

Semi-Supervised Evaluation of Search Engines via Semantic Mapping

Filippo Menczer
Department of Management Sciences
The University of Iowa
Iowa City, IA 52242, USA
filippo-menczer@uiowa.edu

ABSTRACT

Content and link information is used by virtually all search engines to crawl, index, retrieve, and rank Web pages. The correlations between similarity measures based on these cues and on semantic associations between pages is crucial in determining the performance of any search tool. A great deal of research is under way to understand how to effectively extract semantic information from Web pages by mining their text and links. A brute force approach has been used to build semantic maps, that given coordinates based on text and link similarity measures between two pages estimate how the two pages are related in meaning. In this paper we describe a semi-supervised methodology to evaluate the performance of search engines. The relevance of entire hit lists can be estimated by identifying a single relevant page for a given query. We illustrate the evaluation methodology by graphing precision-recall plots for three commercial search engines based on TREC queries. Preliminary results are in line with our intuition of how the different search engines use content and link information to rank hits.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance evaluation (effectiveness)*; H.3.5 [Information Storage and Retrieval]: Online Information Services—*Commercial, Web-based services*

General Terms

Experimentation, Measurement, Performance

Keywords

Semantic mapping, lexical similarity, link similarity, search engines, semi-supervised evaluation, precision-recall plots

1. INTRODUCTION

Search engines are successful when a user's information needs can be matched with the meaning of Web pages. To this end search engines must make assumptions about the correlations between page meaning and observable cues such as keywords, page content, and hyperlinks. Obviously these assumptions are neither completely true (otherwise everybody would be fully satisfied with the current state of the art) nor completely false (otherwise nobody

Copyright is held by the author/owner(s).
WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM xxx.

would be using search engines). In prior work I have formalized some of these assumptions and measured to what extent they are validated or violated by Web data [22, 23, 24]. For example, the relationship between content similarity and link probability sheds light on the processes that govern the growth of the Web and its link structure. Furthermore the relationships between both content and semantic similarity and link probability give some theoretical grounding to ongoing efforts to build effective Web crawlers [22]. These findings stem from a brute-force approach to measuring the relationships between three classes of similarity measures based on textual content, hyperlinks, and meaning of Web pages. A byproduct of this approach is the possibility of mapping, at a fine granularity, measurements of pairwise similarity based on content and links onto estimates of semantic relatedness [24].

The semantic mapping process is summarized in the next section, formalizing content, link, and semantic similarity measures and formulating estimates of precision and recall based on the distribution of semantic relatedness as a function of content and link similarity across billions of Web page pairs. The remainder of the paper introduces a novel methodology to estimate the relevance of arbitrary Web pages to arbitrary queries, based on computing text and link similarity to a single known relevant page and then comparing these measures with pre-compiled semantic maps. This approach amounts to a semi-supervised evaluation method for search engines, whereby precision and recall levels can be estimated from very little initial knowledge of relevant sets — namely a single page per query. Once outlined, the methodology is illustrated by building precision-recall curves for three commercial search engines (Google, Teoma and MSN). The results are consistent with our expectations based on what is known about the different ways in which hits are ranked by these engines. This suggests that the methodology can become a useful tool in comparing and evaluating search engine performance, even if search engines are viewed as black boxes.

The issue of semantic similarity and its relationship with link and lexical proximity has also been explored recently by Chakrabarti *et al.* [5, 7] and by Haveliwala *et al.* [15]. The latter in particular describe an automatic evaluation process closely related to the one proposed here. The final section of this paper discusses the differences and complementarity between the two approaches.

2. SEMANTIC MAPPING

Consider two objects p and q . An object could be a page or a query, but in this paper I will mainly refer to page objects. A search engine must compute a semantic similarity function $\sigma_s(p, q)$ to establish the degree to which the meanings of the two objects are related. The performance and success of a search engine depend

in great part on the sophistication and accuracy of the σ_s approximations implemented in its crawling, retrieval, and ranking algorithms.

While people are good at computing σ_s , i.e. assessing relevance, we can only approximate this function with computational methods. Semantic similarity is generally approximated from two main classes of cues in the Web: lexical cues (textual content) and link cues (hyperlinks). Lexical cues have traditionally been used by search engines to rank hits according to their similarity to the query, thus attempting to infer the semantics of pages from their lexical representation [29]. Similarity metrics are derived from the vector space model [31], which represents each document or query by a vector with one dimension for each term and a weight along that dimension that estimates the term’s contribution to the meaning of the document. The *cluster hypothesis* behind this model is that a document lexically close to a relevant document is also relevant with high probability [33].

The latest generation of Web search tools is beginning to integrate lexical and link metrics to improve ranking and crawling performance through better models of relevance. The best known example is the *PageRank* metric used by Google: pages containing the query’s lexical features are ranked using query-independent link analysis [4]. Links are also used in conjunction with text to identify hub and authority pages for a certain subject [17], guide search agents crawling on behalf of users or topical search engines [25, 26, 5, 27, 32], and identify Web communities [13, 18, 10, 11]. The hidden assumption behind all of these retrieval, ranking and crawling algorithms that use link analysis to make semantic inferences is a correlation between the graph topology of the Web and the meaning of pages, or more precisely the conjecture that one can infer what a page is about by looking at its neighbors. Such a conjecture has been implied or stated in various forms [21, 13, 2, 6, 9, 8, 16, 23, 24].

