

# Comparative Analysis of Semantic Models and Corpora Choice when using Semantic Fields to Predict Eye Movement on Webpages

Benjamin P. Stone (bpstone@psychology.adelaide.edu.au)

School of Psychology, Level 4, Hughes Building, The University of Adelaide  
Adelaide, SA 5005 Australia

Simon J. Dennis (dennis.210@osu.edu)

Department of Psychology, Ohio State University, 1827 Neil Ave  
Columbus, OH 43210 USA

## Abstract

Nine 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 webpages. 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 saliency of different areas displayed on webpages. Latent Semantic Analysis, Sparse Non-Negative Matrix Factorization, and Vectorspace were used to generate similarity comparisons of goal and webpage text in the semantic component of the Semantic Fields model. Surprisingly, Vectorspace was consistently the best performing 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. Three other baseline models: Flat, Non-Flat, and No-Model were included as a point of comparison to evaluate the effectiveness of the Semantic Fields models. In all cases the Semantic Fields models outperformed the baseline models when predicting the participants' eye-tracking data.

**Keywords:** LSA; Vectorspace; SpNMF; Semantic Fields; web pages; eye-tracking, goal-directed visual search

## 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 focussing 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 webpage that a user will focus their attention on, they approach this task in different ways. Display-based research has focused on perceptual aspects of the webpage, components such as element and menu position, color usage, and font style. Alternatively, when attempting to predict user's webpage navigation, semantics-based research matches web user's information needs to the concepts displayed within the textual content of webpages. Display-based and semantics-based research into web user's visual search of web page hyperlinks has indicated that the user's search processes are influenced by: text semantics, element position, aesthetic qualities of elements, and environmental learning (Faraday, 2000, 2001; Ling & Van Schaik, 2002, 2004; Chi et al., 2003; Cox & Young, 2004; McCarthy, Sasse, & Riegelsberger, 2003; Pearson & Van Schaik, 2003; Pirolli & Fu, 2003; Rigutti & Gerbino, 2004; Blackmon, Ki-

tajima, & Polson, 2005; Kaur & Hornof, 2005; Pirolli, 2005; Stone & Dennis, 2007).

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

One major unanswered question when integrating display-based and semantic information is the nature of the semantic model itself, and the data upon which it should rely. As discussed previously, the computational semantics literature provides multiple options for semantic models, and similarly there are several approaches one might take in specifying a corpus relevant to the task (Stone, Dennis, & Kwanten, submitted). In this paper, we aim to shed light on these issues, by comparing different models and corpora on their ability to predict human eye movements during web browsing.

## Semantic Fields (SF)

In a previous article, we presented the LSA-SF model (Stone & Dennis, 2007), which was used to predict the eye movements of 49 participants recorded during goal-oriented search tasks on three websites. The LSA-SF model used LSA to calculate the similarity between a textual representation of the users' goal and each of the textual elements displayed on a webpage. 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 webpage (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).

$$SF(x,y) = \sum_i L_i e^{-\lambda d_{i(x,y)}} \quad (1)$$

Initially, the TASA corpus was used as a knowledge based

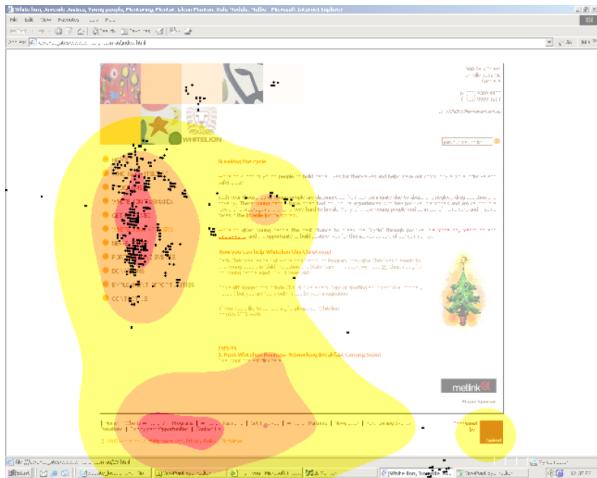


Figure 1: LSA-SF Map with participant's eye tracking data super imposed using black dots.

for the LSA component of the LSA-SF model. 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 & Markovitch, 2007; Stone et al., submitted). Also, while LSA is certainly the best studied statistical semantic model, Stone et al. (submitted) found that other models such as Vectorspace and Sparse Nonnegative Matrix Factorization (SpNMF) out-performed LSA when estimating human judgments of document 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 based (TASA and Wikipedia), when used as components in the generation of Semantic Fields.

