Semantic Models and Corpora Choice when using Semantic Fields to Predict Eye Movement on Web pages

Benjamin Stone∗, Simon Dennis

∗School of Psychology, Level 4, Hughes Building, The University of Adelaide, Adelaide, SA 5005 Australia
Department of Psychology, Ohio State University, 1827 Neil Ave, Columbus, OH 43210 USA

Abstract

Ten models are compared in their ability to predict eye-tracking data that was collected from 49 participants’ goal-oriented search tasks on a total of 1809 Web pages. Forming the basis of six of these models, three semantic models and two corpus types are compared as components for the Semantic Fields model (Stone and Dennis, 2007) that estimates the semantic salience of different areas displayed on Web pages. Latent Semantic Analysis, Sparse Non-Negative Matrix Factorization, and Vectorspace were used to generate similarity comparisons of goal and Web page text in the semantic component of the Semantic Fields model. Surprisingly, Vectorspace was consistently the best performing semantic model in this study. Two types of corpora or knowledge-bases were used to inform the semantic models, the well known TASA corpus and other corpora that were constructed from the Wikipedia encyclopedia. In all cases the Wikipedia corpora out performed the TASA corpora. A non-corpus based Semantic Fields model that incorporated word overlap performed more poorly at these tasks. Three baseline models were also included as a point of comparison to evaluate the effectiveness of the Semantic Fields models. In all cases the corpus-based Semantic Fields models outperformed the baseline models when predicting the participants’ eye-tracking data. Both final destination pages and pupil data (dilation) indicated that participants’ were actively performing goal-oriented search tasks.

Key words: Semantic Fields model, semantic salience, Web navigation, Semantic Models, LSA, SpNMF, Vectorspace, eye tracking, pupil dilation

1. Introduction

The exponential increase in Internet usage over the last decade has motivated psychological researchers to examine Web users’ behavior in this virtual environment. Research focusing on user behavior in Web page environments can generally be delineated into two main streams: display-based and semantics-based research. While both methods to some degree attempt to predict the area on a Web page that a user will focus their attention towards, they approach this task in different ways. Display-based research has focused on perceptual aspects of the Web page, and explores components such as element and menu position, color usage, font style, and animation of graphics (Faraday, 2000, 2001; Ling and Van Schaik, 2002, 2004; McCarthy et al., 2003; Pearson and Van Schaik, 2003; Grier, 2004; Rigotti and Gerbino, 2004). Alternatively, when attempting to predict users’ Web page navigation, semantic-based research matches Web users’ information needs to the concepts displayed within the textual content of Web pages (Blackmon et al., 2002; Chi et al., 2003; Pirolli and Fu, 2003; Brumby and Howes, 2003, 2004; Cox and Young, 2004; Kaur and Hornof, 2005). Overall, display-based and semantics-based research into Web users’ visual search of Web page hyper-links has indicated that the user’s search processes are influenced by factors such as: text semantics, element position, aesthetic qualities of elements, and environmental learning1.

Several researchers have highlighted the importance of combining display-based and semantic information when modeling users’ navigation through Web sites (Blackmon et al., 2002; Chi et al., 2003; Pirolli and Fu, 2003; Kaur and Hornof, 2005; Stone and Dennis, 2007). Research that has combined display and semantic information when predicting Web users’ behavior include the Cognitive Walkthrough for the Web (CWW, Blackmon et al., 2005), the Bloodhound Project (Chi et al., 2003), and the Latent Semantic Analysis - Semantic Fields model (LSA-SF, Stone and Dennis, 2007). For a detailed description of the CWW and Bloodhound Project the reader is directed to Blackmon et al. (2005) and Chi et al. (2003), respectively.

1.1. Semantic Fields (SF)

In a previous article, we presented the LSA Semantic Fields (LSA-SF) model (Stone and Dennis, 2007), which was used to predict the eye movements of 49 participants recorded during goal-oriented search tasks on three Web sites. The LSA-SF model used Latent Semantic Analysis (LSA, Landauer et al., 2007) to calculate the similarity between a textual representation of the users’ goal and each of the textual elements dis-

1For Web site navigation and environmental learning see Pan et al. (2004)
Figure 1: Semantic Fields Map using Vectorspace and a corpus drawn from Wikipedia. Participant’s eye tracking data is super imposed using black dots. While the original SF model only used LSA, the SF models presented in this paper incorporate word overlap, Vectorspace, LSA, and SpNMF semantic models played on a Web page. Using a decay function, these LSA estimates of similarity were then distributed and summed over each pixel position for all of the textual elements contained on a Web page (see Equation 1). Combining the semantic information \( (L) \) with distance \( (d_i(x,y)) \) from its display position using a decay function, enabled the production of maps of information density for each Web page in our study (see Figure 1).