2.1 Similarity measures

The basic idea of semantic mapping is to quantitatively measure the relationships between content, link, and semantic topology of the Web at a fine level of resolution by building maps of σ_s in a space where the coordinates are given by σ_c and σ_l , i.e., by similarity metrics based on lexical content and link cues, respectively. Given the goal of mapping pairwise relationships, the first step was to sample a set of pages representative of the Web at large and for which independent semantic information was available along with the content and link data that is locally accessible by crawling the pages. We started from all the URLs in an RDF dump of the Open Directory Project¹ (ODP). For language consistency we eliminated the “World” branch, which classifies non-English pages. For classification consistency we also eliminated the “Regional” branch, which replicates the main topical tree for many geographical locations. After filtering we were left with 896,233 URLs organized into 97,614 topics. We picked 10,000 URLs uniformly from each of the 15 top-level branches, resulting in a final stratified sample of 150,000 URLs belonging to 47,174 topics. All of these URLs corresponded to working links to HTML pages available via the HTTP protocol. The pages were crawled, preprocessed and stored locally. For efficiency, there was a size limit so that only the first 20 KB of each page were downloaded, with a timeout of 10 seconds.

The second step was a simple brute force approach: for each pair of pages p, q measure $\sigma_c(p, q)$, $\sigma_l(p, q)$, and $\sigma_s(p, q)$. These three measures are described below.

Content similarity Let us define

$$\sigma_c(p_1, p_2) = \frac{\sum_{k \in p_1 \cap p_2} w_{kp_1} w_{kp_2}}{\sqrt{\left(\sum_{k \in p_1} w_{kp_1}^2\right) \left(\sum_{k \in p_2} w_{kp_2}^2\right)}} \quad (1)$$

where (p_1, p_2) is a pair of Web pages and w_{kp} is the frequency of term k in page p . This is the well-known “cosine similarity” function. The use of simple term frequency in place of more sophisticated TF-IDF weighting schemes in Equation 1 is due to the need for a document representation insensitive to different page samples and to different topic subsets. However, common noise words are eliminated [12] and other terms are conflated using the standard Porter stemmer [30]. Then only the most frequent 100 terms in each page are used in computing σ_c .

Link similarity Let us define

$$\sigma_l(p_1, p_2) = \frac{|U_{p_1} \cap U_{p_2}|}{|U_{p_1} \cup U_{p_2}|} \quad (2)$$

where U_p is the set containing the URLs of p ’s outlinks, inlinks, and of p itself. The outlinks are obtained from the pages themselves, while a set of at most 20 inlinks to each page in the sample is obtained by submitting a `link` query with the page URL to a search engine. Link similarity is really a neighborhood function, measuring the degree of clustering between the two pages. A high value of σ_l indicates that the two pages belong to a tightly clustered set of pages. Related measures are often used in link analysis to identify a community around a topic. If $\sigma_l(p_1, p_2) > 0$ there exists an undirected path between p_1 and p_2 of length $\ell \leq 2$ links. The higher σ_l , the greater the probability that there is a directed path between the two pages. Note that definition 2 is a special case of a Jaccard coefficient from set theory, and is also akin to the well known *co-citation* and *bibliographic coupling* measures used in bibliometrics.

Semantic similarity Let us define

$$\sigma_s(p_1, p_2) = \frac{2 \log \Pr[t_0(p_1, p_2)]}{\log \Pr[t(p_1)] + \log \Pr[t(p_2)]} \quad (3)$$

where $t(p)$ is the topic containing p in the ODP, t_0 is the lowest common ancestor of p_1 and p_2 in the ODP tree, and $\Pr[t]$ represents the prior probability that any page is classified under topic t . In practice we compute $\Pr[t]$ offline for every topic t in the ODP by counting the fraction of pages stored at node t in the tree, out of all the pages in the tree (that is, all the 896,233 unfiltered pages, not just the 150,000 sampled pages). This is a straightforward extension of the information-theoretic similarity measure [20]. The path from the root to t_0 is a measure of the meaning shared between the two topics, and therefore of what relates the two pages. Conversely the paths between t_0 and the two page topics are a measure of what distinguishes the meanings of the two pages. This semantic similarity measure clearly relies on the existence of a hierarchical organization that classifies all of the pages being considered. Sampling pages from the ODP guarantees that semantic information for each page is available and somewhat reliable — it is assessed by human editors rather than estimated by automatic content or link analysis methods.