## Method

### Human Data

Eye-movement data generated by 49 university participants on three websites was recorded during nine goal-oriented search tasks. The human eye-movement data used in this research is fully described in Stone & Dennis (2007). There are only two differences in the preparation of the data used in this study to the methods used in the original study.

Firstly, the original study used eye-tracking data recorded on 1842 webpages, in this study only 1809 webpages are used. It was found that 33 page views included in the original dataset were webpages designed by the researcher 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 websites, and their simple one element construction with black text on a white background probably favored the Semantic Fields model.

Secondly, the calibration of the eyedata was performed in a different way in this study. During the course of the participants’ search tasks, after each webpage they viewed, a screen containing nine calibration points was displayed on the monitor. The eye data recorded while the participants were viewing these calibration points was used to adjust their eye data, correcting for head movements that occurred during the experiment. In the previous study, a complex 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.

### Websites

Three websites were chosen from the Internet:

[www.greencorps.com.au](http://www.greencorps.com.au)<sup>1</sup> (Green Corps Australia)

[www.missionaustralia.com.au](http://www.missionaustralia.com.au) (Mission Australia)

[www.whitelion.asn.au](http://www.whitelion.asn.au) (White Lion Australia)

Static versions of these sites<sup>2</sup> were pre-fetched in December 2005 to avoid changes created by website updates. These websites are all similar in the type of service they provide, such that they all offer services to disadvantaged members of the community. The websites were chosen because they were sufficiently complex that searching for information on these sites would be a non-trivial task for participants.

### Models

The SEMMOD<sup>3</sup> semantic models package was used to incorporate the Vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (Kintsch, McNamara, Dennis, & Landauer, 2006), and Sparse Nonnegative Matrix Factorization (Xu, Liu, & Gong, 2003) into the Semantic Fields model.

**The Vectorspace model** (Salton, Wong & Yang, 1975): The Vectorspace model 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 or inverse document frequency (IDF) reduces the impact of high frequency words that appear in many documents in a corpus. Similarities are measured as

<sup>1</sup>The Green Corps Australia URL is no longer used, the Australian Government has change both the website and its URL, which can be viewed here: <http://www.greencorps.gov.au>

<sup>2</sup>The static versions of these websites can be found here: [http://www.psychology.adelaide.edu.au/mall\\_lab/lsa-sf\\_sites/](http://www.psychology.adelaide.edu.au/mall_lab/lsa-sf_sites/)

<sup>3</sup>The SEMMOD semantic models package is release under the GNU Licence and can be found here: [http://mall.psy.ohio-state.edu/wiki/index.php/Semantic\\_Models\\_Package\\_%28SEMMOD%29](http://mall.psy.ohio-state.edu/wiki/index.php/Semantic_Models_Package_%28SEMMOD%29)

the cosines between the resultant vectors for different documents.

**Latent Semantic Analysis** (LSA, Kintsch, McNamara, Dennis, & Landauer, 2006): LSA 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 vector and then added. As with the vector space model, the cosine is used to determine similarity.

**Sparse Nonnegative Matrix Factorization** (SpNMF, Xu, Liu, & Gong, 2003): Nonnegative Matrix Factorization 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. While 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 as a measure of similarity.

## Corpora

**TASA** The Touchstone Applied Science Associates (TASA) corpus was constructed to represent the reading material covered by American students up to first year college. Moreover, the 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, Royer, & Pirolli, 2007).

**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 current to March 2007, however the document number was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. The in-

complete articles removed were identified if they contained the phrases like “help Wikipedia expanding” or “incomplete stub”.

