A software development plan for an Ads Matching system

1. **Features (user stories)**

a) **Site Registration:**

The site registration will consist of a website using HTML and php where a user can register their website for use with our service. The site will have a form for entering the URL of the site, user's name and email address. This information will be used by mySql to populate the user's information in the database and register the website for parsing.

b) **Ad Registration:**

Ad (i.e. advertisement) registration will consist of a website using HTML and php where an advertiser can upload an ad to use our service. The site will have a form entering the website that the ad should direct to, user's name and email address. The site will also be able to upload an image for the ad (or a URL for the image). If an image was uploaded this will be stored on our server and the URL for the location of the location will be entered into our database along with the advertiser's information using mySql.

c) **HTML Parsing:**

The parser component will take the contents of a website and strip out the HTML this part will make use of some regular expressions to remove the HTML tags and leave the rest of the content. After the HTML is removed the content of the site will be run through a language parser that will deconstruct the content into sentences and their component parts. Last semester we made use of the OpenNLP natural language parser library to do our sentence parsing. At the moment we are unsure whether we will continue using this library or need another solution as we ran into performance problems with the OpenNLP parser. After the sentences have been parsed their topic and positive or negative use will be determined with a simple scoring system based on the most frequent use. The top few nouns that are used the most frequently will be used as topics. If there is an adjective modifying a keyword a running total of positive and negative occurrences of the topic is kept. After parsing the parser will pass it's output to the ad matching system to match it with an appropriate ad.

d) **Ad matching**

We will be using Flash to display the ads on the websites. My-SQL will be used to access keywords of ads and the information generated by the parser. PHP will be used for the algorithm which analyzes the ads and parsed website data from the database to determine the best ad for the site. Originally the ad matching project had no way of reliably showing the ads on the website, but now by using flash or some similar medium we can have more control over how the ad is displayed. Flash also allows for flexibility of ad content which could be implemented later.

2. **Platform and Architecture**

Server: Linux (babbage), web server
Client: web browser (IE, Firefox, etc)
Database: MySQL

3. **Programming languages**

PhP (main), HTML, Java (optional for parser)

4. **Time Table (deadlines for major steps)**

Planning: Sept. 15
Design: Sept. 29
Coding: Oct. 20
Testing: Dec. 2
Deployment and documentation: Dec. 13

5. **Task Assignments (assign coordinators and members for tasks)**

Design: all the members, coordinated by Jack
Coding
- Feature 1 – site registration: John
- Feature 2 – ads registration: Mike
- Feature 3 – HTML parsing: Jack
- Feature 4 – ads matching: Peter
- Web design – John
- Data preparation: Mike
- Coding is coordinated by Mike

Testing
- Unit Testing: Jack, John, Mike, Peter for features assigned to them
- Integration and acceptance testing: all members, coordinated by John

Report & Presentation
- By all members, coordinated by Peter

6. **Technical Challenges**

HTML parsing is a significant challenge. No good HTML parser is developed or found yet. Here are some resources that may help us find a solution.

a) Java: http://htmlparser.sourceforge.net/
b) Perl: http://mail.perl.org.il/pipermail/perl/2002-July/000208.html
c) PhP html parser: http://php-html.sourceforge.net/index.php