A Comparison of LSA, WordNet and PMI-IR for Predicting User Click Behavior

Ishwinder Kaur and Anthony J. Hornof
Computer and Information Science
University of Oregon
Eugene, OR 97403, USA
{ishakaur, hornof}@cs.uoregon.edu

Abstract

A predictive tool to simulate human visual search behavior would help interface designers inform and validate their design. Such a tool would benefit from a semantic component that would help predict search behavior even in the absence of exact textual matches between goal and target. This paper discusses a comparison of three semantic systems—LSA, WordNet and PMI-IR—to evaluate their performance in predicting the link that people would select given an information goal and a webpage. PMI-IR best predicted human performance as observed in a user study.

Categories & Subject Descriptors: H.5.2 User Interfaces – evaluation/methodology, graphical user interfaces (GUI), screen design; H.5.4 Hypertext/Hypermedia – Navigation; H.1.2 User/Machine Systems – Human information processing; H.1.1 Systems and Information Theory – Information theory, value of information

General Terms: Design, Experimentation, Human Factors, Verification

Keywords: PMI, LSA, WordNet, semantic relatedness, semantic similarity, computational linguistics

INTRODUCTION

Web users continually apply their semantic knowledge while searching the web. For example, finding "the new bestselling textbook published by a psychology professor" in the website shown in Figure 1 is a web task that taxes semantic knowledge.

Visual interfaces are the predominant method of conveying information during human computer interaction (HCI). Predicting human behavior in tasks involving their use would be useful for the designers of the interfaces. Efforts have been made in the usability simulation and analysis of visual displays. Display Analysis Program (DAP [36]) predicts the mean search time of an alphanumeric display based on the spatial relationship of objects in a layout. UCIE (Understanding Cognitive Information Engineering [23]) evaluates the usability of graphs and tables by predicting the search times and optimal eye movements of the users given an information goal. Though these tools are useful only in specific domains, they are an important first step in using predictive tools to simulate human responses, hence informing the design process.

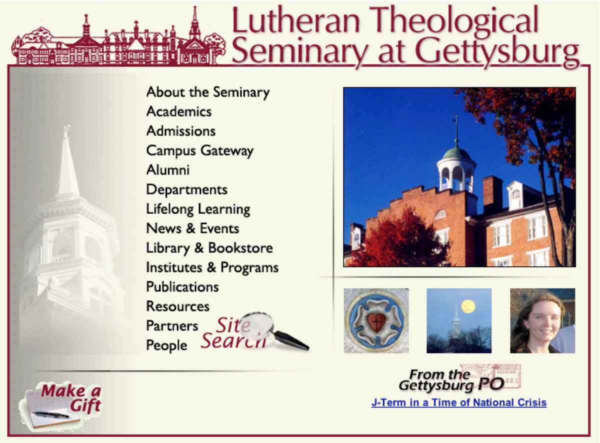


Figure 1. Where would you click to find information about "the new bestselling textbook published by a psychology professor"? (http://www.ltsg.edu, January 26th 04)

The research discussed in this paper is part of a larger effort to build a comprehensive tool for predicting visual search behavior, a tool that user interface designers could use early in a design process to evaluate screen layouts. The tool will be based on cognitive models of visual search, such as those of Byrne [6] and Hornof [14]. All of these models assume that the user knows the exact text of the target. However, exact matches between information goals and link labels are perhaps rare in real world tasks. The vocabulary problem, for example, illustrates that for a single concept there can be no single word or description (however well chosen) that will be used by all people [13]. Semantics affect how visual interfaces are searched [4, 28] and models of visual search will ultimately need to be enhanced with a *semantic system* to predict when users will terminate a search based on an inexact match. The semantic system would assign a quantitative value to the similarity in meaning between words and phrases.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
Copyright 2005 ACM 1-58113-998-5/05/0004...$5.00.

Some models of broader user navigational behavior (such as WUFIS [9]) that guide research and design of web-based information systems and site architectures do incorporate semantics. WUFIS simulates web users with specific information goals, and uses an "information scent" based approach to semantically associate the information on the web page to the information need of the user. Such systems would benefit from an evaluation of a variety of candidate semantic systems.

A number of tools that evaluate some parameters of web usability, such as ease of navigation or lack of confusability, have also incorporated semantics in some form [3, 10]. It would be useful in such cases to have empirical validation for the choice of a semantic system in the context of HCI tasks.

Systems that attempt to model web-surfing behavior, but do not yet have semantic behavior incorporated (for example, the Max Model [25]), would especially benefit from the work presented in this paper, as the rules guiding navigational decisions could then be based on the semantic content of the website along with its structure.

Some systems monitor where users click in real-time to dynamically tailor the information environment and to assist the user by trying to infer their information goal (such as IUNIS [9] and ScentTrails [27]). One such assistive system, Letizia [20], models the content of a document as a list of keywords. The author of Letizia specifically states that "natural language capabilities that can extract some grammatical and semantic information quickly… could greatly improve its [Letizia's] accuracy," [20] suggesting the need for a semantic subsystem in the process of user modeling and behavior prediction.

We would like to evaluate existing semantic systems that could be plugged in and be the semantic component of the tools in question. Given a web page and an information goal, the ideal system would choose the same link label that people would. To promote this engineering approach to usability analysis, the various systems evaluated in the paper are used, as best as possible, "off the shelf" and without enhancements, modifications or extensions.