To enable the creation of sub-space corpora, Lucene<sup>4</sup> (a high performance text search engine) was used to index each document in the Wikipedia corpora. Lucene allows the user to retrieve documents based on customized boolean queries, and wildcard operators like ‘the star’ (\*) which return multiple results from word stems. Like the more well known search engine Google, the documents returned by Lucene are ordered by relevance to a query.

As mentioned above, three search tasks were given to participants for each website. A sub-space of 1000 documents was created for each website using simple Lucene queries that contained keywords from the three tasks set to a specific website. For example, a textual representation of the tasks given to participants for the Green Corps website are:

- 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.

Using keywords from these tasks the following Lucene query was used to create the 1000 document subspace for the Green Corps website:

(“green corps”) OR (“national” AND “program” AND “manager”) OR (“environment\*” AND “benefit\*”) OR (“heritage” AND “benefit\*”) OR (“partner” AND “agency”)

**Appending Webpages as Documents - The creation of TASA-WEB and WIKI-WEB corpora.** When creating corpora for each of the three websites in this study, the textual content of each of the webpages viewed by participants on that website were appended to the TASA and Wikipedia sub-corpora. For example, when creating a corpus for the Mission Australia website, overall 57 unique webpages were viewed by participants during the experiment. So, the textual content from these webpages was used to construct a mini-corpus of 57 documents, which was then appended to both the TASA corpus and Wikipedia sub-spaces for Mission Australia corpora<sup>5</sup>. Highlighting this appended data, the naming convention TASA-WEB and WIKI-WEB is used through out this paper.

## Baseline models to estimate eye-position

Three alternative models were designed as a baseline 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 websites. These baseline

<sup>4</sup>PyLucene, is a Python extension that allows access to the Java version of Lucene: <http://pylucene.osafoundation.org/>

<sup>5</sup>Note: Green Corps and White Lion websites’ textual data was not used when creating corpora for the Mission Australia data.

models are the Flat, Non-Flat, and No-Model.

**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 webpage. Given the total number of eye-points (EP), and the total number of pixels that an eye-point ( $p$ ) could be located in ( $1280 \times 1024$ ), for each page ( $i$ ) that is viewed ( $V$ ) by participants, the Flat model calculates the log-likelihood of the eye-points on any given webpage as  $LL_{webpage}$  (see Equation 2). The sum of these webpage log-likelihoods ( $LL_{FLAT}$ ) calculates the fit of the Flat Model to all eye-points recorded for participants.

$$LL_{FLAT} = \sum_{i \in V} LL_{webpage_i}$$

$$LL_{webpage} = \sum_{p \in EP} \log \left( \frac{1}{1280 \times 1024} \right) \quad (2)$$

**Non-Flat Model** The Non-Flat model is similar to the Flat model, except it gives more weight to the probability estimates for those eye-points found in textual elements<sup>6</sup>. The Non-Flat model is displayed in Equation 3, where for each webpage ( $i$ ) 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 3). The maximized log-likelihood ( $ML$ ) over all webpages 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 webpage than focusing on other areas, moreover the calculation of log-likelihoods for eye-points in these textual areas have been assigned greater weight in accordance with this finding.

$$ML_{NONFLAT} = \sum_{i \in V} ML_{webpage_i}$$

$$ML_{webpage} = A \log \left( \frac{\hat{w}}{\hat{w}N + M} \right) + B \log \left( \frac{1}{\hat{w}N + M} \right) \quad (3)$$

**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 webpage, 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 ( $L_i = 1.0$ ) in the calculation of the Semantic Fields (see Equation 1).