Of all the pairs of pages in the ODP sample, only those with all three well defined similarity measures were considered valid. For

¹<http://dmoz.org>

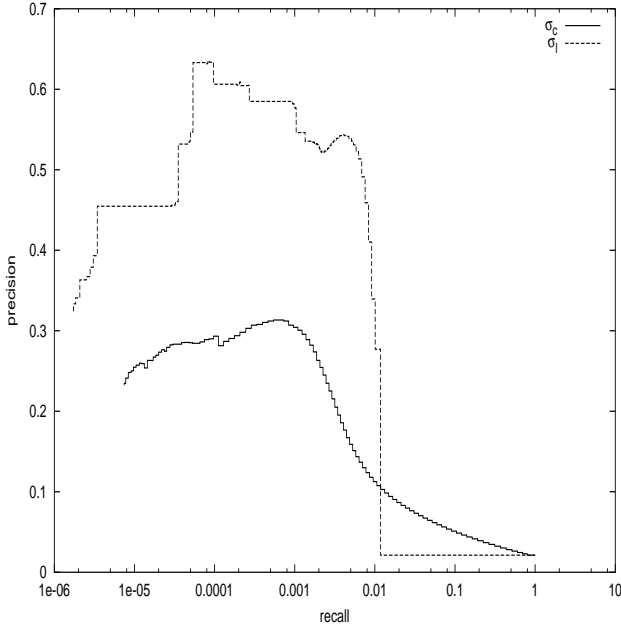


Figure 1: Precision-recall plots for rankings based on content and link similarity. For readability, recall is plotted on a logarithmic scale.

example a pair was discarded if a page timed out making σ_c and σ_l undefined, or if the inlinks of a page were not available making σ_l undefined. There were over 3.8×10^9 valid pairs. Each similarity measure is defined in the unit interval. This was divided into 100 bins, resulting in 10^6 similarity tuples; each pair was assigned to one tuple. By projecting the pairwise data thus collected onto one or more similarity dimensions, we can visualize some emerging topological patterns of the Web.

2.2 Ranking by similarity projection

The first step toward evaluating the effectiveness of lexical and link cues in estimating the meaning of pages is to assess the performance of a theoretical retrieval system that ranks pages based on just one similarity measure. In information retrieval the effectiveness of a document ranking system can be assessed, if the relevant set is known, using the standard *precision* and *recall* measures and the tool of precision-recall plots. While it would be extremely interesting to evaluate how effectively Web pages could be ranked based on, say, content or link similarity, this is generally impossible because relevant sets are unknown in the Web. However, one can think of using a page as a query (as in “query by example” retrieval systems) and ranking all other pages, then using semantic similarity data to assess the ranking. Our data supports this approach, assuming the above procedure is repeated with every page used as an example (this is very similar to the method proposed in [15]). Let us define *projected* precision and recall as follows:

$$P(s) = \frac{\sum_{p,q:\sigma(p,q)\geq s} \sigma_s(p,q)}{|p,q:\sigma(p,q)\geq s|} \quad (4)$$

$$R(s) = \frac{\sum_{p,q:\sigma(p,q)\geq s} \sigma_s(p,q)}{\sum_{p,q} \sigma_s(p,q)} \quad (5)$$

where $\sigma = \sigma_c$ to evaluate content-based ranking and $\sigma = \sigma_l$ to evaluate link-based ranking.

The precision-recall plots in Figure 1 show that ranking by link

similarity produces better precision at very low recall levels, while ranking by content similarity produces better precision at higher recall levels. This is consistent with the use of link analysis in ranking by search engines, since users often look at only a few hits and thus perceive precision as a more immediate performance indicator than recall. A surprising pattern is that for both lexical and link based ranking, the precision-recall curves are not strictly monotonic. Precision increases with recall for very low recall levels. This can be seen as evidence that both text and link similarity become more noisy as they approach 1, i.e., less reliable as predictors of semantic similarity.

2.3 Precision and recall maps

To visualize how accurately semantic similarity can be approximated from content and link cues, we need to map the σ_s landscape as a function of σ_c and σ_l . There are two different types of information about σ_s that can be mapped for any given (σ_c, σ_l) coordinates: (1) averaging highlights the expected values of σ_s and is akin to precision; (2) summing captures the relative mass of semantically similar pairs and is akin to recall. Let us therefore define *localized* precision and recall for this purpose as follows:

$$P(s_c, s_l) = \frac{S(s_c, s_l)}{N(s_c, s_l)} \quad (6)$$

$$R(s_c, s_l) = \frac{S(s_c, s_l)}{S_{tot}} \quad (7)$$

where

$$S(s_c, s_l) = \sum_{p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l} \sigma_s(p,q) \quad (8)$$

$$N(s_c, s_l) = |p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l| \quad (9)$$

$$S_{tot} = \sum_{p,q} \sigma_s(p,q) \quad (10)$$

and (s_c, s_l) is a coordinate value pair for (σ_c, σ_l) .

Figure 2 maps R and P as a function of content and link similarity coordinates. These semantic maps provide for a much richer and more complex account of the information about meaning that can be inferred from text and link cues, comparing to the projections given by Equations 4 and 5.

A number of interesting observations are in order. Note first from the recall map that the distribution of semantic similarity is heavily skewed toward the origin. This is due to the fact that all similarity measures are themselves skewed — given two random pages one would not expect them to be similar in content, clustered in link space, or related in meaning.² Since the majority of pairs occur near the origin, the same holds for most of the semantically related pairs, as shown by the high recall. However all this relevant mass is washed away in a sea of unrelated pairs, so that precision near the origin is negligible. This creates an obvious challenge for search engines: achieving high recall costs dearly in terms of precision and leads to user frustration. While emphasis on precision is a very reasonable approach for a search engine, this data suggests that the cost of such a choice in terms of recall is very high in the Web.

