

---

# Modeling Information Scent: A Comparison of LSA, PMI-IR and GLSA Similarity Measures on Common Tests and Corpora

**Raluca Budiu**

Palo Alto Research Center  
3333 Coyote Hill Rd.  
Palo Alto, CA 94304 USA  
[budiu@parc.com](mailto:budiu@parc.com)

**Christiaan Royer**

Palo Alto Research Center  
3333 Coyote Hill Rd.  
Palo Alto, CA 94304 USA  
[royer@parc.com](mailto:royer@parc.com)

**Peter Pirolli**

Palo Alto Research Center  
3333 Coyote Hill Rd.  
Palo Alto, CA 94304 USA  
[pirolli@parc.com](mailto:pirolli@parc.com)

**Abstract**

In this paper we describe a comparison among three systems that estimate semantic similarity between words: Latent Semantic Analysis [6], Pointwise Mutual Information [17], and Generalized Latent Semantic Analysis [8]. We compare all these techniques on a unique corpus (TASA) and, for PMI and GLSA, we also report performance on a different web-based corpus. The evaluation is carried out through two kinds of tests: (1) synonymy tests, and (2) comparison with human word similarity judgments.

**Keywords**

PMI, LSA, GLSA, semantic similarity, corpus, computational linguistics

**ACM Classification Keywords**

H5.2. 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

## Introduction

In the context of navigation tasks, users must continually assess the semantics of labeled navigation options (e.g., Web links) and judge the relevance and utility of those options. The cues making up such labels have are known as **information scent** [11]. The notion that navigation choices are largely driven by information scent has become a central construct in cognitive models of users [10], usability testing systems [3], and design guidelines [16]. Information scent is perhaps most obvious in the case of browsing a web site for particular information. We often decide what links to follow based on our knowledge about the desired target information and on our intuitions that a particular set of link label words may be relevant to that target information. Theoretical analyses [10] and empirical usability research [16] suggests that information scent is the most important factor in Web navigation. In the field of information visualization [2], research [12] shows that, whereas the particular information visualization often has little effect on user performance, what impacts the user performance most is information scent.

As part of an effort to build user models and automated usability evaluation tools we need engineering techniques that, given a web page or an interface, can predict user judgments based on information scent. Often, in creating such models and tools, it is difficult to model the scent of the information that is displayed on the screen: if the users are looking for *peanut butter*, will they click on the word *jelly*, or on the word *nuts* or perhaps *tropical fruits*? One possible measure of information scent is **semantic similarity**. It is not feasible to ask people for similarity ratings of all words in a language that may be used in an user

interface or on the Web. Fortunately, several techniques that automatically estimate word similarity have emerged. Most of these techniques are based on word co-occurrence in a large corpus: two words are similar depending on how often they co-occur or on how often they occur in the same context. One problem for the modeler is which of these techniques to use? Is there one that is better than another? Attempts to compare these techniques have been made in the past [5, 15, 17]. These comparisons typically involve two kinds of tasks: (1) synonymy tests, in which the systems have to choose the word most similar to a target word; and (2) similarity tests, in which the systems provide their estimate of the similarity between two words; then those numbers are compared with the similarity ratings collected from people.

Unfortunately, any co-occurrence-based estimate of similarity is bound to be dependent on the corpus on which those co-occurrences were computed. Thus, if the corpus contained only Computer Science documents, probably the words *apple* and *computer* would occur in the same contexts often, and thus would be highly similar. However, most people would rate these words as far apart. We believe that one major deficiency of the previous studies has been using these different estimates of similarity on different corpora. Thus, if a study deems a method superior to another, it is not clear how much that is the merit of the method per se, and how much it is the merit of the corpus.

In this paper we examine how three such techniques: Latent Semantic Analysis (LSA)[6], Pointwise Mutual Information (PMI) [17], and Generalized Latent Semantic Analysis (GLSA) [8] compare with each other, when co-occurrence counts are based on the same

corpus. The common corpus is that used for creating the word frequency guide published by Touchstone Applied Science Associates (TASA) [18]. It is one of the corpora available at the official LSA web interface.

Albeit its many advantages, TASA is not a public domain corpus and will become inherently dated. One of the hopes nurtured by many in the information retrieval community is that the large amount of text available on the web could be used as a corpus. Then, the other question that we address in this paper is how the web (or a large sample of it) compares with a carefully constructed corpus such as TASA. Although, for reasons explained below, LSA could not be run on a large web-based corpus, we discuss how the TASA corpus and a larger web-based corpus compare to each other when they are used as base corpora for PMI and GLSA. We also look at whether **stemming** (i.e., whether words such as *read*, *reads*, and *reading* are considered as instances of one word *read* or as different words) makes any difference for PMI and GLSA.