This paper is organized as follows: First we discuss and compare available semantic systems. Then we examine how some of those semantic systems are already used in usability evaluation tools. We then present a survey conducted for independent comparison of three of the systems in the context of HCI and discuss our methods for applying the three systems to predict the human data. Finally, we compare the predictions of the three systems.

SEMANTIC SYSTEMS

There are a number of approaches for computing semantic similarity that might be useful in a predictive tool. These systems are typically discussed in the context of computational linguistics. The approaches can broadly be classified as follows:

1. Statistical: Semantic relationships between terms are captured from the probability of their co-occurrence in a text *corpus*, which is a large collection of documents. Some of these statistical systems are based on a vector space representation of a language.
2. Taxonomical: The contents and relations of a hierarchy of terms are used to derive a quantitative value of similarity between terms. The terms are organized by subsumption with more general concepts subsuming the more specific ones (for example, *automobile* subsumes *car*). Other relations are also defined (for example, *steering wheel* is a part of a *car*).
3. Hybrid: This approach borrows from both of the above approaches. It is based on a taxonomical representation of concepts enhanced by the statistical properties of a text corpus.

Statistical

Statistical approaches to semantic relatedness that are relevant to the context of predictive tools include Latent Semantic Analysis (LSA [12]), Hyperspace Analogue to Language (HAL [24]), Point-wise Mutual Information using Information Retrieval (PMI-IR [37]) and Non Latent Similarity (NLS [7]).

LSA and HAL each process a large textual corpus to create a multi-dimensional semantic space. As the corpora used to build the semantic space can be categorized according to cultural, regional or educational variations, the corpus selection can be used to classify the resulting semantic spaces.

In HAL and LSA, terms are represented by vectors in a semantic space. Term vectors that are closer together in a multi-dimensional semantic space are deemed to have a higher semantic relatedness. Semantic relatedness can include similarity (full or partial synonymy), meronymy (between a part and the whole), hypernymy or IS-A relationships (*dog* IS-A *animal*), and cause-effect relationships. The distinction between similarity and relatedness is important. For example, *dog* and *collar* co-occur in many texts, but ideally a predictive system should determine that they are related but not similar.

The statistical approaches that use semantic spaces have been shown to reflect human performance quite accurately [16]. LSA reflects human acquisition of semantics as it performs comparably to humans in vocabulary tests, subject matter scores, word-sorting tasks, category judgments and lexical priming tasks [17].

Purely statistical approaches will have some inherent potential limitations: First, the type of semantic relatedness is not captured but might be relevant to a predictive tool (e.g., the IS-A relationship might be most useful on websites that hierarchically categorize products like Amazon.com). Second, different "senses" of words are treated as one (e.g., a *bank* could refer to a financial institution or the edge of a river). Third, different word forms are treated as separate words (such as *condition*, *conditional* and *conditioning*). Lastly, part of speech is not taken into account (e.g., *dog* can be a verb and *purchase* can be a noun).

PMI-IR is a statistical approach but it is not based on the concept of semantic spaces. PMI-IR uses the results of information retrieval to compute relatedness between words or phrases in terms of their *mutual information*; that is, the degree of shared content as measured by the probability of co-occurrence versus independent occurrence of terms [11]. PMI has produced results 10% better than LSA on synonym tests [37]. PMI-IR has also out-performed other computational methods for quantifying similarity [34].

A number of other systems have also attempted to improve on LSA's basic approach. NLS, for example, uses a corpus to parse and derive relationships between all word pairs. NLS captures syntactic relationships between the terms, over and above simple co-occurrence. The first order matrix of word relationships is squared to capture transitive, second-order relationships between terms. For example, as *glass* and *heart* are both related to *break*, it implies that *glass* and *heart* have something in common. In a test to predict word associates, NLS performed comparably to LSA for verbs and modifiers, and better than LSA for nouns [7].

Taxonomical

Taxonomical relations are used to derive a number of semantic similarity and relatedness measures. Perhaps the best-known word taxonomy is WordNet [26], which is essentially a lexical database that encodes the relationships of synonymy, hypernymy and meronymy between terms. CYC [19] is also a large knowledge base and common-sense reasoning engine with a taxonomy of concepts and other relations such as causality.

Hand-coded lexical knowledge databases such as WordNet and CYC make it is possible to encode large amounts of otherwise inaccessible human knowledge in a computational format. However, the knowledge acquisition is tedious, subject to the vagaries of human judgement, and not easily scalable to new terms, domains and languages.

Since WordNet is simply a taxonomy, some kind of a system is needed to produce a similarity rating between two words in WordNet. Even though the lengths of links between the various nodes in WordNet are "subjective and vary widely" [Miller, G. A., personal communication], a number of different measures have been developed to compute a quantitative value of semantic similarity between terms in the WordNet taxonomy. **wup** [39] is a node-based measure that computes semantic similarity between terms based on the number of nodes between the terms, their lowest common subsumer, and the root. **lesk** [2] is a measure of semantic relatedness based on word overlap between term definitions, hypernyms, and hyponyms. **path** simply counts the number of edges between two terms.

Another taxonomical system is NetSerf [8]. NetSerf uses a combination of semantic relations from WordNet and relationships extracted automatically from an online Webster's dictionary (mainly IS-A relationships). NetSerf was developed to locate the web archive relevant to a query. NetSerf achieved this by locating the archive description that semantically subsumes the query, such as the archive of *electronics* when looking for *digital cameras*. Results show improvement over unstructured and non-semantic retrieval.