\[
SF(x, y) = \sum L_i e^{-\lambda d_i(x,y)}
\]  

1.2. Focus of this paper

Initially, the TASA corpus was used as a knowledge based for the LSA component of the LSA-SF model (Stone and Dennis, 2007). However, recent research has suggested that a better fit with the knowledge domains required to model human similarity judgments in document comparison tasks may be achieved using document sets which are retrieved from the online encyclopedia Wikipedia (Gabrilovich and Markovitch, 2007; Stone et al., in press). Also, while LSA is certainly the best studied statistical semantic model, Stone et al. (in press) found that other models such as the vector space model (Vectorspace, Salton et al., 1975) and Sparse Nonnegative Matrix Factorization (SpNMF, Xu et al., 2003; Shashua and Hazan, 2005) out-performed LSA when estimating human judgments of paragraph similarity. Based on these findings, in this paper we present a comparison of three semantic models (LSA, SpNMF, and Vectorspace), and two types of knowledge base (TASA and Wikipedia), when used as components in the generation of Semantic Fields. Furthermore, the performance of these revised Semantic Fields models is assessed on the eye-movement dataset presented in Stone and Dennis (2007).

2. Method

2.1. Participants

Eye-movement data generated by 49 university participants, 27 males and 22 females, on three Web sites was recorded during nine goal-oriented search tasks. Participants were recruited from either a first-year pool volunteering in exchange for partial course credit, or other members of Adelaide University who were paid $30 for their time. Participant ages ranged from 16 to 57 (M=22y3m SD=1y). Also, there was a positive skew in the samples age distribution, with 93 percent of participants were younger than 31 years old.

All participants had both previous computer and Internet experience. Self-reported years of Internet experience ranged from 4 to 17 years (M=8y8m, SD=4m), with self-reported frequency of Internet use ranging between 2 to 75 hours per week (M=14h14m, SD=1h52m). Based on these self reports, the group appeared on average to be very experienced users of the Internet.

2.2. Apparatus

2.2.1. Behavioural recording equipment and software

The IETracker program was developed to record both participants’ behaviour during website search tasks and web page display characteristics. These observations are accomplished by programatically controlling, and integrating, Microsoft Internet Explorer (Version 6) and the ViewPoint EyeTracker PC-60 QuickClamp System. User and site specific data collected during this exploratory experiment included: eye-tracking; hyperlink clicking; and web page composition (location, semantics, images, colour, style, and size of web page elements). All data was then stored in a Postgresql database for later analysis.

2.2.2. Web sites

Three Web sites were chosen from the Internet:
www.missionaustralia.com.au (Mission Australia)
www.greencorps.com.au\(^2\) (Green Corps Australia)
www.whitelion.asn.au (White Lion Australia)

Static versions of these sites\(^3\) were pre-fetched in December 2005 to avoid changes created by Web site updates. These Web sites are all similar in the type of service they provide, such that they all offer services to disadvantaged members of the community. The Web sites were chosen because they were sufficiently

\(^2\)The Green Corps Australia URL is no longer used, the Australian Government has change both the Web site and its URL, which can be viewed here: http://www.greencorps.gov.au

\(^3\)The static versions of these Web sites can be found here: http://www.psychology.adelaide.edu.au/mall_lah/lsa-sf_sites/
complex that searching for information on these sites would be a non-trivial task for participants.

The original Stone and Dennis (2007) study used eye-tracking data recorded on 1842 Web pages, in this study only 1809 Web pages are used. It was found that 33 page views included in the original dataset had to be removed for two reasons. Some pages were omitted because data for all calibration points had not been recorded. This stemmed from participant head movements and overt glances during the calibration procedure. Also, several Web pages had been designed to catch user clicks on PDF files. These “catch-pages” only contained one textual element that informed the participants that they had either found their goal or had not and should click the “back” button. These pages have been removed because they were not part of the original Web sites, and their simple one element construction with black text on a white background probably favored the Semantic Fields model.

2.3. Procedure

Participants were required to search for the following three target pieces of information on each of the three web sites.

Mission Australia:
1. Who is currently the Chief Operating Officer of Mission Australia?
2. You are interested in working for Mission Australia. Search their website for the current job vacancies available at Mission Australia.
3. You are currently researching homelessness in young people and have heard that Mission Australia has recently published a report called “The voices of homeless young Australians”. Search the Mission Australia website for this report into youth homelessness.

Green Corps:
1. You want to know more about Green Corps management. Find out who is the National Program Manager of Green Corps.
2. Find what environmental and heritage benefits are contributed by Green Corps.
3. Find the online Expression of Interest form to apply to become a Green Corps Partner Agency.