In the remainder of the paper we present briefly the three techniques, then we present the two corpora and tasks on which we evaluate them. We end with the results of our evaluations and conclusions.

### Latent Semantic Analysis LSA

To compute word similarity, LSA [6] uses a large collection of documents. It first builds a word by document matrix, with each entry  $w_{ij}$  in that matrix representing the number of occurrences of the word  $i$  in the document  $j$ . Then LSA applies singular value decomposition (SVD) to compute a lower rank ( $K$ ) approximation of the original matrix.  $K$  is called

**number of factors.** The similarity of two words in LSA is computed as the cosine of their corresponding vectors in the reduced dimensionality matrix. Here we report LSA values obtained from the LSA website: <http://lsa.colorado.edu>. According to [6], LSA was able to score 64.4% in the TOEFL synonymy test. It was also successfully used in a number of other applications such as essay scoring and web page evaluation [1].

### Pointwise Mutual Information PMI

The Pointwise Mutual Information (PMI) (e.g., [7]) between two words  $A$  and  $B$  (roughly) captures how likely it is to find  $B$  in a text given that you know that the text contains  $A$ . In our tests we use a window of 16 words to capture co-occurrence. A web interface for computing PMI is available at <http://qlsa.parc.com>.

Turney [17] used PMI and the corpus behind the AltaVista search engine to estimate word similarities. He looked at various measures of co-occurrence and then used synonymy tests to compare the results obtained using AltaVista's 350 million web pages corpus and the results obtained from LSA, run on an encyclopedia of 30,473 articles. PMI resulted in performance comparable or better (between 62.5% and 73.75%) than LSA on TOEFL.

### Generalized Latent Semantic Analysis GLSA

GLSA [8] is a technique developed at PARC and starts with a large document corpus  $C$  and a set of words  $V$  (called **terms**). Then it computes a word-by-word matrix in which each entry  $w_{ij}$  represents the PMI (as computed above, with a 16-word window) between words  $i$  and  $j$  in the vocabulary  $V$ . Then, as for LSA, a SVD is applied to the resulting matrix and the similarity

$$PMI(A,B) = \log \frac{p(A,B)}{p(A)p(B)} \approx \frac{C(A,B)}{C(A)C(B)}$$

$p(A,B)$  is the probability that words  $A$  and  $B$  co-occur in the same document;  $p(X)$  is the probability that  $X$  occurs in a document;  $C(A, B)$  is the number of documents in which  $A$  and  $B$  co-occur (possibly within a given distance, called *window size*);  $C(X)$  is the number of documents in which  $X$  occurs.

A question in a **synonymy test** presents a word and asks the test taker to choose the word most similar to it among several (usually four) alternatives. Synonymy tests are often used for testing vocabulary knowledge of non-native English speakers.

between the two words is the cosine of the corresponding vectors in the reduced matrix. In the case of GLSA there are two parameters: (1) the **number of factors**, representing the rank to which the matrix is reduced (similar to LSA) ; and (2) the **number of terms** to be included in the vocabulary  $V$ . In the experiment reported in [8], GLSA obtained 86% performance on the TOEFL test. A web interface for GLSA is available at <http://qlsa.parc.com>.

### Corpora

We evaluate PMI, LSA and GLSA on two corpora. The first is, as discussed in the introduction, the TASA corpus. On the LSA website, the TASA corpus is labeled as "General reading up to 1<sup>st</sup> year college". The TASA corpus has been created in 1995 from "60,527 samples of text from 6,333 textbooks, works of literature, and popular works of fiction and nonfiction used in schools and colleges throughout the United States." It contains 17,274,580 tokens corresponding to 154,941 different types (i.e., words).

The second corpus used for our tests was gathered by web crawling by the Stanford WebBase project (<http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/>). Throughout the paper we refer to the first 8 million pages from this project as the Stanford corpus. The Stanford corpus contains 32,932,036 different types. Due to the high number of documents in this corpus and to the fact that LSA computes a word by document matrix, it was not possible to run LSA on the Stanford corpus. We report only PMI and GLSA results on that corpus. However, we did run all of LSA, PMI and GLSA on the TASA corpus, so we can fairly compare them at least on that corpus.

