matching and link structure, but the semantic gap is still not bridged. Previous studies have shown that users' surfing on the web exhibit coherent intentions (or browsing semantics) and that these intentions can be learned and used for the prefetching of related web pages Ibrahim and Xu [4]. In our approach, the semantics of a web page can be derived statistically through analyzing the browsing paths of users toward this page. For this reason, we also refer to these emergent semantics of a page as *dynamic semantics*.

We use the technique of latent semantic analysis Deerwester et. al. [2] to determine the semantics of web pages, derived from their textual features, image features, and structural features. We have previously shown that latent semantic analysis can discover that certain sets of different image features co-occur with the same textual annotation keywords Zhao and Grosky [7]. Using this technique, we represent each document in a reduced dimensional space, where each dimension corresponds to a concept, each concept representing a set of co-occurring keywords and image features. Resulting keyword searches for web pages become much more efficient after this transformation; more efficient than if we used latent semantic analysis on just the keywords. Building on this insight, an important aspect of our current work is to bring multimedia information into the definition of web-page semantics.

We envision the following scenario. A web page does not have a fixed semantics, but multiple semantics that vary over time. Each element of a webpage's multiple semantics corresponds to a group of users who visit this particular page through similar browsing paths, and is defined in terms of the contents (both textual and visual) of these pages, the structural layout of these pages, and the amount of time spent browsing these pages. As different users visit a given page, its semantics changes. Similarly, as users browse the web, they incrementally build up semantic profiles. By comparing a user's browsing semantics with the semantics of pages on the web, we can make intelligent suggestions as to what web pages the user should further examine. Similarly, a user can use our system to query for page suggestions. However, instead of giving only textual keywords as input, he can also use visual and structural cues. Also, users can issue query-by-examples; that is, they can give the search engine examples of acceptable web pages.

The main motivation for this work is our belief that short sub-paths of a user's browsing path through the web exhibit uniform semantics, and that these semantics can be captured, easily represented, and used to our advantage. Paths that are more heavily used will be given higher weight in determining the appropriate semantics. This is a social theory of semantics, akin to collaborative filtering Shardanand and Maes [6], where the web pages similar to a given web page, *w,* are the pages that are liked by other users who also liked *w*.

This paper is a preliminary sketch of some ideas we are currently implementing in this area. In the next section, we explain our general approach.

2 OUR APPROACH

Our belief is that a user's browsing path through the web exhibits what we call *semantic coherence.* That is, while the entire user's path exhibits multiple semantics, especially pages far from each other on the path, neighboring pages, especially the portions close to the links taken, are semantically close to each other. Our tasks are to characterize the

contiguous sub-paths of a user's browsing path that exhibit similar semantics and detect the *semantic break points* along a user's browsing path where the semantics change appreciably, as well as to categorize the semantics of each web page based on the totality of users' browsing paths.

We now formalize this concept. As is usually done, we treat the web as a directed graph, the nodes corresponding to web pages and the edges corresponding to links. Each edge will be labeled by the link identifier of its corresponding link. We define a *browsing path* as a sequence $<n_1, e_1, n_2, e_2, n_3, \ldots, n_{q-1}, e_{q-1}, n_q>$, where

1. For $1 \le j \le q$, n_j is a node corresponding to web page P_j.
2. For $1 \le j \le q-1$, e_j is an edge corresponding to link L_j, which is a link from page P_j to page P_{j+1}.

From the complete set of web pages under consideration, we extract a global set of textual keywords, as well as a set of visual keywords and structural keywords. We are experimenting with various techniques for this task.

For each multiset[1], M, of sub-paths that we are to analyze, we then form three matrices: a term-path matrix, an image-path matrix, and a structure-path matrix.

TP_{ij}, the $(i,j)^{th}$ element of the term-path matrix, TP, is determined by the strength of the presence of the i^{th} textual keyword, t_i, along the j^{th} browsing path, $<n_{1,j}, e_{1,j}, n_{2,j}, e_{2,j}, n_{3,j}, \ldots, n_{p-1,j}, e_{q-1,j}, n_{q,j}>$, as well as how many times this browsing path occurs in M. The strength of t_i is determined by how many times this term occurs on pages $P_{1,j}, \ldots, P_{q,j}$, how much time the user spends examining the page(s) where t_i occurs, and how close each occurrence of t_i on page $P_{k,j}$ is to both the outgoing anchor position of link $L_{k,j}$ and the incoming anchor position of link $L_{k-1,j}$, if the latter exists.