Hybrid

Hybrid systems include **res**, **jcn**, **lin** and **lch**, all of which are based on the WordNet taxonomy, but also use a corpus for calculating similarity and relatedness values.

res [30] is a WordNet based measure that uses probability-based information content and has been demonstrated to perform better than **path** at predicting human similarity judgements between terms. The similarity between two terms is defined as the information content of their most specific subsumer (calculated by taking the negative logarithm of the probability of its occurrence). One limitation of such an approach is that all terms within a branch will have the same similarity value with another term as long as the subsumer remains the same. For example, *dog* and *cat* have the same similarity value as *dog* and *panther* because both pairs have the same subsumer, *carnivore*.

lin [21] overcomes this limitation by taking the information content of the participating terms into account. The similarity is then defined as the information content of the most specific subsumer divided by the information content of the two terms. **jcn** [15] is based on lexical edge counting, like **path.** Unlike **path**, the edges are weighted by frequency-based information content. The edge strength is determined by comparing the conditional probability of the child node given that the parent exists. The WordNet taxonomy has a non-uniform network density and highly variable node depth, and this measure tries to counter the effects of these problems. **lch** [18] combines semantic information from WordNet with local syntactic information to expand the size of sparse training data.

The next section discusses how two semantic systems have been directly applied to predicting the usability of human computer interfaces.

SEMANTIC SYSTEMS IN USABILITY TOOLS

This section discusses user interface analysis tools that make use of some of the previously discussed semantic systems to predict usability of web interfaces. These interface analysis tools would likely benefit from a comparison of the underlying semantic systems, as this paper aims to provide.

Bloodhound [10] is a usability tool that relies on an underlying semantic system. Bloodhound automatically analyzes the navigability of a website by building on the theory of information foraging [29]. Information foraging is based on the concept of information scent, which is the perceived utility of an information source and is related to the degree of similarity between (a) proximal cues afforded by the distal information source and (b) the information goal. Bloodhound computes this information scent based on the spreading activation using the probabilities of contextual occurrences of terms in a corpus. The information scent estimate is very similar in theory to PMI, discussed earlier.

The approach used in Bloodhound has limitations. Since the web navigation simulator built into Bloodhound is sensitive to the choice of task query keywords, an analyst is required to choose the "correct" keywords to get the system to work in the desired manner. Also, the corpus used by the semantic system is typically the content of the very web site being tested. This choice of corpus might be appropriate for predicting user navigational behavior if the semantic structure of the website is consistent with the expectations of the users. But since the site is still being evaluated, the corpus based on the website could possibly be more representative of the designer's rather than the users' semantic expectations. A general-purpose or domain-specific corpus that captures the semantic acquisition of the target population and that is independent of the search documents may be more useful.

The Cognitive Walkthrough for the Web (CWW [3]) builds on the Cognitive Walkthrough technique to evaluate the usability of a website. CWW uses LSA to compute the information scent for the links on a webpage in comparison to the user goal. This methodology has been used to find and repair usability problems with webpage designs, such as the unfamiliarity or confusability of link labels and group headings.

A potential problem with using LSA for predicting usability is that, because LSA has a tendency to overestimate similarity, some of the predicted problems need to be subjectively discounted. Also, the use of the thresholds in CWW (e.g. LSA term vector length of less than 0.8 for a section heading implies that it is unfamiliar and a cosine of more than 0.6 between two headings implies that they are confusable) would benefit from more empirical validation. These limitations might be overcome by using a semantic system that is demonstrated to predict what a user would select, not just what would be confusing. This paper attempts to identify such a system.

The usability tools discussed in this section use statistical techniques for semantic input. Apart from these techniques, other semantic systems such as NLS, PMI, HAL, and WordNet-based measures have been studied in various experiments, especially in the field of information retrieval, to see how they compare to human ratings of similarity, synonymy, antonymy, analogy relationships and detection of malapropisms. Researchers have used these systems to quantify semantic similarity in a manner that is consistent with the human view to a lesser or greater extent but, quite surprisingly, they have not to our knowledge been compared in the context of predicting user behavior in HCI.

SURVEY OF USER BEHAVIOUR

We wanted to test a number of semantic systems to compare their performance in predicting user selection behavior. To collect human data that could be used to compare these systems, we conducted a survey in which we asked participants what link they would select on a webpage given an information goal. For example, we asked them where they would click to find a publication from a professor in the Psychology department on the webpage in Figure 2.

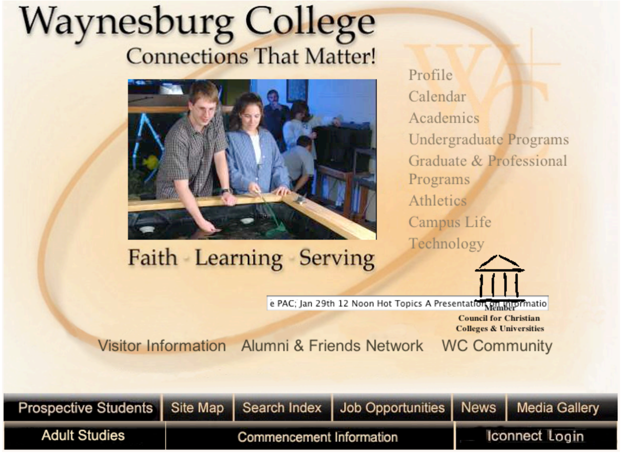


Figure 2. A sample webpage (http://www.waynesburg.edu/ Jan 26th 04)

The data collected in this survey were then used to determine (a) if people tend to click the same links given the same information goal on the same web page, and (b) which semantic system best predicts their choice.