White Lion:
1. Find out who is the current President of White Lion
2. You are interested in becoming a mentor for young people. Find out how to become one of White Lions mentors.
3. You are interested in financial viability of White Lion as a business. Find out which Government Departments are supporters of the White Lion organization.

Each task was read aloud to the participants twice before they commenced their search. Moreover, they were asked to signal the experimenter (with a hand gesture) at any time they wanted the search task repeated aloud. A three by three Latin square design was used to control for order effects in the display of each website. Also, a nested three by three Latin square design was used for the same purpose to guide the presentation order of each of the target tasks.

After an initial calibration procedure, using the Viewpoint eye tracking software, participants were given their search task in the manner described above and commenced their search. Given the physical structure of the Viewpoint Eye-Tracker (which uses a mounted camera with forehead and chin rests) and some participant’s predisposition to fidget, it was necessary to perform additional eye-tracking calibration during the search tasks. This additional calibration required the participants to focus on targets in nine separate regions of the monitor. Moreover, these targets were automatically displayed on the screen by the IETTracker program after each hyperlink clicked during the participant’s search task. To elaborate on this point further, if a participant clicked through four pages in their search for the target information, then four extra calibration procedures were performed during this search task; each calibration performed after leaving the page that was clicked on and before the next page was displayed to the participant.

The calibration of the eye-tracking data was performed in a different way in this study compared to Stone and Dennis (2007). In the previous study, an algorithm was used to adjust the eye-positions relative to participants’ movement during experimentation. However, to make this process more transparent in the current study, the participants’ eye-points were repositioned using the average offsets recorded over all nine calibration points.

Finally, participants were instructed that when they believed that they had found the target information that they should then click on the ‘HOME’ icon (which is an image of a house on Internet Explorer’s menu bar). This was followed by one more round of automated calibration. On average, it took participants three page views to find their goal on the White Lion Web site, four page views on the Green Corps Web site, and five page views on the Mission Australia Web site.

2.4. Semantic Fields Models

Four semantic models were incorporated into the Semantic Fields model\(^4\): word overlap, Vectorspace model (Salton et al., 1975), Latent Semantic Analysis (Landauer et al., 2007), and Sparse Nonnegative Matrix Factorization (Xu et al., 2003).

2.4.1. Word Overlap

Simple word overlap was used as a baseline in this research, and also because of its solid performance at paragraph level comparisons (Stone et al., in press). Term frequencies are calculated for each document, and similarities are then measured as cosines (see Equation 2) of the resulting document vectors.

---

\(^4\)The SEMMOD semantic models package was used to incorporate the Vectorspace model, Latent Semantic Analysis, and Sparse Nonnegative Matrix Factorization into the Semantic Fields model. The SEMMOD semantic models package is release under the GNU License and can be found here: http://mall.psy.ohio-state.edu/wiki/index.php/Semantic_Models_Package_%28SEMMOD%29
2.4.2. The Vectorspace model

The Vectorspace model (Salton et al., 1975) assumes that terms can be represented by the set of documents in which they appear. Two terms will be similar to the extent that their document sets overlap. To construct a representation of a document, the vectors corresponding to the unique terms are multiplied by the log of the frequency within the document and divided by their entropy across documents and then added. Using the log of the term frequency (TF) within documents identifies higher frequency or important words in those documents. While dividing by the entropy reduces the impact of high frequency words that appear in many documents in a corpus. Similarities are measured as the cosines between the resultant vectors for different documents.

2.4.3. Latent Semantic Analysis (LSA)

LSA (Landauer et al., 2007) started with the same representation as the Vectorspace model - a term by document matrix with log entropy weighting. In order to reduce the contribution of noise to similarity ratings, however, the raw matrix is subjected to singular value decomposition (SVD). SVD decomposes the original matrix into a term by factor matrix, a diagonal matrix of singular values and a factor by document matrix. Typically, only a small number of factors (e.g., 300) are retained. To derive a vector representation of a novel document, term vectors are weighted, multiplied by the square root of the singular value of the term frequency (TF) within documents identifies higher frequency or important words in those documents. While dividing by the entropy reduces the impact of high frequency words that appear in many documents in a corpus. Similarities are measured as the cosines between the resultant vectors for different documents.