An analogous, but much weaker effect is seen for very high content and link similarity. Here we observe a moderate peak in recall, which does not correspond to a peak in precision. There are a few highly similar and highly clustered pages here, but they are not necessarily highly related.³

²All similarity measures actually have roughly exponential distributions [24].

³This cluster is mainly due to adult “spammer” sites.

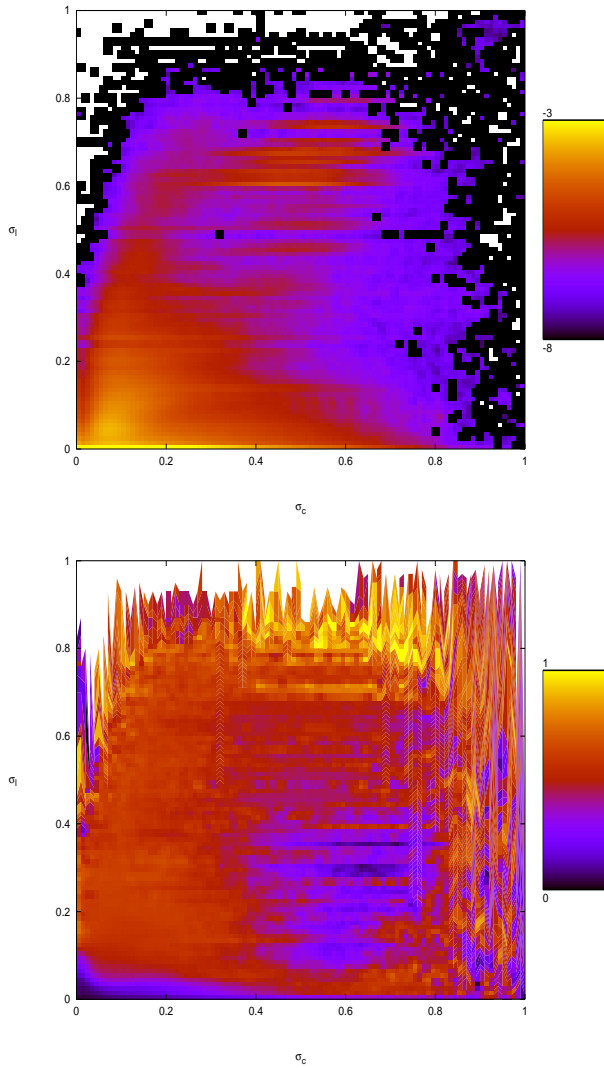


Figure 2: Semantic maps of recall (top) and precision (bottom) for all pairs of sample Web pages. For readability, recall is visualized on a logarithmic scale between 10^{-8} and 10^{-3} or above.

A few observations can be made from the precision map. First, for very high content similarity there is significant noise making it difficult to get a clear signal. There are many relevant pages in this region, but they cannot be identified from link or content analysis. Maybe these are cases where authors do not know of related pages or do not want to point to the competition. There are also many pairs in this region that are not semantically related. This explains the poor performance of the first generation of search engines, which ranked solely by content similarity — an unreliable signal in its highest range. Second, there is a clearly strong signal for medium-high content similarity and high link similarity (the yellow area). Such peaks correspond to text and link similarity values such that there are very few pairs, but those few correspond to highly related pages. These hits can be identified by a search engine by first selecting pages in the right content similarity range, then distilling the most relevant ones by link analysis. Finally, the presence of a local minimum in precision (the purple area) is peculiar and deserves further study.

3. SEMI-SUPERVISED EVALUATION

The semantic maps give us a tool to estimate the relatedness between any two pages. If we know what one page is about, we can estimate the degree to which another arbitrary page is about the same topic. This can be done by automatic measurements of similarity based on locally observable cues: the content and links of the two pages. Extending the idea, if we know one relevant page for a query, we can estimate the relevance of any given set of pages. This suggests a straightforward application of semantic mapping — evaluating search engines.

In machine learning, semi-supervised methods are techniques for extracting knowledge from data when only a very small fraction of the data is labeled so as to provide examples to the learning system [3, 28]. Typically in supervised learning a sufficient number of labeled examples are available (the training set). When a training set is not available, a few examples can be labeled by hand and then more examples can be labeled automatically by a bootstrapping process. Finally the resulting (approximated) training set is used in a supervised setting.

Here we liberally borrow the semi-supervised metaphor and apply it to an evaluation task rather than a learning task. As mentioned above, evaluation of retrieval systems requires knowledge of relevant sets, which are unavailable on the Web. (See [1] for a recent review of search engine evaluation techniques.) The traditional approach is to have users manually assess the relevance of all documents in a collection. This is impossible in the Web due to its large size and dynamic nature. Measuring precision would require user assessment of all retrieved pages, a very expensive procedure. However, even if it were feasible to assess the relevance of all pages indexed by a search engine, recall still could not be measured because of the many relevant pages potentially unknown to the search engine due to its limited coverage [19].