Participants

Twenty-four native English speakers participated in the survey as paid participants. The sample included 12 female and 12 male participants, aged 18 to 56 years (median 20 years) and having 4 to 12 years (median 8 years) of experience using the web. Each session took approximately 50 minutes and each participant was paid $10.

Apparatus

Participants were given paper printouts of the home pages of 45 colleges and universities randomly selected from a list of nearly 2,000 [22]. We collected the 45 goal statements from an email survey of students and alumni who access college and university webpages daily. The query posed in the email survey was "if you were searching a college or a university webpage, what would you be trying to find?"

Our choice of webpages was constrained by their size so that the whole screenshot would fit on single sheet of paper. We wanted to stress ecological validity in our experiment and hence we did not otherwise limit our selection based on the visual design features even though the design elements may affect the links selected, along with the underlying semantic structure.

Procedure

Pairing each of the 45 webpages with 3 randomly selected goal statements resulted in 135 webpage-goal combinations. These 135 combinations were divided into three sets such that each web page and each goal statement appeared only once in a set. Each of the 24 participants was given one of the three sets, resulting in 8 participants per set. Hence, each of the 135 webpage-goal combinations was processed by 8 of the 24 participants, leading to a mixed (between- and within-subjects) design.

Analysis

For each webpage-goal combination, the participants wrote down one or two link labels on the web page that they would click to find that goal. Participants' responses were studied to find *clustering*, which we define as the measure of consistency in the participant responses. That is, for each goal and web page combination, did the eight people who searched for the goal on the webpage choose the same link?

Clustering measures the extent to which multiple people would select the same link given the same goal and the same web page. It is an important measure for the purpose of evaluating the quality of prediction, because for a system to adequately predict the choice made by the user, we need to ensure that there is indeed a set of "correct" links that users pick. Figure 3 shows the responses of the participants plotted on the web pages for samples of high and low clustering webpages. It is evident that a very high agreement between participants exists in the first case, whereas in the second case, the responses are less clustered.

Each of the participants who searched for a specific goal on a particular webpage had effectively one vote, which they could either choose to give to a single link, or split equally between two links. The measure of consistency, or the clustering C, for the webpage-goal pair is defined as the normalized root mean square (NRMS) vote:

$$C = \frac{\sqrt{\dfrac{\sum_{i=1}^{n} x_i^{2}}{n}}}{\sum_{i=1}^{n} x_i} \qquad (1)$$

In Equation 1, x_i refers to the total, non-zero votes given to the i^{th} link that appears on the list of n competing links answered by the participants. As an example, for the goal statement G2 in Figure 3(a), all the participants chose the same response, hence x_i are $\{x_1 = 8\}$ whereas for the goal G2 in figure 3(b), the eight votes are split between five different links, and hence x_i are $\{x_1 = 3.5, x_2 = 2.5, x_3 = 1, x_4 = 0.5, x_5 = 0.5\}$. Table 1 shows the clustering measures for the goals G1, G2 and G3 in Figures 3(a) and 3(b).

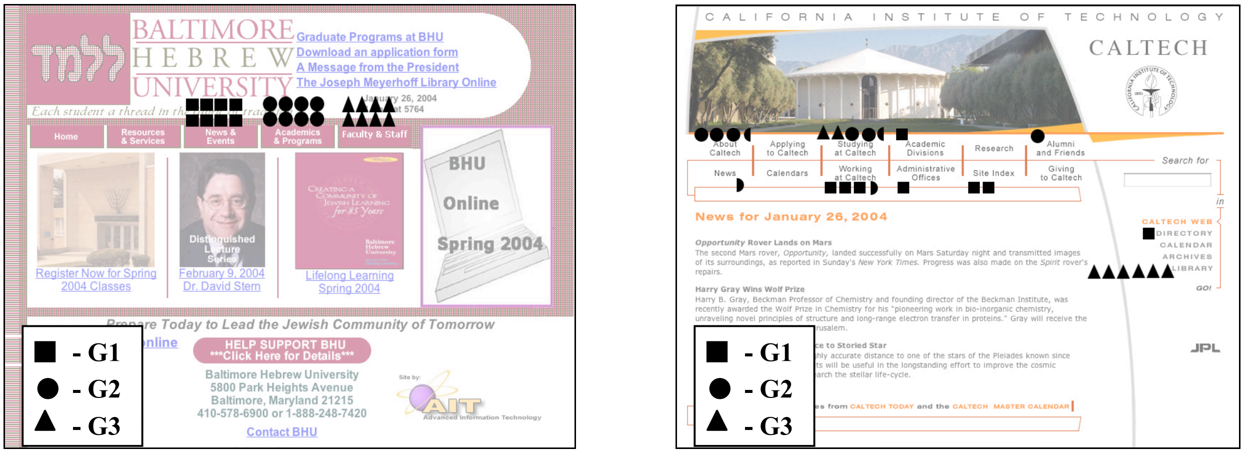
Table 1: The clustering measures of the responses to webpage-goal combinations plotted in Figure 3

Webpage Goal	Figure 3(a)	Figure 3(b)
G1	1.000	0.223
G2	1.000	0.250
G3	1.000	0.559

Survey Results

For the 135 webpage-goal combinations, the clustering measure showed a mean value of 0.49 with a standard deviation of 0.22. Note, that the value of 0.5 (or greater) on the clustering measure implies that the users agree on two links or fewer as their choice for the given goal statement.

