Finding semantic needles in haystacks of Web text and links[*]

Filippo Menczer
School of Informatics and Department of Computer Science
Indiana University
Bloomington, IN 47408
fil@indiana.edu

Abstract

Content and links are used to search, rank, cluster and classify Web pages. Here I analyze and visualize similarity relationships in massive Web datasets to identify how content and link analysis should be integrated for relevance approximation. Human-generated metadata from Web directories is used to estimate semantic similarity. Highly heterogeneous topical maps point to a critical dependence on search context.

Keywords: Web Mining, Web Search, Web Content and Link Analysis.

1 Introduction

When searching the Web, we want to find information relevant to a certain query. When classifying pages, we want to identify the topic that best matches the meaning of a document. When clustering, we want to group pages into collections of related documents. And when harvesting resources, we want to distill the most important pages about a particular subject. All these activities, which occupy a good part of the time we spend online, are based on notions — relevance, meaning, relatedness, aboutness — of *semantic* relationships between concepts identified by Web objects such as pages, queries, and topics. The information discovery applications we use to carry out these tasks use surrogate similarity measures stemming from the analysis of observable features — content and links — in order to estimate the semantic relationship between Web objects. For example, a search engine typically ranks hits based on some function of the keywords in the user query, the text of each Web page and pages linked to it, and the link structure of the Web. Link and content analysis are combined within the secret recipes that make or break search engines.

Despite the commercial and social impact of search engines and their ranking functions, we lack systematic studies to analyze the relationships between similarity measures based on Web content and links, and between these and semantic similarity. In this paper I review ongoing efforts to fill such a void, in particular studying relationships between pairs of Web pages. There are two

[*] Final version to appear in *IEEE Internet Computing,* May/June 2005. Copyright 2005 IEEE.

approaches to obtain measures of semantic similarity. One is by user studies in which human subjects are asked to, say, evaluate the relevance of a page to a query, or the match between a page and a topic. The second approach is to employ available human-generated metadata such as manually maintained directories and ontologies. While user studies represent a golden standard, their cost limits the number of measurements that can be obtained. For example, exhaustive mining of pairwise relationships in a tiny sample of 1000 pages would require a million evaluations due to the quadratic nature of the effort. Conversely, millions of pages are classified in broad hierarchical directories by human editors. While they do not cover a very large portion of the Web, such directories are featured prominently by major search engines such as Google and Yahoo owing to the accuracy of the human judgments on which they are based. Here I show that such directories give us a precious and readily available resource to measure semantic similarity for massively dense page matrices; the analysis of this data can generate hypotheses to be tested through more targeted user studies.

Consider two objects p and q (an object can be a page or a query, for example; let us refer to objects as pages for simplicity) and a semantic similarity function $\sigma_s(p,q)$ to establish the degree to which the meanings of the two are related. While people are good at computing σ_s, e.g. assessing relevance, a search engine must approximate this function computationally. 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. Therefore understanding the limitations of such approximations is crucial for the design of better search tools for the Web.

This paper outlines quantitative measures of relationships between content, link, and semantic topology of the Web at a fine level of resolution. The basic idea is to measure the correlations between σ_s, σ_c, and σ_l where the two latter functions are similarity metrics based on lexical content and link cues, respectively. This way one can begin to estimate the quality of cues about meaning that one can obtain from local text and link analysis, and explore whether and how σ_c and σ_l should be combined to better approximate σ_s. Another goal is to analyze the sensitivity of these relationships to the topical context of a user's information needs.

2 Sidebar: Background

Semantic similarity is generally approximated from two main classes of cues in the Web: lexical cues (textual content) and link cues (hyperlinks). Content similarity metrics traditionally used by search engines to rank hits are derived from the vector space model, 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 semantically related with high probability [16].

The latest generation of search engines integrate content and link metrics to improve ranking and crawling performance through better models of relevance. The best known example is Google: pages are retrieved based on their content and ranked based on, among other factors, the *PageRank* measure, which is computed offline by query-independent link analysis [2]. Links are also used in conjunction with text to identify hub and authority pages for a certain subject [8] and to guide smart crawling agents [15], among other applications. Finally link analysis has been applied to identify Web communities [5]. The hidden assumption behind all of these retrieval, ranking, crawling and clustering 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. This *link-cluster hypothesis* has been implied or stated in various forms (e.g., [4]). It has also been confirmed empirically by the observed decay in content similarity as one crawls away from a seed page, showing that content similarity is strongly anticorrelated with link distance [14]. That type of analysis has important limitations, however. For one, knowledge of link distance requires exhaustive breadth-first search, which makes it expensive to crawl very far from the seed page (say, more than three links away). In addition, the choice of seed pages can bias the crawl dynamics considerably; for example starting from a popular hub or authority page such as a Yahoo category will give quite different results than starting from some obscure personal homepage.

Navigation models for efficient Web crawling have provided a context for the study of functional relationships between link probability and forms of content or semantic similarity [11]. The dependence of link topology on content similarity has also been used to interpret the Web's emergent link degree distribution through local, content-driven generative models [11, 13].

Finally, several studies have related link and content similarity in the context of hypertext document classification. In particular, a text-based classifier trained on a topic taxonomy derived from a Web directory has been used to study the structure of Web topics in relationships to links, namely degree distributions within topics, topic convergence on directed walks, and link-based vs. content-based Web communities [3]. The same Web directory was used to evaluate strategies for similarity search on the Web [7]. These studies share the use of human-edited Web directories as a proxy for explicit user measurements of semantic similarity.