2.4.4. Sparse Nonnegative Matrix Factorization (SpNMF)

Nonnegative Matrix Factorization (Xu et al., 2003) is a technique similar to LSA, which in this context creates a matrix factorization of the weighted term by document matrix. This factorization involves just two matrices - a term by factor matrix and a factor by term matrix - and is constrained to contain only nonnegative values. The nonnegative matrix factorization has been shown to create meaningful word representations using small document sets, in order to make it possible to apply it to large collections we implemented the sparse tensor method proposed by Shashua and Hazan (2005). As in LSA, log entropy weight term vectors were added to generate novel document vectors and the cosine was used to measure of similarity.

2.5. Corpora

2.5.1. TASA

The Touchstone Applied Science Associates (TASA) corpus was constructed to represent the reading material that is covered by American students up to their first year of college. The 35471 documents contained in the TASA corpus range over nine content areas: language arts and literature, social science, science and math, fine arts, home economics and related fields, trade, service and technical fields, health, safety and related fields, business and related fields, popular fiction and nonfiction (Budiu et al., 2007).

\[
\cos \theta = \frac{v_1 \cdot v_2}{||v_1|| \ ||v_2||}
\]  

(2)

2.5.2. Wikipedia

Wikipedia was used as a generic set of documents from which smaller targeted sub-spaces could be sampled and compiled. Wikipedia is maintained by the general public, and has become the largest and most frequently revised or updated encyclopedia in the world. Critics have questioned the accuracy of the articles contained in Wikipedia, however research conducted by Giles (2005) did not find significant differences in accuracy of science based articles in Wikipedia and similar articles contained in the Encyclopedia Britannica. In total, 2.8 million Wikipedia entries were collected and are current to March 2007. However, this document number was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. The incomplete articles removed were identified if they contained the phrases like “help Wikipedia expanding” or “incomplete stub”.

To enable the creation of sub-space corpora, Lucene5 (a high performance text search engine) was used to index each document in the Wikipedia corpus. Lucene allows the user to retrieve documents based on customized Boolean queries. These queries can include wild-card operators like ‘the star’ (*) to retrieve multiple results from word stems. Like the more well known search engine Google, the documents returned by Lucene are ordered by their relevance to a query.

The three search tasks given to the participants for each Web site were enumerated above in Section 2.3. Based on keywords selected from these tasks, the following Lucene queries (in italics) were constructed to retrieve the 1000 document sub-spaces from the Wikipedia corpus for each website.

<table>
<thead>
<tr>
<th>Mission Australia:</th>
<th>(“mission australia”) OR (“chief operating officer”) OR (“employment”) OR (“homeless&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Corps:</td>
<td>(“green corps”) OR (“national” AND “program” AND “manager”) OR (“environment” AND “benefit”) OR (“heritage” AND “benefit”) OR (“partner” AND “agency”)</td>
</tr>
<tr>
<td>White Lion:</td>
<td>(“white lion”) OR (“company president”) OR (“organization* president”) OR (“mentor*” AND “young” AND “people”) OR (“government department*” AND “support&quot;)</td>
</tr>
</tbody>
</table>

2.5.3. Appending Web pages as Documents - The creation of TASA-WEB and WIKI-WEB corpora.

When examining semantic models’ ability to perform similarity estimates on paragraphs, Stone et al. (in press) found that including stimulus paragraphs into the semantic models’ knowledge base (or corpus) provided an effective method of corpus improvement. Performance gains obtained in this way most likely result from context being given to novel stimulus

---

5PyLucene is a Python extension that allows access to the Java version of Lucene: http://pylucene.osafoundation.org/
terms that were not contained in the original corpus. To illustrate this point, the TASA corpus may not contain the phrase “Green Corps”, but may contain the words “contribute” and “community”. By appending documents (i.e., the text contained in Green Corps Web site) to the TASA corpus, the semantic model can more accurately make associations between these terms. Based on this Stone et al. (in press) finding, when creating corpora for each of the three Web sites in this study, the textual content of each of the Web pages viewed by participants on that Web site were appended to the TASA and Wikipedia corpora. For example, on the Mission Australia Web site, overall 57 unique Web pages were viewed by participants during the experiment. So, the textual content from these Web pages was used to construct a mini-corpus of 57 documents, which was then appended to both the TASA corpus and Wikipedia subspaces for Mission Australia corpora. Highlighting this appended data, the naming conventions TASA-WEB and WIKI-WEB are used throughout this paper.

2.6. Baseline models to estimate eye-position

Three alternative models were designed as baselines to assess the success of the Semantic Fields models when predicting participants’ eye-positions when they are engaged in goal-oriented search tasks on the three Web sites. These baseline models are the Flat, Non-Flat, and No-Model.