Similarly, IP_{ij} [SP_{ij}], the $(i,j)^{th}$ element of the image-path matrix, IP [SP], is determined by the strength of the presence of the i^{th} visual [structural] keyword, v_i, along the j^{th} browsing path, as well as how many times this browsing path occurs in M.

In the following discussion, we will be referring to an overall keyword-path matrix, KP. This matrix will be either TP, TP concatenated with IP, TP concatenated with SP, or TP concatenated with both IP and SP. If TP is an $r \times t$ matrix and IP is an $s \times t$ matrix, then the concatenation of TP with IP is defined as the $(r+s) \times t$ matrix, KP, where,

$$KP_{i,j} = \begin{cases} TP_{i,j} \text{ if } i = 1,\ldots,r \\ IP_{i-r,j} \text{ if } i = r+1,\ldots r+s \end{cases}$$

The other concatenations are similarly defined.

After suitably normalizing KP based on the distribution of the various keywords over the browsing path collection, we then perform latent semantic analysis on KP, which we assume to be an $m \times n$ matrix. That is, we do the following,

1. We use the singular value decomposition to find matrices U, Σ, and V, such that $KP = U\Sigma V^T$, and
   - U is an $m \times r$ matrix, where r = rank(KP), while Σ is an $r \times r$ matrix, and V is $r \times n$
   - U^TU and V^TV are both equal to the $n \times n$ identity matrix
   - $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0)$, where r = rank(KP) and the singular values are $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_r > 0$
2. Choosing[2] an appropriate $k < n$, we define
   - U_k as the submatrix of U consisting of its first k columns
   - Σ_k as the submatrix of Σ consisting of its first k rows and columns
   - V_k as the submatrix of V consisting of its first k columns
3. Defining $A_k = U_k\Sigma_kV_k^T$, we then have that A_k is the best rank k approximation to A. As is well known, A_k reveals the latent structure of A, combining linear combinations of keywords into new concepts that are more meaningful.

An obvious question concerns the nature of the sub-paths we are going to use in the above calculations. One of the approaches we are experimenting with has the following iterative structure:

1. Choose an initial set of subpaths for the above calculation
2. Calculate webpage semantics in terms of the results of the above calculation
3. Define the semantic breakpoints of user browsing paths as explained below
4. Recalculate webpage semantics based on these semantic breakpoints, as explained below

[1] Some sub-paths occur more than once, as many users trod the same byways and the same user travels the same way many times.

[2] An interesting technique for finding an optimal value of k can be found in [ZMS98].

5. Recalculate the semantic breakpoints of user browsing paths based on the new webpage semantics
6. Keep iterating steps 4-5 until a stopping criterion, based on the convergence of the webpage semantics and the semantic breakpoints, is satisfied

For Step 1, we choose an initial set, Σ, of subpaths to be all subpaths of length less than or equal to some fixed, arbitrary value, λ, which has the property that no subpath in Σ which ends at a webpage, w, is a subsequence of another subpath in Σ which also ends at webpage w. Each of these subpaths can be represented by a point, found via latent semantic analysis, in a reduced dimensional space. Call this set of points **P**. The semantics of a web page, w, can then be defined, in Step 2, as the subset of **P** corresponding to the sub-paths that end at page w.

The semantic breakpoints of a subpath of a user's browsing path, $<n_1, e_1, n_2, e_2, n_3, \ldots, n_{q-1}, e_{q-1}, n_q>$, Step 3, may then be found as follows. For each of the webpages on this browsing path, $P_1, \ldots, P_q$, calculate its semantics as defined above. Suppose these semantics are $\mathbf{P}_1, \ldots, \mathbf{P}_q$, respectively, where each $\mathbf{P}_j$ is a set of points, as mentioned above. We can formulate a distance function on point sets, *d*, such that, if $d(\mathbf{P}_i, \mathbf{P}_j)$ is less than some threshold, we input both $\mathbf{P}_i$ and $\mathbf{P}_j$ to an intersection-like operator, *intersect,* producing a new, combined semantics. We then execute the following,