The survey data show that people do tend to cluster at a link when looking for the same piece of information on the same web page. Figure 3 shows the user data for the highest and lowest clustering webpage-goal combinations.



(a)

(b)

Figure 3: Participant responses spatially plotted for the (a) highest clustering and (b) lowest clustering webpage-goal combinations. The different data points refer to different goal statements. For example, the square shapes plot the responses of the eight participants who pursued goal G1 on each webpage. The half moons indicate split votes.

It would be useful to identify a semantic system that could predict where this clustering will occur so that the system could be built into tools used by web designers and analysts. The clustering data will now be used to evaluate candidate systems for such applications.

APPLYING THE SYSTEMS TO PREDICT USER BEHAVIOR

This section explains how the systems were used to predict the user behavior reported in the previous section. The goal is to determine which system most accurately predicts the human link selection behavior. WordNet (**jcn**, **path**, **wup**, **lch**, **lin**, **res**), LSA and PMI were all applied to predict the links selected by humans. The study attempts to evaluate these three semantic systems, as best as possible, off the shelf, with as little tweaking or enhancement as possible.

WordNet: Since the majority of the entries in WordNet are single words and not phrases (only 0.74% of the words in all the goal statements and link labels formed a part of the compound words found in WordNet), the measures based on WordNet do not directly return a quantitative value of similarity/relatedness between two phrases. On the other hand, all of the goal statements and many of the link labels in our survey are phrases, sometimes even complete sentences.

There are a number of different approaches that could be used to calculate phrase similarity using the WordNet based measures. We considered three approaches: (a) Ambroziak and Woods' [1] approach of organizing phrases in a hierarchy using generality and subsumption, (b) Richardson and Smeaton's [31] approach of averaging over all the word similarity values in a phrase, and (c) Sutcliffe et al.'s [33] technique of finding the single best match for a word from each phrase. We used approach (c) because, unlike (a), the mechanism was defined in sufficient detail for independent implementation and unlike (b), it is not biased against longer link labels.

WordNet measures were computed using the WordNet::Similarity package [38]. Recall that a few of the measures use a text corpus for computing frequency of occurrence and the related information content of terms. A potential problem with using a text corpus is that polysemous words (with multiple unrelated meanings) might get exaggerated frequency counts, thus skewing predictions. To help combat this problem, all WordNet measures use a corpus called SemCor that is hand tagged with WordNet senses [32] and is included in the standard distribution of the WordNet similarity package.

LSA: LSA measures were computed using the University of Colorado, Boulder's AutoCWW server. The first semantic space used was based on the generic TASA corpus containing 37,651 documents from general texts up to first year of college. Each paragraph in the text was treated as a separate document.

LSA provides a built-in method of computing the similarity between word phrases based on the underlying vector space model. Through empirical analysis, this approach was found to be superior to the indirect approach of computing phrase similarity from word similarity (as described previously in the section on WordNet). Hence the indirect approach was not used in the final evaluation of LSA.

To explore how the choice and size of corpus affects the results from LSA, we also built our own domain-relevant semantic space using articles from the *Chronicle of Higher Education* [35] as our corpus. We created these additional semantic spaces for LSA in part because it is relatively easy to do so, thus showcasing an off-the-shelf feature of LSA. Two versions of the *Chronicle* corpora were created and used. One version contains 112 articles, (3,990 "documents"). The other set used 363 articles (12,669 "documents").

PMI-IR: PMI-IR similarities were obtained using the Waterloo Multi Text System (WMTS), which is a distributed information retrieval system with one terabyte of corpus data. Turney's original study concerning PMI-IR used the AltaVista search engine to pose queries to the web in real time, but in the current study, a static corpus including approximately 77 million web pages collected by a web crawler in 2001 has been used.

We used the PMI-IR system in two versions. The first version, **PMI-w**, computed word similarity for each of the words in the goal statement and the link label. It then used the phrase similarity measure discussed for the WordNet measures to compute effective similarities.

The PMI window of reference of size 20 was used for the first version of the system (20 was the default window size in Turney's 2001 experiment). The second version, **PMI-p**, computed similarity using exact string comparison of the goal statement and the link labels. Since the probability of occurrence of exact phrase strings is less than that of individual words, the window of reference in this case was chosen to be the whole document.

LSA and PMI-IR were thus tested with different corpora. Even though it would be interesting to compare the two systems using the same underlying domain-specific corpus, the point of this research is not so much to evaluate the underlying theories, but rather existing systems built on those theories, and it is less straightforward to train PMI-IR on a new corpus than it is LSA.

SYSTEM PREDICTIONS

This section compares the predictions of the various systems. For each of the 135 webpage-goal pairs, participants in the survey responded with a set of links they would click on. These human preferences are taken to be *target responses*. The target responses are weighted by the number of votes received by each of them in the survey. For the same webpage-goal pair, each semantic system was used to rank the link labels from most similar to least similar; the top-ranked labels are taken as the *system responses*. The overlap between the target responses and the system responses are termed the *correct predictions*.