2.6.1. Flat Model

The Flat model is the simplest model, it assumes that the eye-position has equal likelihood of being focused on all pixels contained on the Web page. Given the total number of eye-points (EP), and the total number of pixels that an eye-point (p) could be located in (1280 x 1024), for each page (i) that is viewed (V) by participants, the Flat model calculates the log-likelihood of the eye-points on any given Web page as

\[
LL_{\text{Webpage}} = \sum_{p \in \text{EP}} \log \left( \frac{1}{1280 \times 1024} \right)
\]

(3)

2.6.2. Non-Flat Model

The Non-Flat model is similar to the Flat model, with the exception that it gives more weight to the probability estimates for those eye-points found in textual elements\(^7\). The Non-Flat model is displayed in Equation 4, where for each Web page (i) that is viewed (V) by participants, N is the number of pixels in text elements, M is the number of pixels outside text elements, A is the number of eye-points in text elements and B is the number of eye-points outside text elements. Furthermore, \(\hat{w}\) is the optimized probability of an eye-point being in a text element (see Equation 4). The maximized log-likelihood (ML) over all Web pages viewed by participants occurred at a MLE of \(\hat{w} = 3.41\) for this sample. Therefore, participants were 3.41 times more likely to focus their eyes on the textual elements on a Web page than focusing on other areas. In accordance with this finding, the MLE has been used to calculate Non-Flat model log-likelihoods and assigned a greater weighting to eye-points recorded in these textual elements than the non-textual elements.

\[
ML_{\text{NONFLAT}} = \sum_{i \in V} ML_{\text{Webpage}}
\]

\[
ML_{\text{Webpage}} = A \log \left( \frac{\hat{w}}{\hat{w}N + M} \right) + B \log \left( \frac{1}{\hat{w}N + M} \right)
\]

(4)

2.6.3. No-Model

The No-Model condition has been created to test the theory that the Semantic Fields model is driven only by the structure of the Web page, and that the semantic models do not add to the Semantic Fields model’s capacity to predict participants’ eye-positions. It takes the same parameters as the Semantic Fields model, however the semantic model coefficient is kept constant at one \((L_s = 1.0)\) in the calculation of the Semantic Fields (see Equation 1).

---

\(^6\)Note: Green Corps and White Lion Web sites’ textual data was not used when creating corpora for the Mission Australia data.

\(^7\)An example of text elements (including images that have a textual description) on a Web page are shaded in red in Figure 2.
2.6.4. Calculating the log-likelihood for Semantic Fields and No-Model conditions

The log-likelihoods for the Semantic Fields models and No-Model are calculated in the same fashion. The Semantic Field Value $S_{F_{Webpage}}$ (Equation 1) for each eye-point ($p$) is divided by the summed total of the Semantic Field Values for all pixels that Web page ($S_{F_{Webpage}}$). This process calculates the probability that a single eye point is viewed. Then, the log of these probabilities ($LL_{Webpage}$) is calculated and summed over all eye-points ($EP$) on a Web page viewed by a participant. This process of summing the log-likelihoods is repeated for each Web page ($i$) that was viewed ($V$) by participants to calculate the overall log-likelihood ($LL_{SF}$) for both the Semantic Fields and No-Model conditions (see Equation 5).

$$LL_{SF} = \sum_{i \in V} LL_{Webpage}$$

$$LL_{Webpage} = \sum_{p \in EP} \log \left( \frac{S_{F_{p}}}{S_{F_{Webpage}}} \right)$$  \hspace{1cm} (5)

3. Results

3.1. Did the participants complete their tasks successfully?

When designing the nine tasks, the researchers had identified specific destination Web pages that matched the information required by each task. Table 1 displays the number of times participants finished on the same Web page that was chosen by the experimenter. Overall, there was a 92.93 percent overlap between the expectations of the experimenter and the final landing pages chosen by participants. However, only half the participants completed the second task on the Green Corps Website to the expectations of the researchers.

It seems likely that finding the environmental and heritage benefits contributed by Green Corps was more open to subjective judgments of completion by participants than were elicited by the other questions. On reviewing the data, it appears that simply finding the words ‘environment’ and ‘heritage’ on the same page may have prompted some participants to end their search. Although all tasks had been constructed with a specific end page in mind (see Figures 5-13 in Appendix A), on reflection the other eight tasks appear to have more specific goals. For example, finding the name of committee member or a specific file on the Web site. While Green Corps Task 2 is a limitation of this study, overall all tasks the vast majority of participants found the target Web pages chosen by the experimenters.

