Who is SMILing on the Web?

Gal Ashour, Byron Dom, John Golden, Jan Pieper and Savitha Srinivasan
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120 USA
E-mail:ashour@haifa.ibm.com, dom, jgolden, jhpieper, savitha@almaden.ibm.com

Dick Bulterman
Oratrix Development
Valeriusplein, 30 in Amsterdam, The Netherlands
E-mail:Dick.Bulterman@oratrix.com

INTRODUCTION

The amount of streaming media available on the web is orders of magnitude less than the amount of text and the resources for locating relevant information contained in streaming media [4], such as search engines, web directories and portals are not as well developed as those for text. The indexing challenges offered by the unstructured, spatio-temporal nature of media together with the lack of universal media standards largely accounts for this. The W3 Synchronized Multimedia Integration Language (SMIL) [6] standard partially addresses these issues with the support for creation and presentation of complex streaming media across multiple formats. Our focus here is the characterization of web pages that are *SMILing* or *streaming complex media presentations*. Since exhaustive enumeration of web pages is impractical, we take a sampling approach. Experimental validation shows that less than 1% of web pages contain links to streaming media, of which 65% is RealMedia, 11% is Windows Media, 2% is SMIL and 22% is other formats. We also investigate the distribution of links from web pages to streaming media and find it skewed to small numbers of both in (to media files) and out (from web pages) links.

METHODOLOGY

We base our estimates for the web on a set of 60 million pages whereas the latest estimate of web size of which we are aware [2] put the size at approximately 1 billion pages. To gain confidence in our estimates we performed the following sampling experiment. We detect web pages that link to or embed streaming media on a set of 2 million pages, and then again on the entire set of 60 million pages. Web pages that link to or embed streaming media were found to be of the following types:

Static page: This type of page contains the name and extension of the media file and its location.

Client-side dynamic page: This type of page dynamically creates a browser page that contains media links using a scripting language (i.e. JavaScript, VBscript etc.).

Server-side dynamic page: This type of page may return a page that links to streaming media in response to user click.

We view the detection of streaming media as binary classification, which can be associated with a function $f(x) = y$ where $x_1, x_2, ..x_d$ are the features in a given document x. Thus a document is represented by $x = (x_1, x_2, ..x_d)$ and $y = \{1, 0\}$ represents discrete target values where 1=streaming media page and 0=not a streaming media page. Each page is represented by a single feature vector of weighted terms. A simple boolean classifier using a decision list was developed with a high level of empirical inductive bias towards the terms that were observed to be high-confidence indicators of streaming media. We are interested in the in-links and out-links, the *hub* score of the media pages, and the *authority* score of the associated media files. The definitions we use for these terms are virtually identical to those used in [3]. The hub score of media page u is equal to the sum of the authority scores of media files v that it links to: $h(u) = \sum_{u \to v} a(v)$, and the authority score of media file v is equal to the sum of the hub scores $h(u)$ of all the media pages u that link to it: $a(v) = \sum_{u \to v} h(u)$.

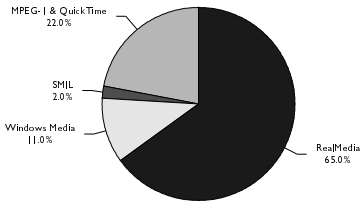

Figure 1: Percentages of media files of different media formats on the web

EXPERIMENTAL FINDINGS

We measured the number of pages that contain or embed streaming media on an initial sample size of 2 million pages. We repeated this experiment on a larger sample size of 60 million pages. We observed that approximately 0.6% of web pages contained links to streaming media on the initial sample size of 2 million web pages. On the larger sample size of 60 million pages we observed a 0.56% effect, which complies with the prediction. Figure 1 summarizes our findings related to streaming media obtained from 60 million web pages. We observe that RealMedia is the dominant media format in our sample data set, and so we focus on RealMedia files for the link analysis[5,7]. The two-dimensional histograms of Figure 2 show the degree of linkage between media pages and media files using in-link/outlink counts and hub/authority scores. Not suprisingly the distributions are heavily skewed toward the low end (1 link per page/file). The degree to which this is true for the in-link distribution may be due to the fact that most media files are attached to a single introductory web page and other web pages refer to that page rather than the media file itself. We also report similar analyses for *hub* and *authority* scores.