A varying *window size* is used when comparing the system responses to the target responses to determine the correct predictions. A window size of *N* represents that the top *N* system responses are considered in the comparison, and the remaining items in the ranked list are ignored. If any of the top *N* responses match any of the target responses, the system is said to have made a correct prediction.

Humans often respond with more than one link for an information goal on a web page, this being evident from that fact that survey respondents collectively chose 3 links or more on 59% of the webpage-goal combinations. To allow the system to make a prediction of actual human behavior (multiple links), rather than just the most popular link selected, it is important to consider a variable window size instead of just choosing the top link label. Ideally, we would like the system to come up with more correct predictions towards the top of the ordering and we would like the performance to improve as the window size increases and more choices are considered.

We used the following three dependent variables, which were computed as a function of the semantic measure used and the window size *N*, to compare the performance of WordNet, LSA and PMI-IR:

Percent Correct is the average, across all the webpage-goal pairs, of the votes received by correct predictions as a percentage of the total number of votes for the target responses (that is, achieved votes vs. maximum possible votes).

Precision is the average, across all the webpage-goal pairs, of the correct predictions divided by the system responses (that is, accuracy of prediction).

Recall is the average, across all the webpage-goal pairs, of the correct predictions divided by the target responses (that is, potency of prediction).

Figure 4 shows the percent correct for the three systems—WordNet, LSA and PMI-IR. Repeated measures MANOVA demonstrated that for all the three dependent variables the difference between measures was significant, F(10, 125) = 8.456, $p < 0.01$, as was the difference between window sizes, F(4, 131) = 58.727, $p < 0.01$.

In other words, if we plot all the six WordNet measures, the three LSA corpora, and the two PMI-IR variants on the same graph, for any of the three dependent variables, the differences between these eleven measures were statistically significant. Also, the window size has a significant effect on the performance of measures.

We compared six measures based on WordNet (**jcn**, **lin**, **lch**, **res**, **wup** and **lesk**). Figure 4(a) shows three representative measures. **jcn** performed best. **wup** performed worst. Though the difference between **jcn** and **lch**, **lin** and **wup** was statistically significant (p = 0.037, p = 0.038, p = 0.004 respectively, for the most part the difference between various WordNet measures was not significant.

Figure 4(b) plots the performance of LSA with three corpora. LSA-TASA uses the general reading corpus up to first year of college; LSA-c and LSA-C respectively comprise 112 and 363 articles from the *Chronicle of Higher Education* [35]. The difference between the three measures is not statistically significant.

Figure 4(c) shows the performance of the two variants of PMI-IR: PMI-p is the comparison between the complete text strings of the goals and the link labels. PMI-w computes the similarity indirectly from the similarity of the constituent words. The difference between measures is statistically significant ($p < 0.01$).

Measures from each of the three systems—WordNet, LSA and PMI-IR—that performed best were then compared against each other. In other words, the "top" plot in each graph in Figure 4 was compared. This comparison is shown in Figure 5(a). Figure 5(b) and 5(c) show the precision and recall, respectively, of the three best measures.

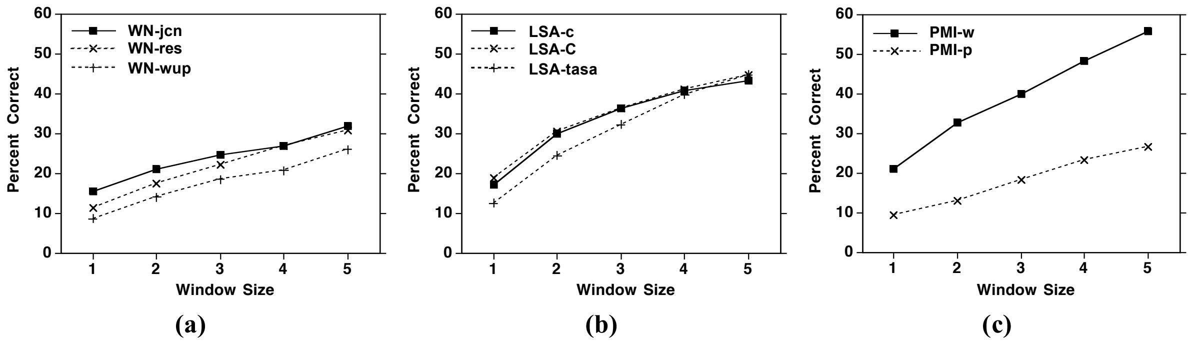


Figure 4: Percent correct as a function of window size for the three systems, including (a) three measures based on WordNet, (b) LSA with the three corpora, and (c) two variants of PMI-IR. Window size of *N* represents that the top *N* link labels selected by the system were considered

In all three graphs in Figure 5, there is a trend that PMI-w performed the best, and in all three cases *post hoc* contrast analysis showed that the difference is significant.

Figures 5(b) also shows that precision degrades with window size, which is good news because it implies that the top responses of the system have a better chance of being "correct". Figure 5(c) shows that recall improves with window size, which is also good as the increased number of responses considered represent the human selection behavior more accurately. Ideally, we would want recall to level off after a certain point, demonstrating that there is an ideal window size of use for each system beyond which precision degrades for no real improvement in recall. It can be noticed in Figure 5(c) that although LSA does have a tendency to level off, PMI-IR does not.