Using the semi-supervised approach, suppose a single highly relevant page r is known for a given query. Such a page can be identified by hand at relatively low cost. Through the semantic maps, we can imagine bootstrapping a virtual relevant set made of pages highly related to r . The idea is to estimate the semantic similarity between the retrieved (hit) set and such a virtual relevant set, from measures of lexical and link similarity. Finally we can approximate the precision and recall of the entire hit set.

3.1 Methodology

For a query q , identify a highly relevant page r_q . Then consider each page p in the hit set H_q obtained by a search engine in response to q . Let $s_c(p, q) = \sigma_c(p, r_q)$ (from Equation 1) and $s_l(p, q) = \sigma_l(p, r_q)$ (from Equation 2). To estimate the precision and recall of a hit set we can use the localized semantic similarity measures computed from our Web sample to build the precision and recall maps (cf. Equations 6 and 7).

Formally, let $H_q^m = \{p_1 \cdots p_m\} \subset H_q$ be the set of top m hits in H_q . Then estimate the set precision and recall for H_q^m as follows:

$$P(H_q^m) \approx \frac{\sum_{i=1}^m \frac{S(s_c(p_i, q), s_l(p_i, q))}{N(s_c(p_i, q), s_l(p_i, q))}}{m} \quad (11)$$

$$R(H_q^m) \approx \frac{\sum_{i=1}^m S(s_c(p_i, q), s_l(p_i, q))}{S_{tot}} \quad (12)$$

where the $S(\cdot)$ function is looked up from Equation 8, $N(\cdot)$ from Equation 9, and S_{tot} is the constant in Equation 10. Averaging over

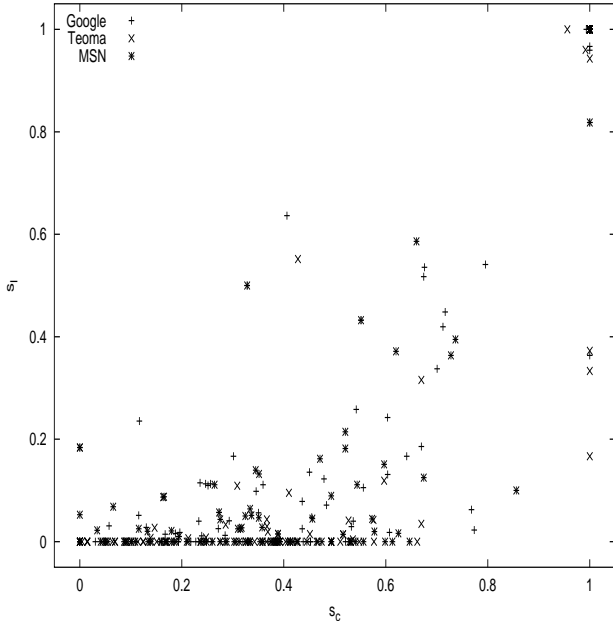


Figure 3: Lexical and link similarity coordinates of the top three hits returned by the three search engines for each of the 30 TREC queries. The points in the top right-hand corner correspond to the cases in which the relevant homepage is ranked among the top three hits.

a set of queries Q we finally obtain:

$$P(m) = \frac{1}{|Q|} \sum_{q \in Q} P(H_q^m) \quad (13)$$

$$R(m) = \frac{1}{|Q|} \sum_{q \in Q} R(H_q^m). \quad (14)$$

In other words, the relevance of every hit is assessed by estimating its semantic similarity to the known relevant page. This estimation is done by measuring the lexical and link similarity between the hit and the relevant page. These measures are used as coordinates into the general precision and recall maps.

The precision and recall levels thus computed, $P(m)$ and $R(m)$, can be plotted against each other using m as a coordination level [33].

3.2 Comparison of three search engines

To test the proposed methodology, I applied it to the evaluation of three large commercial search engines: Google, Teoma, and MSN. The choice was due to the fact that Google is accessible through the Google Web API⁴ and the other two are accessible through HTTP agents in compliance with the Robot Exclusion Standard. Most other commercial search engines disallow access by agents.

A set Q of 30 queries were randomly chosen among the TREC⁵ 2001 homepage finding topics. These are realistic Web queries because they are short, and they are also selected in such a way as to guarantee that there exists at least one Web “homepage” that is ideally relevant for each query. Such relevant homepages were manually identified for each of the 30 queries, as shown in Table 1.

