Combining Link and Content Analysis to Estimate Semantic Similarity

Filippo Menczer*
School of Informatics and Department of Computer Science
Indiana University, Bloomington
fil@indiana.edu

ABSTRACT

Search engines use content and link information to crawl, index, retrieve, and rank Web pages. The correlations between similarity measures based on these cues and on semantic associations between pages therefore crucially affects the performance of any search tool. Here I begin to quantitatively analyze the relationship between content, link, and semantic similarity measures across a massive number of Web page pairs. Maps of semantic similarity across textual and link similarity highlight the potential and limitations of lexical and link analysis for relevance approximation, and provide us with a way to study whether and how text and link based measures should be combined.

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

General Terms: Measurement

Keywords: Web search, semantic maps, content and link similarity, precision, recall

1. INTRODUCTION

Search engines typically combine analysis of Web page content and links to retrieve and rank hits in response to user queries. While there is a large body of literature on both text and link analysis,[1] it is not known how these should be combined to achieve optimal retrieval performance. Here I present preliminary data on how similarity measures based on content and link analysis might be combined in order to best approximate a measure of semantic similarity induced by manual classification of pages into a hierarchical directory. More generally, I explore what page content and links say about each other, and what they say about the meaning of pages.

The connection between Web lexical and link cues, and between either of these and semantic characterizations of pages, has been previously studied in the context of hypertext document classification [3], topic distillation [1], and navigation [4].

2. METHODOLOGY

In the present approach the relationships between content, link, and semantic topology in the Web is studied empirically at a fine level of resolution. The idea is to measure the correlations between similarity measures driven by content, link, and semantic evidence.

The first step is to sample a set of pages that are representative of the Web at large and for which independent semantic information is available along with content and link data locally accessible by crawling the pages. The Open Directory (ODP) was used to sample 10,000 URLs uniformly from each of the 15 top-level branches, resulting in a set of 150,000 URLs belonging to 47,174 topics. The pages were crawled, preprocessed and stored locally.

The second step is a brute force approach: for each pair of pages p, q measure three similarities all defined in $[0, 1]$:

Content similarity $\sigma_c(p, q) = (\vec{p} \cdot \vec{q})/(\|\vec{p}\| \cdot \|\vec{q}\|)$ where $\vec{p}, \vec{q}$ are the representations of the pages in word vector space, after removing stop words and stemming. This is actually the "cosine similarity" function, traditionally used in information retrieval.

Link similarity $\sigma_l(p, q) = |U_p \cap U_q|/|U_p \cup U_q|$ 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 inlinks to each page in the sample is obtained from a search engine. This Jaccard coefficient measures the degree of clustering between the two pages, with a high value indicating that the two pages belong to a clique.

Semantic similarity $\sigma_s(p, q) = \frac{2 \log \Pr[t_0(p,q)]}{\log \Pr[t(p)] + \log \Pr[t(q)]}$ where $t(p)$ is the topic containing p in ODP, t_0 is the lowest common ancestor of p and q in the ODP tree, and $\Pr[t]$ represents the prior probability that any page is classified under topic t. This information theoretic measure uses entropy to compare how much meaning is shared by two topic nodes compared to what distinguishes them. It reduces to the familiar tree distance measure for a perfectly balanced tree. This measure relies on the existence of a hierarchical organization such as ODP that classifies all of the pages being considered. However, the ODP ontology is more complex than a simple tree; it has various types of cross-reference links between categories. Here I sidestep this issue by reducing the directory to a single, prototypical tree.

A total of 3.8×10^9 page pairs yielded valid $(\sigma_c, \sigma_l, \sigma_s)$ tuples. These were divided into 10^6 bins (100 bins per similarity measure). From this 3D histogram information a number of interesting statistics and visual maps can be derived. The Pearson's correlation coefficients between pairs of similarity metrics are $\rho(\sigma_c, \sigma_l) = 0.10$, $\rho(\sigma_c, \sigma_s) = 0.11$, and $\rho(\sigma_l, \sigma_s) = 0.08$. These are weak but very significant correlations when considering the number of pairs.

All three metrics appear to have a roughly exponential distribution. Most pairs tend to have very small values for all similarity measures; given two random pages we do not expect them to be lexically similar, closely clustered, or semantically related. The very small number of pairs with high similarity values is the main reason for the low correlation coefficients.

*Funded in part by NSF CAREER Grant No. IIS-0348940.

[1] See [2] for a review of the literature.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

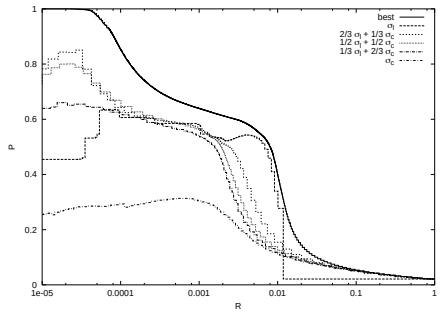

Figure 1: Precision-recall plots for rankings based on various linear combinations of content and link similarity.