The percent correct and recall plots, Figure 5(a) and 5(c), look quite similar because the dependent variables are related in that both compute how accurately the system is predicting the target responses. The difference is that percent correct is weighted by the votes received by the target responses whereas recall treats them as equal. As an example, consider that for a webpage-goal combination, the target responses are A and B with 3 and 1 votes respectively. If one semantic system gives A as a response and another gives B, the recall values are same for both systems, that is 0.5 (1 out of 2 recalled), but the percent recall is 75% for the first system (correct for 3 out of 4 votes) and 25% for the second (correct for 1 out of 4 votes).

DISCUSSION

This comparative study between various semantic systems suggests that PMI-IR is best at predicting the link most people will select given a web page and an information goal. Even though the ability to work with a large dynamic corpus is one of the strengths of PMI-IR, it is possible that the use of a different, more domain- specific corpus might improve results. This conjecture may be evaluated in our future work.

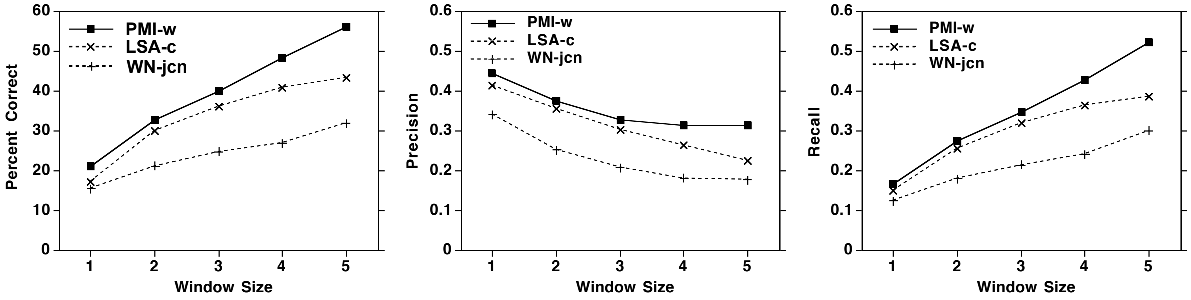
Even though we did not limit our webpages based on visual design features, and hence the design elements may have affected the participants' choice of links, the various systems still did a decent job of predicting responses. This suggests that semantics did play a large role in link selection in the survey.

Other interesting observations include that the performance of LSA improves slightly, though not significantly, with the use of a more domain specific corpus, but that increasing the size of the corpus from 3,990 to 12,669 documents (by a factor of more than 200%) does not improve prediction. It is interesting to see that a relatively modest domain-relevant corpus leads to a better performance than a larger but more general corpus, though the improvement was not statistically significant.

The WordNet measures, in spite of the human-encoded taxonomy, were not as good at predicting the human selections. The lower performance of WordNet measures could perhaps be attributed to the phrase similarity measure that was used, but the fact remains that the PMI-IR system worked best with the same similarity measure.

Within the WordNet measures, **jcn** performed best at predicting the links people would select, although the difference was not statistically significant. This finding supports the results of a previous comparative study of WordNet measures [5] in which the authors compared the performance of five measures including **res**, **lin**, **jcn**, and **lch** to detect and correct malapropisms introduced to Wall Street Journal articles. The study demonstrated that **jcn** performed best of all the measures, as it was able to report and correct the greatest number of malapropisms.

The current study, though useful and relevant, is not enough to make a categorical statement about the "best" system for predicting human click behavior. Other systems such as HAL and NLS also need to be tested, but they were not readily available at the time of the study. Also, synergistic techniques for combining semantic systems might perform better than any of the systems individually. This was demonstrated for a number of statistical systems in the context of predicting synonyms in the Test of English as a Foreign Language (TOEFL) [34].



(a) **(b)** **(c)**

Figure 5: (a) Percent correct, (b) precision and (c) recall as functions of window size for the best measures of the three semantic systems. For all dependent variables, PMI-IR performed best.

The visual search behavior of *satisficing*, in which people select the first option that is "good enough" to pursue without determining the optimal choice, might affect the link that people will select. Brumby and Howes [4] suggest that people do not look at all links on a page, but instead assess unexamined links based on sampling a subset. Perhaps the highest similarity link might be overlooked during human search. We may have reduced satisficing in the survey by presenting the webpages on paper and specifically asking the participants to consider all links and pick the best.

It remains to be seen if the semantic systems can inform the prediction of visual search patterns based on similarity values. Previous work [4, 28] suggests that menu search not only depends on the semantic similarity between the information being searched and the "correct" link, but also on the semantic similarity between the information being searched and the remaining "incorrect" link labels. We plan to study this effect in future studies.

CONCLUSION

In this paper we have compared three semantic systems, WordNet, LSA and PMI-IR, for predicting the link that humans would select given a goal on a webpage. The human behavior was captured in a user survey. Our study suggests that PMI-IR is the best system for predicting what the people would select.

Automated usability evaluation and prediction needs a comprehensive theory incorporating various aspects of visual search. Ecologically valid semantic analysis, such as in this paper, would contribute to such a theory, leading to a more accurate prediction of where people would click on actual webpages. Tools for web interfaces based on predictive modeling of user navigational behavior, such as analytic tools for studying user navigational behavior and intelligent agents for assisting users during web-based tasks, could also benefit from such a theory.

This paper demonstrates that comparing techniques that are not normally considered side-by-side can help to identify the best candidate or system to draw from another discipline to solve important HCI problems, which in this case helps the field progress towards developing comprehensive predictive theories.