### Tests

When assessing similarity measures, two techniques are typically used. One is measuring the performance on a synonymy test, computed as the percent of correct answers. We use three synonymy tests: TOEFL, first used in [6], ESL [17], and Reader's Digest Word Power Vocabulary Test (RD). TOEFL and ESL are intended for foreign students. TOEFL has 80 questions and ESL 57 questions. The RD test has 103 synonymy questions, some of them containing multiple words. Another way of measuring the performance of a similarity metric is to compare with similarity ratings collected from people. Given a list of word pairs and some estimates of similarity, a rank correlation is run between the estimates and the data collected from human participants. We use several existing data sets: (1) Rubenstein and Goodenough (RG) ratings [14], containing 65 pairs; (2) Miller and Charles (MC) ratings [9], containing 30 word pairs also present in RG; (3) Resnick (R) ratings [13], containing the same pairs as MC; (4) the Word Similarity Test Collection (WS353) [4] of 353 pairs including the 30 pairs in MC; (5) Rohde, Gonnerman and Plaut's (RGP) ratings (RGP) [15], containing 400 word pairs.

### Evaluation

As discussed before, we only evaluated LSA on the unstemmed TASA corpus, using the LSA interface at <http://lsa.colorado.edu>. The GLSA and PMI implementations are those at <http://qlsa.parc.com>. The results are reported in Table 1. On the both the unstemmed and stemmed TASA corpus, GLSA is consistently better than both LSA and PMI. The only exception is the RGP test set on the unstemmed TASA corpus, where LSA performs best. When using the stemmed and the unstemmed versions of TASA, we

For LSA and GLSA, a parameter that affects the performance is the number of factors. We varied that number between 10-300 (maximum allowed by the interface) for LSA and between 10-2000 for GLSA. Contrary to previous reports, the variation with the number of factors was very noisy and we did not find a consistent performance peak around a certain number of factors for neither technique. Due to space reasons, we only report the best performance obtained and the corresponding number of factors.

All the results for GLSA were obtained using 32,000 terms. We also studied how GLSA performance changed with the number of terms. For a range between 4000 and 32000 terms, we found that the performance increased with the number of terms and reached a plateau between 16000 and 32000 terms.

mostly notice an increase in performance for the stemmed TASA for both PMI and GLSA. For this corpus GLSA seems less sensitive to stemming than PMI though (9% average increase for PMI and only 1% average increase for GLSA). For the web-based Stanford corpus, GLSA outperforms PMI on the synonymy tests, but PMI is generally better for the word similarity tests (except for RGP on Stanford unstemmed, where GLSA is better). Most of the time GLSA's performance is lower on the stemmed Stanford corpus than on the unstemmed Stanford corpus (on average 1.5% lower); this is also true for PMI (on average 1% lower). PMI is much better on the bigger Stanford corpus than on the TASA corpus. However, interestingly, for GLSA, although the synonymy tests seem to get some benefit from using the Stanford corpus (6.3% improvement in performance), the similarity tests are 13.4% worse with the larger Stanford corpus.

### Conclusions

Information scent is a major construct in navigation tasks that is often estimated by measures of semantic similarity. We examined three such measures (LSA, PMI, and GLSA) and how they compare to each other on several synonymy and word similarity tests. Since all these measures rely on word co-occurrence in a large document set, it is important to evaluate them on the same corpus. Another question of interest is whether a large, easily accessible web-based corpus, such as the Stanford corpus, can substitute a carefully formed corpus such as TASA. The current study found that GLSA outperforms LSA and PMI on the TASA corpus, but not on Stanford (except for synonymy tests), where PMI is better. PMI works best with a large web-based corpus, stemmed or not, possibly

Table 1. Performance of LSA, PMI and GLSA: percent correct on synonymy tests (TOEFL, ESL, RD) and rank correlation (\*100) on word similarity tests (WS353, MC, R, RG, RGP). The corpora used (TASA and Stanford) were either unstemmed (US) or stemmed (S). The smaller-font number in each cell (for LSA and GLSA) represents the number of factors used to obtain the best performance.