⁴<http://www.google.com/apis>

⁵<http://trec.nist.gov>

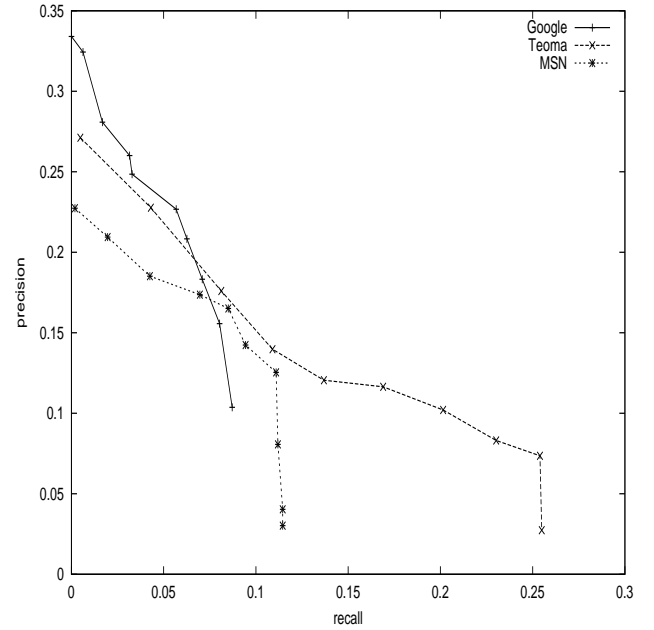


Figure 4: Precision-recall plots for the three search engines, based on semi-supervised evaluation.

Each of the 30 queries was fed to each of the search engines, and the top 10 hits returned by each search engine were stored. For each search engine and query, the lexical and link similarity coordinates (s_c , s_l) of each hit were computed by comparing the hit with the query’s relevant homepage according to Equations 1 and 2. For lexical similarity, 10 inlinks were obtained for each hit using the Google Web API. To illustrate this step, Figure 3 plots some of the top hits in similarity space.

For each search engine, query, and level m ($1 \leq m \leq 10$), the similarity coordinates of the top m hits were then mapped into semantic similarity values to obtain the estimates of query precision and recall according to Equations 11 and 12. Finally, these values were averaged across the $|Q| = 30$ queries to obtain the mean precision and recall estimates according to Equations 13 and 14. The resulting $P(m)$ and $R(m)$ were used to draw the precision-recall plots in Figure 4.

The precision-recall plot for Google is consistent with our intuition about this search engine’s emphasis on precision. The PageRank score allows Google to achieve very high precision at low recall levels. This seems to be a winning strategy since many users are most interested in getting a few highly relevant hits, and few users search past the few highest-ranked hits. While Google’s precision decreases rapidly with increasing recall, it does not become as low as that of the other two search engines we evaluated. However, Google’s recall remains low compared to the other two search engines. Teoma’s performance is complementary to Google’s. At low recall its precision is lower, but eventually Teoma achieves significantly higher recall — over twice the maximum recall obtained by Google or MSN. Like Teoma, MSN is outperformed by Google’s precision at low recall and eventually it achieves higher recall than Google. However, MSN is consistently outperformed by Teoma at every precision/recall level.

Precision and recall can also be combined into a single perfor-

Table 1: TREC queries used in the evaluation of the three search engines, with URLs of relevant homepages.

Query (q)	Homepage (r_q)
Linux Documentation Project	http://www.tldp.org/
DogHouse Technologies, Inc. Victoria, Australia	http://www.dogtech.com/
New England Journal of Medicine	http://www.vicnet.net.au/
CNET	http://content.nejm.org/
Kennedy Space Center	http://www.cnet.com/
Australian Democrats	http://www.ksc.nasa.gov/
Sheraton Hotels Latin America	http://www.democrats.org.au/
Haas Business School	http://www.geographia.com/sheraton/
Manly, Australia	http://www.haas.berkeley.edu/
Harvard Graduate Student Council	http://www.manlyweb.com.au/
Texas Department of Human Services	http://hcs.harvard.edu/gsc/
Cable Wireless, Inc.	http://www.dhs.state.tx.us/
Brent Council	http://www.cw.com/
Solaris certification	http://www.brent.gov.uk/
Ada information clearinghouse	http://suned.sun.com/US/certification/solaris/
Lockwood Memorial Library	http://www.adaic.org/
Maine Office of Tourism	http://ublib.buffalo.edu/libraries/units/lml/
HKUST Computer Science Dept.	http://www.visitmaine.com/home.php
Wah Yew Hotel	http://www.cs.ust.hk/
NCREL	http://www.fastnet.com.au/hotels/zone4/my/my00198.htm
Ohio Public Library Information Network	http://www.ncrel.org/
SUNY Buffalo	http://www.oplin.lib.oh.us/
Sarah Jane Johnson United Methodist Church	http://www.buffalo.edu/
Chicago NAP	http://hwmin.gbgm-umc.org/sarahjohnsonny/
Graduate Theology Union	http://www.aads.net/main.html
Beachwood Motel, Orchard Beach, Maine	http://www.gtu.edu/
Chemistry, Leeds University	http://www.beachwood-motel.com/
Donald W. Shaw Real Estate	http://www.chem.leeds.ac.uk/
Douglas Barclay Law Library	http://www.donaldshaw.com/
	http://www.law.syr.edu/lawlibrary/lawlibrary.asp