3.2. Were the participants paying attention?

It is difficult to understand the psychophysiological connection between pupil dilation and cognitive load. However, researchers have been reporting this phenomenon for nearly a century now. The German psychiatrist Bumke, in his 1911 review of the relevant German literature, proposed that “in general every active intellectual process...produces pupil enlargement” (as cited by Hess, 1972, p. 492). A fairly consistent pattern of findings has been reported in the literature, with recordings of participants’ pupil sizes increasing during more difficult tasks when compared to those recorded during easier tasks. This effect is often reported when participants are engaged in mathematical problem solving of varying difficulty (Hess and Polt, 1964; Schaeffer et al., 1968; Boersma et al., 1970; Ahern and Beatty, 1979; Steinhaus et al., 2000; Stone et al., 2004). In a preparatory study, Stone et al. (2004) found that under similar experimental conditions to those used in this study (eye-tracker, CRT monitor, room and lighting), participants’ pupil sizes were larger whilst engaged in more difficult mathematical subtraction tasks when compared to their pupil size during easier adding tasks. Moreover, this pattern was consistently found when participants undertaking these tasks were directed to focus on twenty-five evenly spaced targets, that were positioned inside the cells of a five-by-five grid on the CRT monitor. This study indicated that greater cognitive load could be detected using the pupil width measure across a wide range of focal points on the CRT monitor.

Table 1: Percentage of overlap between the expected landing Web page chosen by the experimenter and those chosen by the 49 participants.

<table>
<thead>
<tr>
<th>Website</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mission Australia</td>
<td>100.00</td>
<td>95.92</td>
<td>95.92</td>
<td>97.28</td>
</tr>
<tr>
<td>Green Corps</td>
<td>95.65</td>
<td>53.06</td>
<td>97.92</td>
<td>82.21</td>
</tr>
<tr>
<td>White Lion</td>
<td>100.00</td>
<td>100.00</td>
<td>97.92</td>
<td>99.31</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>92.93</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3: Standardized pupil width during participants’ fixations while they were performing goal-oriented Web page navigation. Time spent searching each page is delineated into deciles.

Assuming pupil dilation is a measure of the participants’ cognitive load, then larger standardized values will reflect greater cognitive processing by the participant. In Figure 3, it is clearly shown that as participants spend more time on each Web page, the width of the pupil increases. This suggests that as participants spent more time on each page, their cognitive load also increased. It is likely that these cognitive load increases are related to the decision processes undertaken by the participants as they assess an appropriate navigation path to complete their search goal (e.g., what information is important on the Web page, and whether to click on to a more relevant page).

In the after the ninth decile pupil width decreases in all cases. This probably reflects the time period after which participants’ have either found the target or have clicked on a link to pursue and are waiting for the browser to load the next page. In both of these cases the next page participants are taken to is the calibration page. Moreover, in both scenarios one would expect a reduction in participants’ cognitive load during this time when compared to that produced by participants during time spent reading and assessing page content for information relevant to search goals.

3.3. Ten models compared using the Bayesian Information Criterion

The log-likelihoods of ten models (see Equations 3-5) constructed to predict participants eye movements are compared

---

9 As mentioned previously, the calibration display (set on a gray background) was presented prior to each Web page viewed by participants.

10 It is not possible to verify this in our data, as a time stamp was not generated for the mouse click action.
in this section using the Bayesian Information Criterion (BIC, Schwarz, 1978) for assessing model fit. The ten models include seven Semantic Fields (SF) models. One of these SF models, word overlap, does not use a knowledge base. Evaluations are only made on the co-occurrences between words in Web page element text and the textual description of user goals. The other six SF models are more complex and use backgrounding documents, half of these SF models used the TASA-WEB corpus while the other half used WIKI-WEB corpora as a knowledge base. Furthermore, three semantic models (LSA, SpNMF, and Vectorspace) were used to compare goal text to element text in these six SF models. So, for the six more complex SF models, there is a two (corpus) by three (semantic model) experimental design. The other three models compared here are the Flat, Non-Flat, and Non-Model conditions. While no parameters are set in the log-likelihood calculations for most of these models, the maximized log-likelihood equation for the Non-Flat model has one parameter (\( \hat{\omega} \)). The BIC is an appropriate method for comparing the fit of these models to the eye-data, because it adjusts for the number of parameters going into the model. Moreover, higher BIC scores indicate a better fitting model to the data.