	TASA US			TASA S		Stanford US		Stanford S	
	LSA	PMI	GLSA	PMI	GLSA	PMI	GLSA	PMI	GLSA
<b>TOEFL</b>	60 140	23	<b>71</b> 251	20	<b>71</b> 99	53	<b>75</b> 494	48	<b>71</b> 408
<b>ESL</b>	44 300	22	<b>61</b> 602	31	<b>55</b> 396	54	<b>58</b> 881	46	<b>66</b> 242
<b>RD</b>	39 120	22	<b>56</b> 21	28	<b>58</b> 25	42	<b>67</b> 231	46	<b>68</b> 732
<b>WS353</b>	60 280	57	<b>65</b> 104	64	<b>67</b> 135	<b>71</b>	53 85	<b>71</b>	50 187
<b>MC</b>	75 250	65	<b>79</b> 392	73	<b>84</b> 35	<b>78</b>	62 140	<b>77</b>	57 234
<b>R</b>	71 220	61	<b>81</b> 32	72	<b>85</b> 35	<b>84</b>	61 141	<b>82</b>	55 233
<b>RG</b>	64 250	61	<b>77</b> 71	71	<b>77</b> 36	<b>75</b>	66 163	<b>74</b>	66 205
<b>RGP</b>	<b>63</b> 300	47	56 52	56	<b>59</b> 77	52	<b>55</b> 325	<b>57</b>	50 1138

because it is exclusively based on word co-occurrences with no attempt to remove any noise that may be present in such data and thus a large corpus may include more representative examples. For word similarity tests, GLSA works better with a carefully formed corpus such as TASA, but for synonymy tests, Stanford leads to better performance. Stemming a large corpus does not appear to help a lot, but stemming the TASA corpus does improve the

performance of PMI (probably by providing more examples). In conclusion, PMI seems like a winning technique for word similarity: it can be easily computed from a web-based corpus and the performance is very good. GLSA (which essentially combines PMI with SVD) may be better suited on a small, representative corpus such as TASA. For applications that involve small, domain-specific corpora, GLSA (and possibly LSA) may be a better choice.

### Acknowledgements

We would like to thank TASA and Robert Millard for giving us permission to use their corpus; Marilyn Hughes Blackmon for making the TASA corpus available to us; Doug Rohde for sharing his similarity data set. Portions of this research have been funded by ARDA/NIMD Contract No. MDA904-03-C-0404.

### References

- [1] Blackmon, M.H., Kitajima, M. and Polson, P.G. Web interactions: Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. *Proc. CHI 2005* (2005).
- [2] Card, S., Mackinlay, J. and Schneiderman, B. Information visualization. Morgan Kaufmann, 1999.
- [3] Chi, E., Rosien, A., Suppattanasiri, G. et al. The bloodhound project: Automating discovery of web usability issues using the infoscent simulator. *Proc. CHI 2003* (2003).
- [4] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan Z., Wolfman, G. and Ruppin, E. Placing search in context: The concept revisited. *ACM Transactions of Information Systems* 20, 1(2002), 116-131.
- [5] Kaur, I. and Hornof, A. J. A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. *Proc. CHI 2005*, ACM Press (2005).
- [6] Landauer, T. K. and Dumais, S. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104 (1997), 211-240.
- [7] Manning, C. and Schutze, H. Foundations of statistical natural language processing. MIT, 1999.
- [8] Matveeva, I., Levow, G., Farahat, A. and Royer, C. Terms representation with Generalized Latent Semantic Analysis. *Proc. RANLP 2005* (2005).
- [9] Miller, G. and Charles, W. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6, 1(1991), 1-28.
- [10] Pirolli, P. Rational analyses of information foraging on the web. *Cognitive Science* 29, 3 (2005), 343-373.
- [11] Pirolli, P. and Card, S. Information foraging. *Psychological Review* (1999).
- [12] Pirolli, P., Card, S. and Van Der Wege, M. The effect of information scent on searching information visualizations of large tree structures. *Proc. AVI 2000* (2000).
- [13] Resnick, P. Using information content to evaluate semantic similarity. *Proc. IJCAI 1995* (1995).
- [14] Rubenstein, H. and Goodenough, J. Contextual correlates of synonymy. *Communications of the ACM* 8, 10 (1965), 627-633.
- [15] Rohde, D.L.T., Gonnerman, L.M., and Plaut, D.C. An improved model of semantic similarity based on lexical co-occurrence. Submitted to *Cognitive Science*.
- [16] Spool, J., Perfetti, C. and Brittan, D. Designing for the scent of information. UI Engineering (2004).
- [17] Turney, P. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *Proc. ECML 2001* (2001).
- [18] Zeno, S., Ivens, S., Millard, R., and Duvvuri, R. The educator's word frequency guide. Touchstone Applied Science Associates (TASA), Inc., 1995.