<sup>6</sup>An example of text elements (including images that have a textual description) on a webpage are shaded in red in Figure 2

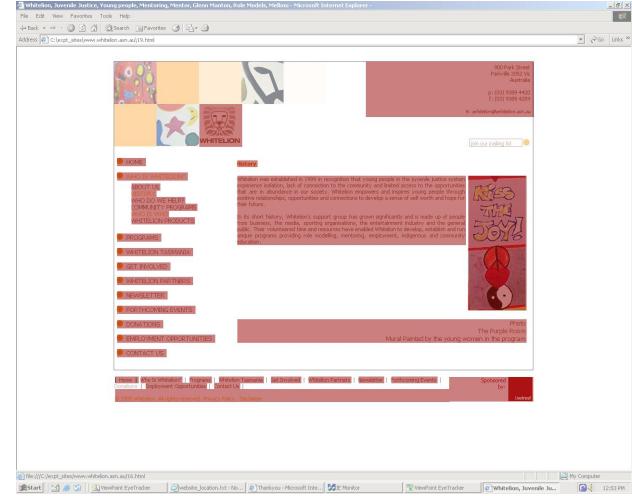


Figure 2: Textual webpage elements are highlighted in red, images that have “ALT” or descriptive text are included.

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

$$LL_{webpage} = \sum_{p \in EP} \log \left( \frac{SF_p}{SF_{webpage}} \right) \quad (4)$$

**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<sup>7</sup> ( $SF_{(x,y)}$ , see Equation 1) for each eyepoint ( $p$ ) is divided by the summed total of the Semantic Field Values for all pixels that webpage ( $SF_{webpage}$ ). 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 webpage viewed by a participant. This process of summing the log-likelihoods is repeated for each webpage ( $i$ ) 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 4).

## Results

The log-likelihoods (LL) of nine models constructed to predict participants eye movements are compared in this section using the Bayesian Information Criterion (BIC, Schwartz, 1978) for assessing model fit. The nine models include six Semantic Fields (SF) models, 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 SF models. Therefore,

<sup>7</sup>No-Model holds the semantic model coefficient constant at 1.0, so the value it returns for a pixel can still be thought of as a Semantic Field Value

for the SF models, there is a two (corpora) by three (semantic model) experimental design. The other three models compared here are the Flat, Non-Flat, and Non-Model conditions. While the calculation of most model LL only takes one parameter, the maximized log-likelihood (ML) calculated for the Non-Flat model has two. The BIC is the most appropriate method of comparing the fit of these LL to the eye-data, because it adjusts for the number of parameters going into the model. Moreover, higher BIC score indicate a better fitting model to the data.

Table 1: Comparison of Bayesian Information Criteria (BIC) statistics calculated from LL generated for all nine models.

Corpora	Model	BIC	Difference
WIKI-WEB	Vectorspace	-10983963	0
WIKI-WEB	SpNMF	-10989551	-5588
WIKI-WEB	LSA	-10990633	-6670
TASA-WEB	Vectorspace	-10993390	-9427
TASA-WEB	SpNMF	-11001041	-17078
TASA-WEB	LSA	-11001286	-17323
*	No-Model	-11005111	-21148
*	Non-Flat	-11287790	-303827
*	Flat	-11417562	-443599

\* No corpus used

The results displayed in Table 1 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 Semantic Fields models than the other models, therefore the Semantic Field 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 webpage 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.

## Discussion

While the Semantic Fields models provided the best fit to the human eye-tracking data in this study, there were some interesting performance differences within each of the Semantic Fields models that were introduced by the manipulation of corpora and semantic models. Corpora like the TASA have been hand-picked to broadly represent the expected general knowledge of a first-year American college student. However, the findings in this study indicate that for some semantic models (LSA, Vectorspace 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.

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, Laham, & Landauer, 1999) is 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 dimension reduction is of benefit when performing textual comparisons of user goals and webpage 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 saliency on webpages.

As outlined previously, incorporating additional components such as colour and contrast information may improve the performance of the Semantic Fields model (Stone & Dennis, 2007). Another possible approach would be to use the Explicit Semantic Analysis (ESA) model (Gabrilovich & Markovitch, 2007) to make textual comparisons in the Semantic Fields model. The ESA model also uses Wikipedia as a knowledge base and has performed very well on document comparison tasks. Using human judgments of the document similarity of 50 headlines and précis taken from the ABC online news mail service that were collected by Lee, Pincombe, and Welsh (2005), Gabrilovich and Markovitch (2007) report that ESA has generated estimates of document similarity that correlate strongly (0.72) with human performance on this task. Given the ESA's strong performance at this task, in future research on the Semantic Fields model may use the ESA to generate estimates of similarity between webuser's goals and textual elements contained on webpages.