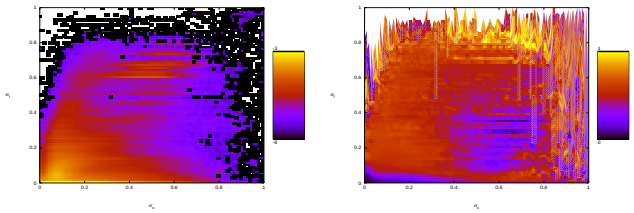

Figure 2: Semantic maps of recall R (left) and precision P (right) for all pairs of sample Web pages. For readability, R is visualized on a log scale between 10^{-8} and 10^{-3} or above.

3. SEMANTIC PROJECTIONS

In information retrieval the effectiveness of a document ranking system can be assessed by plotting *precision* versus *recall*, assuming relevant sets are known. The data collected here allows one to evaluate how effectively Web pages are ranked based on content or link similarity by using σ_s as a surrogate for relevance assessments and each page as a query (as in "query by example"). Let us define *linear-projected* precision and recall as follows:

$$P(\alpha,\beta) = \frac{\sum_{p,q:\alpha\sigma_l(p,q)+(1-\alpha)\sigma_c(p,q)\geq\beta} \sigma_s(p,q)}{|p,q:\alpha\sigma_l(p,q)+(1-\alpha)\sigma_c(p,q)\geq\beta|} \quad (1)$$

$$R(\alpha,\beta) = \frac{\sum_{p,q:\alpha\sigma_l(p,q)+(1-\alpha)\sigma_c(p,q)\geq\beta} \sigma_s(p,q)}{\sum_{p,q}\sigma_s(p,q)} \quad (2)$$

where α is the slope of the projection line and β is the intercept, which can be used as a ranking parameter.

The projected precision-recall plots in Figure 1 are based on linear combinations of the form $\alpha\sigma_l+(1-\alpha)\sigma_c$. Since most of the σ_s "mass" occurs near the origin ($\sigma_c = \sigma_l = 0$), recall is visualized on a log scale. Ranking by link similarity alone ($\alpha = 1$) produces better precision at low recall levels, while ranking by content similarity alone ($\alpha = 0$) produces better precision at high recall levels. This is consistent with the use of link analysis in ranking by search engines, since most users only look at a few hits. However, combinations of content and link similarity yield better compromise rankings. In particular, using any amount of content information in addition to link analysis improves precision at both low recall levels (where link information alone is noisy) and high recall levels (where link information alone is useless). Thus a search engine could improve the quality of its results by appropriately combining query-independent link analysis with content analysis.

4. SEMANTIC MAPS

To visualize the relationship between different similarity measures, let us map σ_c and σ_l into σ_s. For any given (σ_c, σ_l) coordinates, averaging σ_s is akin to precision while summing is akin to recall. Let us define *localized* precision and recall as follows:

$$P(s_c,s_l) = \frac{\sum_{p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l}\sigma_s(p,q)}{|p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l|} \quad (3)$$

$$R(s_c,s_l) = \frac{\sum_{p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l}\sigma_s(p,q)}{\sum_{p,q}\sigma_s(p,q)}. \quad (4)$$

Figure 2 maps R and P as functions of content and link similarity coordinates. The majority of semantically related pairs occur near the origin, as shown by the high R, because the distributions of content and link similarity are so heavily skewed. However all this relevant mass is washed away in a sea of unrelated pairs, so P near the origin is negligible. This underscores that while emphasis on precision is a very reasonable approach for a search engine, it costs dearly in terms of recall.

Focusing on the precision map, for very high σ_c there is significant noise making it hard to get a clear signal from link cues. The many relevant pages in this region cannot be identified from link analysis. These may be 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, so very high σ_c is not a reliable signal. For medium-high σ_c and low σ_l we observe a surprising basin of low precision. Such an inversion likely corresponds to pairs of pages that are more semantically related than σ_s reveals, a symptom of the limitations implicit in reducing the ODP ontology to a particular tree. In the same σ_c range, precision reaches many high peaks for maximum σ_l. Here we note few pairs of highly related pages. A normative strategy suggested by these maps is to mine through pages with medium-high σ_c, then use link analysis to distill the most relevant pages.

5. CONCLUSION

Understanding how semantic information can be mined from the content and links of Web pages is key for advancing search technology. This paper reports on the first attempt to approximate semantic associations by mining content and link information from billions of pairs of Web pages. The preliminary results presented highlight the importance of appropriately combining different sources of evidence for page meaning. However, any simple combination of of σ_c and σ_l (linear or not) will result in both false positives and false negatives because of the many local optima. The approach proposed in this paper should be validated and extended by considering alternative definitions of content, link or semantic similarity, different hierarchical classifications or more general ontologies, and different cues altogether.

6. REFERENCES

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. 21st ACM SIGIR Conference*, pages 104–111, 1998.

[2] S. Chakrabarti. *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann, San Francisco, 2003.

[3] S. Chakrabarti, B. Dom, P. Raghavan, *et al.* Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks*, 30(1–7):65–74, 1998.

[4] F. Menczer. Lexical and semantic clustering by Web links. *Journal of the Am. Soc. for Information Science and Technology*, Forthcoming.