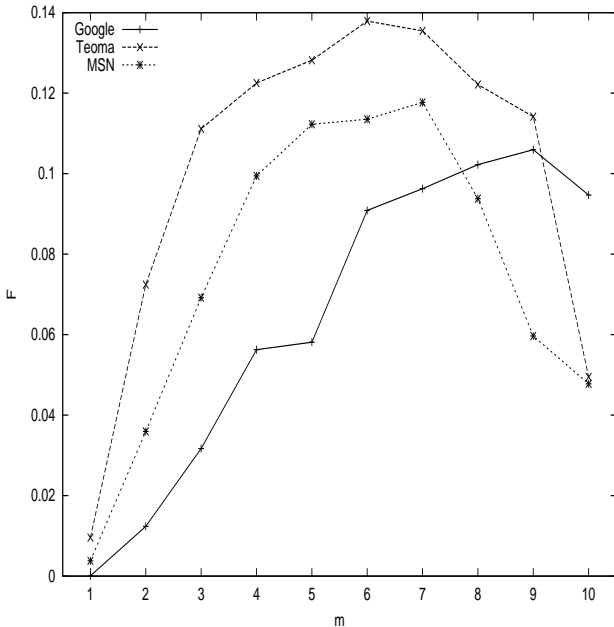


Figure 5: F-measure versus coordination level m for the three search engines, based on semi-supervised evaluation.

mance metric called *F-measure* [1] defined as the harmonic mean:

$$F(m) = 2 \frac{P(m)R(m)}{P(m) + R(m)}. \quad (15)$$

Figure 5 plots the F-measure for the three search engines. We see that for $m < 10$ Teoma dominates based on this metric. The low recall achieved by Google hurts its performance from this perspective. However, because Google maintains high precision, its F value surpasses the other two search engines for $m = 10$.

4. DISCUSSION

Understanding how semantic information can be mined from the content and links of Web pages is key for advancing search engine technology. This paper reported on a large-scale, brute-force effort to map semantic association in a topological space generated from content and link based metrics. The semantic maps visualize a massive amount of data collected from billions of pairs of Web pages and thus provide us with a valuable snapshot of how the meaning of a page can be estimated from its content similarity and link clustering to other pages.

The use of semantic maps to estimate relevance and thus evaluate search engines is a natural application of this compiled knowledge. This paper outlined a methodology to perform such an evaluation with only minimal effort from human users who need to identify few relevant pages. Of course the method's accuracy should improve when more than one page is labeled by users as relevant. However I have shown how to apply the methodology with a single relevant page labeled per query.

The preliminary results reported here are encouraging in that they are consistent with our daily experience using popular search engines. The conclusion is that among the three engines evaluated, Google appears as the best choice if precision is the most important aspect of performance. For example the queries used in our evaluation are based on homepages, therefore they represent search requests in which high precision seems to be both important and attainable. On the other hand, our evaluation suggests that Teoma would be a better choice when both precision and recall matter.

Haveliwala *et al.* [15] recently proposed an automatic evaluation process that is closely related to the one introduced here. The authors evaluate and compare different ranking schemes in the absence of user relevance assessments by measuring the correlation between (i) the ranking obtained by an algorithm to be evaluated, and (ii) a “ground truth” ordering obtained using a semantic similarity measure based on a Web hierarchy (actually they used the ODP as well). There are a number of differences that make the two approaches complementary in goals, scope, and implementation. First, the goals are different because we are interested not in evaluating known ranking schemes over known document sets, but rather unknown schemes employed by search engines over unknown crawl sets. Second, the scope is different because our approach can estimate both precision and recall from a few hits, while ranking correlation yields a single quality measure from the similarity scores between pages in the directory. Also, in our approach any query can be used, as opposed to using pages from the directory as queries. On the other hand, our approach does require some supervision (labeling at least one relevant page) and thus is not completely automatic. Notice however that our approach could be automated by using topic names or pages from a directory as queries — although using the same directory for both semantic maps and queries would introduce bias and should therefore be avoided. Third, there is an important difference in how semantic similarity is computed in the two approaches. The entropy-based measure that I propose (Equation 3) accounts for the distribution of pages across nodes in the directory tree. The two metrics coincide if the tree is perfectly balanced, which typically is not the case in reality. Notwithstanding these differences, both approaches are trying to address a common and important question: how to make the evaluation of Web retrieval systems feasible without renouncing the quality of user assessments.

The next step in this project is one of “meta-evaluation,” i.e. evaluating formally the semi-supervised evaluation methodology by gauging its outcome against that of traditional evaluation schemes based on direct user assessments of retrieved page sets. Semi-supervised evaluation estimates both precision and recall, therefore user studies will need to account for the fact that users may not have full knowledge of relevant sets.

Another direction for future work is to use *topical* semantic maps, that is, similarity data based on pairs of pages within a particular topic. Authors of pages on different topics use text and links in different ways, so similarity maps differ significantly across topical areas [24]. Therefore if it were possible to associate a query not only with a relevant page but also with a topic, one might obtain better relevance estimates using semantic maps specific to that topic, as long as the topic was present within the directory. The idea of restricting computation to sets of pages within topics has also been proposed for metrics such as PageRank [14].

The obvious advantage of the proposed evaluation scheme is that it is cheap and easy to implement once the semantic maps are compiled. New semantic maps can be compiled, based on local similarity measures that differ from the ones explored here. For example one might consider alternative definitions of content, link or seman-

tic similarity, various term weighting schemes, different hierarchical classifications or more specialized ontologies, or different cues altogether.