## Summary

Nine models were used to try and predict the eye-movements of 49 participants who were performing goal-oriented search tasks on three websites. All six Semantic Fields models produced a better fit than three baseline models at this task. Both choice of corpora and semantic model produced main effects when estimating the human eye-tracking data. Such that, the corpora drawn from Wikipedia out-performed the TASA, and Vectorspace consistently outperformed both SpNMF and LSA. Particularly encouraging was finding that the Semantic Fields models outperformed the No-Model condition, indicating 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.

### Acknowledgements

Many thanks to Dan Navarro for taking the time to read through this paper and for his excellent suggestions.

### References

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In *Chi '05: Proceedings of the sigchi conference on human factors in computing systems* (pp. 31–40). New York, NY, USA: ACM.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. In *Chi '02: Proceedings of the sigchi conference on human factors in computing systems* (pp. 463–470). New York, NY, USA: ACM.

Budiu, R., Royer, C., & Pirolli, P. L. (2007). Modeling information scent: a comparison of lsa, pmi and glsa similarity measures on common tests and corpora. In *8th riao conference*. Pittsburgh, PA.

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., et al. (2003). The bloodhound project: automating discovery of web usability issues using the infoscent $\pi$  simulator. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 505–512). New York, NY, USA: ACM.

Cox, A., & Young, R. M. (2004). A rational model of the effect of information scent on the exploration of menus. In *Proceedings of the sixth international conference on cognitive modelling* (pp. 82–87). Mahwah, NJ: Lawrence Erlbaum Associates.

Faraday, P. (2000). Visually critiquing web pages. In *Proceedings of the 6th conference on human factors and the web*.

Faraday, P. (2001). Attending to web pages. In *Chi '01: Chi '01 extended abstracts on human factors in computing systems* (pp. 159–160). New York, NY, USA: ACM.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 6–12). Menlo Park, CA: AAAI Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438, 900–901.

Kaur, I., & Hornof, A. J. (2005). A comparison of lsa, wordnet and pmi-ir for predicting user click behavior. In *Chi '05: Proceedings of the sigchi conference on human factors in computing systems* (pp. 51–60). New York, NY, USA: ACM.

Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (2006). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation on models of text document similarity. In *Proceedings of the 27th annual conference of the cognitive society* (pp. 1254–1259). Mahwah, NJ: Lawrence Erlbaum Associates.

Ling, J., & Van Schaik, P. (2002). The effect of text and background color on visual searches of web pages. *Displays*, 23, 223–230.

Ling, J., & Van Schaik, P. (2004). The effect of link format and screen location on visual search of web pages. *Ergonomics*, 47, 907–92.

McCarthy, J., Sasse, M. A., & Rigelsberger, J. (2003). Could i have the menu please? an eye tracking study of design conventions. In *Proceedings of hci 2003* (pp. 401–414). London, UK: Springer-Verlag.

Pearson, R., & Van Schaik, P. (2003). The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. *International Journal of Human-Computer Studies*, 59, 327–353.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, 29, 343–373.

Pirolli, P., & Fu, W. (2003). Snif-act: A model of information foraging on the world wide web. In *9th international conference on user modeling* (pp. 45–54). Johnstown, PA: Springer-Verlag.

Rigutti, S., & Gerbino, W. (2004). Navigating within a web site: the webstep model. In *Proceedings of the sixth international conference on cognitive modeling* (pp. 378–379). Mahwah, NJ: Lawrence Erlbaum Associates.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613–620.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on machine learning* (pp. 792–799). New York, NY.

Stone, B., & Dennis, S. (2007). Using lsa semantic fields to predict eye movement on web pages. In *Proceedings of the 29th annual conference of the cognitive society* (pp. 665–670). Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, B., Dennis, S., & Kwanten, P. J. (submitted). *Comparing methods for document similarity analysis*.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international acm sigir conference on research and development in information retrieval* (pp. 267–273). New York, NY: ACM Press.