**Table 2: Comparison of Bayesian Information Criteria (BIC) statistics calculated from log-likelihoods generated for all ten models.**

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Model</th>
<th>BIC</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIKI-WEB</td>
<td>Vectorspace</td>
<td>-10983963</td>
<td>0</td>
</tr>
<tr>
<td>WIKI-WEB</td>
<td>SpNMF</td>
<td>-10989551</td>
<td>-5588</td>
</tr>
<tr>
<td>WIKI-WEB</td>
<td>LSA</td>
<td>-10990633</td>
<td>-6670</td>
</tr>
<tr>
<td>TASA-WEB</td>
<td>Vectorspace</td>
<td>-10993390</td>
<td>-9427</td>
</tr>
<tr>
<td>TASA-WEB</td>
<td>SpNMF</td>
<td>-11001041</td>
<td>-17078</td>
</tr>
<tr>
<td>TASA-WEB</td>
<td>LSA</td>
<td>-11001286</td>
<td>-17323</td>
</tr>
<tr>
<td>-</td>
<td>No-Model</td>
<td>-11005111</td>
<td>-21148</td>
</tr>
<tr>
<td>-</td>
<td>Overlap</td>
<td>-11094311</td>
<td>-110348</td>
</tr>
<tr>
<td>-</td>
<td>Non-Flat</td>
<td>-11287790</td>
<td>-303827</td>
</tr>
<tr>
<td>-</td>
<td>Flat</td>
<td>-11417562</td>
<td>-443599</td>
</tr>
</tbody>
</table>

The results displayed in Table 2 are presented in descending order of their BIC scores. There were three interesting trends displayed in the results. Firstly, BICs were higher for the six corpus-based Semantic Fields models than the other models, therefore the Semantic Fields models provided a better fit for the eye tracking data than the other baseline models. Moreover, as would be expected, the simple Flat model performed the worst, followed by the Non-Flat model, and then the No-Model; the latter two baseline models were expected to perform better as they contain information about the structure of the Web page display. Secondly, corpus choice appeared to affect SF model performance. The SF models using the WIKI-WEB corpora outperformed the TASA-WEB corpora in all instances. Thirdly, the semantic models that were used in the SF models produced a main effect. Using either corpus as a knowledge base, Vectorspace was consistently the best performing model, followed by SpNMF and then LSA.

It is unsurprising that the word overlap model performed badly. Unlike the corpus-based semantic models, word overlap will only find similarity when there is an overlap of words between textual elements on a Web page and the textual description of a task. Alternatively, the models such as Vectorspace, LSA and SpNMF have a greater pool of words and their associates contained in the corpus to draw on when making these similarity estimates. Finding even a small degree of similarity, based on an associated word (i.e., not the exact matching of words required by the word overlap model), will generate SF model ‘heat’ from its source location within the structure of a Web page. That is, if there is no association found between the task and the text in a Web page element, then the presence of that element (even if it is a menu item), cannot inform the SF model. Furthermore, the importance of including structural information into the SF model is highlighted by the better performance of the No-Model condition (which only uses Web page structure as opposed to semantics) in comparison to the word overlap model.

### 3.4. How well does the VEC-SF model using the WIKI-WEB corpus predict the eye data?

During complex tasks that require reading, such as goal-oriented Web page search, the location of participants’ eye fixations are generally an indicator of goal-specific attentional processes (Rayner, 1998, p. 375). To accomplish their task, participants engaged in visual search must seek out locations of goal-relevant information on the display. If the Semantic Fields model is able to capture some of the variability in participants’ eye movements, then one would expect that Semantic Field values located at participants’ eye focal points will increase as participants’ spend more time viewing a page.

As reported above, the Semantic Fields model that used Vectorspace with the WIKI-WEB corpus provided the best performance when estimating the location of participants’ eye-points during web navigation. To provide the reader with a visual illustration of the Semantic Fields generated using both Vectorspace and the WIKI-WEB corpus, Semantic Field maps for the target pages for all nine tasks are displayed in Figures 5-13 in Appendix A. Figure 4 displays the mean SF values using Vectorspace with the WIKI-WEB corpus for each page viewed by participants. Following the same line of reasoning that was outlined in Section 3.2, time spent on page has been divided into deciles to avoid possible confounds produced by individual differences in participants’ task performance. On all three Websites, it appears that after an initial orienting phase that participants’ eye-movements move towards areas of greater SF values. As mentioned in the pupil section, it is possible that the sharp drop off in the last 10% of time spent on the page, captures eye movement between the click of a link and the browser moving onto the next page.

### 4. Discussion

In this paper we have evaluated the Semantic Fields models’ ability to estimate the location of 49 participants’ eye move-
ments during goal-oriented search tasks on three website. It was found that all corpus-based Semantic Fields models performed better than other models that depended solely on display characteristics. Particularly encouraging was finding that the Semantic Fields models outperformed No-Model condition. The No-Model condition is essentially the Semantic Fields model with textual similarity estimates held constant at one. Incorporating the same decay function as the Semantic Fields model, the No-Model condition produces heat solely based on the location of Web page elements. That the Semantic Fields model outperforms the No-Model condition indicates that the semantic component of the Semantic Fields model is providing more to the fit of this model to the eye-tracking data than can be produced by the display component alone. While the Semantic Fields models provided the best fit to the human eye-tracking data in this study, there were some interesting performance differences between each of the corpus-based Semantic Fields models. These performance differences were introduced by the manipulation of both corpora and semantic models used by the Semantic Fields model.