1. i = q
2. $\mathbf{T}_i = \mathbf{P}_i$
3. c = true
4. while (c and i > 1)
5. if $d(\mathbf{T}_i, \mathbf{P}_{i-1}) \geq$ threshold then c=false
6. else {$\mathbf{T}_{i-1} = intersect(\mathbf{T}_i, \mathbf{P}_{i-1})$
7. i = i-1}

When finished, we determine the value of i. If i = 1, then there are no semantic breakpoints in the subpath, whereas if i > 1, the semantic breakpoint in the subpath closest to webpage P_q lies between P_{i-1} and P_i. See Figure 2 for an illustration of this process.

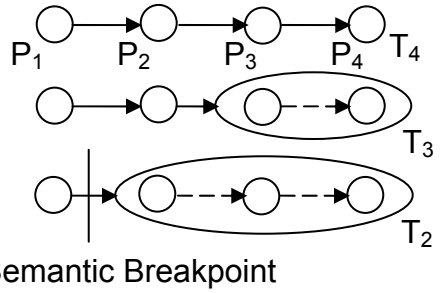

Figure 2 – Semantic Breakpoint Calculation

In Step 4, for a given webpage, P_q, we consider all subpaths ending at P_q. For each subpath, we find the semantic breakpoint, if any, as above. If there are no semantic breakpoints in the given subpath, that subpath remains unchanged. If, however, the semantic breakpoint closest to P_q lies between P_{i-1} and P_i, we transform the given subpath to the subpath starting at P_i and ending at P_q. We then use these transformed subpaths to recalculate P_q's semantics. See Figure 3 for an illustration of this process.

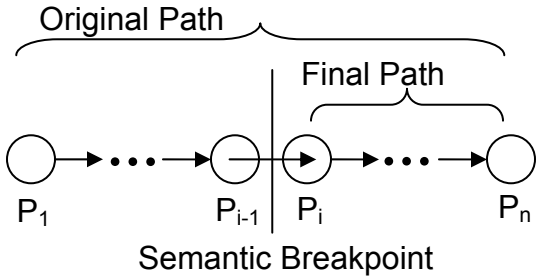

Figure 3 – Recalculating Webpage Semantics

3 CONCLUSION

We have presented our notion of webpage semantics, which are an emergent property of their use. This is similar to the approach of Santini, Gupta and Jain [5] for image databases. We do not deny, however, that webpage authors can also contribute webpage semantics via the classic semantic web. Our approach, however, follows from our belief that webpages are accessed, used, appreciated, and enjoyed for myriad reasons, some of which are quite different from what the webpage's author intended or realized.

We are implementing a proof-of-concept system for our approach. Without some form of cooperation, it is virtually impossible to get valid web usage data from surfers on the web. If there were a new standard to capture this information, it would be possible to send out spiders across the web to gather this data, just as keywords are captured from web pages to enable web search engines to function. In lieu of this, we are working with a large website instead of the entire web and deriving the necessary data from the access logs.

Semantic of web pages derived in the last two objectives can be applied to web information retrieval in a number of ways. It can be used to improve the effectiveness of web search engines, to enhance the organization of web servers, as well as to pre-fetch documents for clients to tolerate web access latency.

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American.* May 2001.

[2] S. Deerwester, S.T. Dumais, G.W. Furnas, et. al. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* 41(6):391-407. 1990.

[3] V. Gudivada and V.V. Raghavan. Content-Based Image Retrieval Systems. *IEEE Computer,* 28(9):18-22. 1995.

[4] T.I. Ibrahim and C.-Z. Xu. Neural Net Based Pre-fetching to Tolerate WWW Latency. *Proceedings of the IEEE 20th International Conference on Distributed Computing Systems.* Taipei, Taiwan:636-643. 2000.

[5] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*. 13(3):337-351. 2001.

[6] U. Shardanand and P. Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth. *Proceedings of the ACM conference on Human Factors in Computing Systems,* Denver, Colorado:210-217. 1995.

[7] R. Zhao and W.I. Grosky. Narrowing the Semantic Gap – Improved Text-Based Web Document Retrieval Using Visual Features. *IEEE Transactions on Multimedia.* 4(2):189-200. 2002.

---

[i] To appear in Sigmod Record, December 2002