ACKNOWLEDGMENTS

We thank Dr Peter Turney of the NRC, Canada for the access to the WMTS for PMI analysis; Dr Marilyn Blackmon for the access to the AutoCWW server at Colorado, used for LSA analysis; the team of Dr Ted Pederson at University of Minnesota Duluth for access to WordNet; and Tim Halverson for constant support. This research is supported by the National Science Foundation through grant IIS-0308244 to the University of Oregon with Anthony Hornof as the principal investigator.

REFERENCES

1. Ambroziak, J., & Woods, W. A. (1998). Natural Language Technology in Precision Content Retrieval. *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*.
2. Banerjee, S., & Pedersen, T. (2003). Extended gloss over-laps as a measure of semantic relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 805 - 810.
3. Blackmon, M., Kitajima, M., & Polson, P. (2003). Repairing usability problems Identified by the Cognitive Walkthrough for the Web. *Proceedings of ACM CHI 2003: Conference on Human Factors in Computing Systems*, 497-504.
4. Brumby, D., & Howes, A. (2004). Good enough but I'll just check: Web-page search as attentional refocusing. *Proceedings of the 6th International Conference on Cognitive Modeling*, 46-51.
5. Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources*, 2nd meeting of the North American Chapter of the Association for Computational Linguistics.
6. Byrne, M. D. (2001) ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*. 55, 41-84.
7. Cai, Z., McNamara, D. S., Louwerse, M., Hu, X., Rowe, M., & Graesser, A. C. (2004). NLS: A Non-Latent Similarity Algorithm. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. 180-185.
8. Chakravarthy, A. S., & Haase, K. B. (1995). NetSerf: Using semantic knowledge to find Internet information archives. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 4 - 11.
9. Chi, E. H., Pirolli, P., Chen, K., Pitkow. J. (2001) Using Information Scent to Model User Information Needs and Actions on the Web. *Proceedings of ACM CHI 2001: Conference on Human Factors in Computing Systems,* 490-497.
10. Chi, E. H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., Cousins, S. (2003) The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent(tm) Simulator. *Proceedings of ACM CHI 2003: Conference on Human Factors in Computing Systems.*
11. Church, K, Gale, W., Hanks, P., Hindle, D. (1991) Using Statistics in Lexical Analysis, in Zernik (ed.) *Lexical Acquisition: Exploiting OnLine Resources to*

Build a Lexicon, 115-164, Lawrence Erlbaum Associates Publishers.

12. Dumais, S. T., Furnas, G., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using Latent Semantic Analysis to improve access to textual information. *Proceedings of ACM CHI '98: Conference on Human Factors in Computing Systems*, 281-285.

13. Furnas, G., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964-971.

14. Hornof, A. J. (2004) Cognitive Strategies for the Visual Search of Hierarchical Computer Displays. *Human Computer Interaction*. 19, 183-223

15. Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the Conference on Research in Computational Linguistics*.

16. Landauer, T. K., & Dumais, S. T. (1997). A solution to the Plato's Problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

17. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

18. Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: an electronic lexical database* (pp. 265-283).

19. Lenat, D. (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure, *Communications of the ACM*, 38(11), 33-38.

20. Lieberman, H. (1995) Letizia: An Agent That Assists Web Browsing. *Proceedings of the International Joint Conference on Artificial Intelligence*.

21. Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of International Conference on Machine Learning*.

22. List of colleges and universities accessed at http://www.clas.ufl.edu/CLAS/american-universities.html

23. Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human Computer Interaction*, 8, 353-388.

24. Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28, 203-208.

25. Lynch, G., Palmiter, S., & Tilt. C. (1999). The Max Model: A standard web site user model, *Human Factors and the Web*, downloadable at http://zing.ncsl.nist.gov/hfweb/proceedings/lynch/

26. Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39-41.

27. Olston, C. & Chi, E. H. (2003). ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transactions on Computer Human Interaction*. 10(3), 177-197.

28. Pierce, B. J., Parkinson, S. R., & Sisson, N. (1992). Effects of semantic similarity, omission probability and number of alternatives on computer menu search. *International Journal of Man-Machine Studies*, 37(5), 653-677.

29. Pirolli, P., & Card, S. K. (1999). Information Foraging. *Psychological Review*, 106, 643-675.

30. Resnik, P. (1999). Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 95-113.

31. Richardson, R., & Smeaton, A. F. (1995). Using WordNet in a Knowledge-Based Approach to Information Retrieval. Working Paper.

32. Sense tagged corpus available at http://multisemcor.itc.it/semcor.html

33. Sutcliffe, R. F. E., Boersma, P., Bon, A., Donker, T., Ferris, M. C., Hellwig, P., et al. (1995). Beyond keywords: Accurate retrieval from full text documents. *Proceedings of the 2nd Language Engineering Convention*.

34. Terra, E., & Clarke, C. L. A. (2003). Frequency Estimates for Statistical Word Similarity Measures. *Proceedings of Human Language Technology Conference*. North American chapter of the Association for Computational Linguistics annual meeting, 244-251.

35. The Chronicle of Higher education http://chronicle.com/

36. Tullis, T. S. (1988). A system for evaluating screen formats: research and application. In R. Hartson & D. Hix (Eds.), *Advances in Human Computer Interaction*. 2, 214-286.

37. Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning*, 491-502.

38. WordNet Similarity CPAN reference http://search.cpan.org/dist/WordNet-Similarity/

39. Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics*, 133 – 138.