LSA has been the focus of much of the statistical lexical semantics research in recent years. That LSA has successfully been used to grade essays (Foltz et al., 1999) is a testament to the overall usefulness of this model. It was therefore surprising that a much simpler model like Vectorspace, would consistently out-perform more complex models such as LSA and SpNMF when generating similarity comparison of text in this study. It is also interesting to note that Vectorspace is the first step in the calculation of both LSA and SpNMF, which begs the question as to whether the extra complexity introduced by these latter models when employing dimensionality reduction is of benefit when performing textual comparisons of user goals and Web page content. Furthermore, the simplicity of Vectorspace’s calculation allows for quick and efficient construction of “on-the-fly” semantic knowledge spaces that could be incorporated into more applied models of semantic salience on Web pages.

Figure 4: Semantic Field values (Vectorspace with the WEB-WIKI corpus) calculated for participant eye-points during goal-oriented Web page navigation. Time spent searching each page is delineated into deciles.
Drawing targeted corpora for the semantic models from the larger corpus of Wikipedia provided better knowledge bases for the semantic models than a more traditional corpus like TASA. The TASA corpus has been hand-picked to broadly represent the expected general knowledge of a first-year American college student (Dennis, 2007). However, the findings in this study indicate that for some semantic models (Vectorspace, LSA and SpNMF), semi-automated corpora generation using Wikipedia provides a better base to compare the similarity of textual information. That said, the generation of Wikipedia sub-spaces in this research was based on very simple Boolean queries. Greater focus on the formulation of these Lucene queries may increase the performance of semantic models when calculating text similarity and thereby conceivably produce better estimates of eye-tracking data by the Semantic Fields model.

The pupil width measure indicated that participants’ cognitive load increased as they spent more time on each Web page. Overall, most participants ended their goal-oriented search on the same pages that were initially identified by the experimenters when constructing the search tasks. Taken together, these findings support the notion that participants were actively searching out their target goals during this experiment. Given that participants were predominantly “on-task” during this experiment, it was also pleasing to find that the corpus-based Semantic Fields models were able to outperform the other models under these circumstances.

5. Conclusions

Both Web page semantics and display characteristics determine the success with which a user will be able to find information on a Web page. The Semantic Fields model incorporates both of these characteristics, and was found to provide better estimates of participants’ eye-movements during goal-oriented search than could be generated by solely display-based models. Choices of both the semantic model and knowledge base affected the performance of the semantic component that is used by the Semantic Fields model. Contrary to expectations, a relatively simple semantic model, Vectorspace, outperformed more complex semantic models that employ dimensionality reduction. Also, better approximations to the knowledge required to successfully estimate textual similarity were produced by targeted corpora drawn from Wikipedia when compared to those found using the more generic TASA corpus. Overall, the Semantic Fields model that used both Vectorspace and a targeted corpus drawn from Wikipedia, was found to be the best performing model when estimating participants’ eye movements during goal-oriented search tasks in this study.

References

Ling, J., Van Schaik, F., 2004. The effect of link format and screen location on
visual search of web pages. Ergonomics 47 (8), 907–921.

A. Goal pages with Semantic Fields maps generated using Vectorspace and WIKI-WEB

Figure 5: Mission Australia - Task 1, “Who is currently the Chief Operating Officer of Mission Australia?”

Figure 6: Mission Australia - Task 2, “You are interested in working for Mission Australia. Search their website for the current job vacancies available at Mission Australia.”
Figure 7: Mission Australia - Task 3, “You are currently researching homelessness in young people and have heard that Mission Australia has recently published a report called ‘The voices of homeless young Australians’. Search the Mission Australia website for this report into youth homelessness.”

Figure 8: Green Corps - Task 1, “You want to know more about Green Corps management. Find out who is the National Program Manager of Green Corps.”

Figure 9: Green Corps - Task 2, “Find what environmental and heritage benefits are contributed by Green Corps.”

Figure 10: Green Corps - Task 3, “Find the online Expression of Interest form to apply to become a Green Corps Partner Agency.”
Figure 11: White Lion - Task 1, “Find out who is the current President of White Lion.”

Figure 12: White Lion - Task 2, “You are interested in becoming a mentor for young people. Find out how to become one of White Lions mentors.”

Figure 13: White Lion - Task 3, “You are interested in financial viability of White Lion as a business. Find out which Government Departments are supporters of the White Lion organization.”