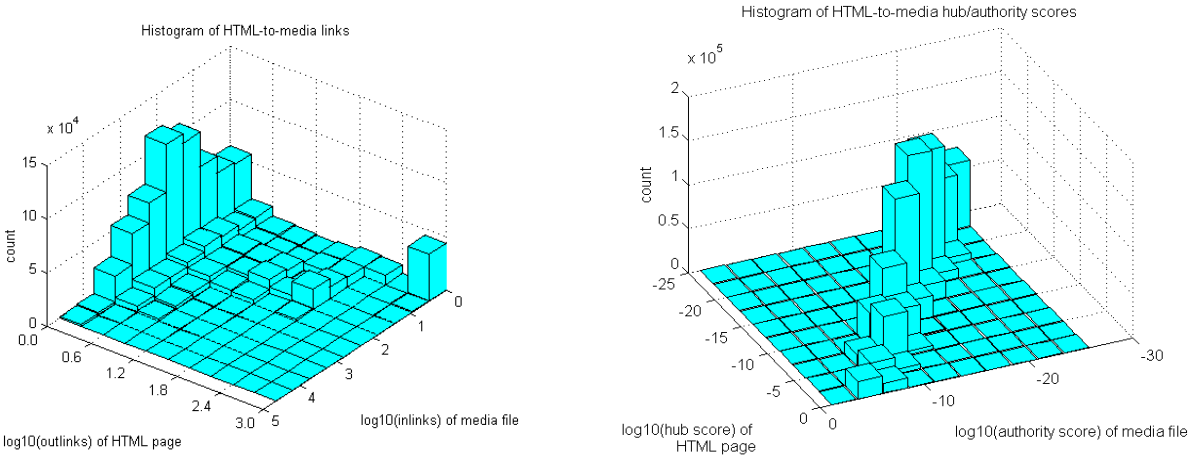

Figure 2: Histogram of in-link/out-link counts and hub/authority scores of media pages and media files

CONCLUSIONS

This study was motivated by the requirement for effective resources to locate relevant information in streaming media. Our approach is to predict and validate relevant characteristics using a probabilistic model for the entire web. There are several useful observations from this initial snapshot of the web: First, less than 1% of web pages contain streaming media links, of which 65% is RealMedia, 11% is Windows Media format, 2% is SMIL and 22% other formats. Although there is little SMIL content on the web, more than half of the streaming media content providers use metafiles. These may be considered to be implicitly SMIL since they support some level of sophisticated streaming media. Analysis of links from web pages to streaming-media files showed distributions highly skewed towards small numbers of links. That is few pages had large numbers of links to media and few media files had links from large numbers of web pages. This is more or less as expected. The most interesting pages/files are those with large numbers of in links/out links and even more so those with high hub and authority scores.

ACKNOWLEDGMENT: We thank A. Tomkins and K. McCurley for providing us with the web crawl data.

REFERENCES

1. Google Inc., “Google Launches World's Largest Search Engine". google.com press release. 6/26/2000. http://www.google.com/pressrel/pressrelease26.html
2. Kleinberg, J. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms. 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.
3. Lawton, G. Video Streams into the Mainstream. In IEEE Computer Magazine, July 2000.
4. SMIL Quick Reference, RealNetworks, Inc., See URL at http://service.real.com/help/ library/guides/production/htmfiles/smilref.htm. December 1998.
5. Rutledge, L., van Ossenbruggen, L., Hardman, L., Bulterman, D. Anticipating SMIL 2.0: The Developing Cooperative Infrastructure for Multimedia on the Web. In Proceedings of WWW-8, Toronto, Canada, May 1999.
6. Windows Media Player, All about Windows Media metafiles, See URL at http://msdn.microsoft.com/workshop/imedia/ windowsmedia/crcontent/asx.asp. Microsoft Corp., April 2000.