The equally obvious disadvantage of semi-supervised evaluation is that it is just an estimation method. I would argue that it is the “right” approach to estimation because it is grounded in independent assessments that are originally made by human editors, spanning a wide range of topics — if not the exact queries on which the engines are evaluated. However any estimation method is prone to a number of bias sources including page sampling, topic sampling, editor expertise, hierarchical topic structure, linguistic limitations, lexical distributions, link noise, labeling errors, query selection, and so on. Such limitations must be acknowledged and carefully considered before the proposed evaluation scheme is used to draw conclusions about the actual performance of search engines.

5. ACKNOWLEDGMENTS

I am grateful to Padmini Srinivasan, Nick Street, and Gautam Pant for their helpful comments. Thanks go to the Open Directory Project for making their data publicly available. This work is funded in part by NSF CAREER Grant No. IIS-0133124.

6. REFERENCES

- [1] R. Belew. *Finding Out About: A Cognitive Perspective on Search Engines and the WWW*. Cambridge University Press, Cambridge, 2000.
- [2] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. 21st ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 104–111, 1998.
- [3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann, 1998.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.
- [5] S. Chakrabarti. *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann, San Francisco, 2003.
- [6] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks*, 30(1–7):65–74, 1998.
- [7] S. Chakrabarti, M. Joshi, K. Punera, and D. Pennock. The structure of broad topics on the Web. In D. Lassner, D. De Roure, and A. Iyengar, editors, *Proc. 11th International World Wide Web Conference*, pages 251–262, New York, NY, 2002. ACM Press.
- [8] B. Davison. Topical locality in the Web. In *Proc. 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–279, 2000.
- [9] J. Dean and M. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, 1999.
- [10] G. Flake, S. Lawrence, and C. Giles. Efficient identification of Web communities. In *Proc. 6th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.
- [11] G. Flake, S. Lawrence, C. Giles, and F. Coetzee. Self-organization of the Web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

- [12] C. Fox. Lexical analysis and stop lists. In *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.
- [13] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Proc. 9th ACM Conference on Hypertext and Hypermedia*, pages 225–234, 1998.
- [14] T. Haveliwala. Topic-sensitive PageRank. In D. Lassner, D. De Roure, and A. Iyengar, editors, *Proc. 11th International World Wide Web Conference*. ACM Press, 2002.
- [15] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the Web. In D. Lassner, D. De Roure, and A. Iyengar, editors, *Proc. 11th International World Wide Web Conference*. ACM Press, 2002.
- [16] M. Henzinger. Link analysis in Web information retrieval. *IEEE Data Engineering Bulletin*, 23(3):3–8, 2000.
- [17] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [18] S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks*, 31(11–16):1481–1493, 1999.
- [19] S. Lawrence and C. Giles. Accessibility of information on the Web. *Nature*, 400:107–109, 1999.
- [20] D. Lin. An information-theoretic definition of similarity. In J. Shavlik, editor, *Proc. 15th Intl. Conference on Machine Learning*, pages 296–304, San Francisco, CA, 1998. Morgan Kaufmann.
- [21] F. Menczer. ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery. In *Proc. 14th International Conference on Machine Learning*, pages 227–235, 1997.
- [22] F. Menczer. Growing and navigating the small world Web by local content. *Proc. Natl. Acad. Sci. USA*, 99(22):14014–14019, 2002. www.pnas.org/cgi/doi/10.1073/pnas.212348399.
- [23] F. Menczer. Lexical and semantic clustering by web links. *IEEE Trans. on Knowledge and Data Engineering*, Submitted, 2002. Shorter version available as Computing Research Repository (CoRR) Technical Report [arXiv.org:cs.IR/0108004](http://arXiv.org/cs.IR/0108004).
- [24] F. Menczer. Mapping the semantics of Web text and links. *IEEE Journal on Selected Areas in Communications*, Submitted, 2002. <http://dollar.biz.uiowa.edu/~fil/Papers/maps.pdf>.
- [25] F. Menczer and R. Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the Web. *Machine Learning*, 39(2–3):203–242, 2000.
- [26] F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan. Evaluating topic-driven Web crawlers. In D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, editors, *Proc. 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 241–249, New York, NY, 2001. ACM Press.
- [27] F. Menczer, G. Pant, and P. Srinivasan. Topic-driven crawlers: Machine learning issues. *ACM TOIT*, Submitted, 2002. <http://dollar.biz.uiowa.edu/~fil/Papers/TOIT.pdf>.
- [28] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proc. 9th Intl. Conf. on Information and Knowledge Management (CIKM-2000)*, 2000.
- [29] B. Pinkerton. Finding what people want: Experiences with the WebCrawler. In *Proc. 1st International World Wide Web Conference*, 1994.
- [30] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [31] G. Salton and M. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.
- [32] P. Srinivasan, G. Pant, and F. Menczer. A general evaluation framework for topical crawlers. *IEEE Trans. on Knowledge and Data Engineering*, Submitted, 2002.
- [33] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979. Second edition.