3 Similarity classes

One can map different classes of similarity measures and their relationships using a brute force approach, across all pairs of pages in a representative sample of the Web. A stratified sample of 150,000 pages classified in 47,174 topics across 15 top level categories of the Open Directory Project[1] (ODP) was obtained for this purpose. The ODP provides us with a hierarchical ontology to automatically estimate semantic similarity from human classification metadata which is readily available in RDF format.

For each pair of pages p, q one can measure content similarity $\sigma_c(p, q)$, link similarity $\sigma_l(p, q)$, and semantic similarity $\sigma_s(p, q)$. Too many content and link based similarity measures have been proposed in the information retrieval, machine learning, data mining and knowledge discovery literature to list here (see [6] for a partial review). While I have selected a few representative measures to test the approach, one can use this framework to evaluate any measure, or any combination of measures, and see how closely these approximate semantic similarity — whether the latter is based on user studies or automatically extracted from classification metadata.

[1] `http://dmoz.org`

The methodology and the three measures analyzed here are described in more detail elsewhere [12]. Briefly, the sample pages were crawled, their textual content and outlinks were extracted, and their inlinks were obtained via Google.[2]

As a representative content similarity measure, I considered the traditional cosine similarity based on the vector space model:

$$\sigma_c(\vec{p}, \vec{q}) = \frac{\|\vec{p} \cdot \vec{q}\|}{\|\vec{p}\| \cdot \|\vec{q}\|}. \tag{1}$$

Different weighting schemes such as binary, term frequency, and TF-IDF did not seem to have appreciable quantitative effects on the analysis that follows.

As a representative content similarity measure, I considered the simple Jaccard coefficient:

$$\sigma_l(p, q) = \frac{|U_p \cap U_q|}{|U_p \cup U_q|} \tag{2}$$

where U_p is p's undirected link neighborhood (outlinks, inlinks, and p itself). This is a measure of the local clustering between two pages; a high value of σ_l indicates that the two pages belong to a clique. This measure generalizes co-citation and co-reference in bibliometrics [1], and is related to the ways links are analyzed to identify Web communities [5].

Finally, semantic similarity can be roughly estimated using the information-theoretic measure:

$$\sigma_s(p, q) = \frac{2 \log \Pr[t_0(p, q)]}{\log \Pr[t(p)] + \log \Pr[t(q)]} \tag{3}$$

where $t(p)$ is the topic containing p in the 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 [9]. This measure relies on the existence of a hierarchical ontology that classifies all of the pages being considered, and is designed to compensate for the fact that the tree can be unbalanced both in terms of its topology and of the relative size of its nodes. For a perfectly balanced tree σ_s corresponds to the familiar tree distance measure. However, the ODP is not really a tree due to the presence of cross-reference links; the limitations of the tree assumption are discussed below, along with ongoing work to generalize this measure.

4 Correlations

The pages sampled from the ODP and crawled yielded 3.8×10^9 pairs for which the three similarities were measured. Each of the measures is defined in the unit interval; this was divided into 100 bins. Each of the resulting 10^6 bins was used to store the number of page pairs with values corresponding to the bin's $(\sigma_c, \sigma_l, \sigma_s)$ coordinates. From this information a number of interesting statistics and visual maps can be derived.

The Pearson's correlation coefficients between pairs of similarity metrics are $r(\sigma_c, \sigma_l) = 0.10$, $r(\sigma_c, \sigma_s) = 0.11$, and $r(\sigma_l, \sigma_s) = 0.08$. These are not as strong correlations as one might have

[2] http://www.google.com/apis/

Table 1: ODP top level topics and summary statistics from pairwise similarity analysis. Pearson's correlation coefficients above the all-pairs values (shown in the last row for reference) are in bold. All differences are statistically significant.

Topic	Pairs	$r(\sigma_c, \sigma_l)$	$r(\sigma_c, \sigma_s)$	$r(\sigma_l, \sigma_s)$
Adult	2.5×10^6	**0.15**	0.11	0.07
Arts	2.1×10^7	0.06	0.11	**0.09**
Business	0.6×10^7	0.06	0.08	**0.13**
Computers	2.3×10^7	0.08	**0.12**	**0.22**
Games	1.4×10^7	**0.12**	0.04	0.05
Health	2.3×10^7	0.10	**0.16**	0.07
Home	2.2×10^7	**0.50**	**0.35**	**0.16**
Kids & Teens	2.0×10^7	**0.13**	**0.16**	0.08
News	3.2×10^6	**0.39**	**0.30**	**0.48**
Recreation	1.6×10^7	0.08	**0.16**	**0.11**
Reference	2.5×10^7	**0.16**	**0.19**	**0.27**
Science	2.3×10^7	**0.11**	**0.14**	0.05
Shopping	1.9×10^7	0.06	0.11	**0.14**
Society	1.7×10^7	0.08	**0.14**	0.07
Sports	1.9×10^7	**0.14**	**0.19**	**0.09**
All-Pairs	3.8×10^9	0.10	1.11	0.08

predicted, although statistically they are very significant ($p \ll 10^{-307}$) due to the large number of pairs. The latter two r-values quantify the validity of the cluster and link-cluster hypotheses, suggesting that both content and link analysis (poorly) predict meaning. The three metrics have roughly exponential distributions peaked at zero; given two random pages one does expect them to be lexically similar, closely clustered, or semantically related. The very small number of pairs with high similarity values explains the low correlation coefficients.

More interesting are the correlations for pairs of pages within topics, shown in Table 1. All r-values are significantly positive although most are small. A few interesting exceptions are in the "Home," "News," and "Reference" categories. For example "Home" has the highest correlation between content and semantic similarity; the majority of these sites are about recipes, where terms (e.g., ingredients) are good clues about meaning. It is also reassuring that news sites do a good job linking to related pages.

5 Semantic maps

To visualize how accurately semantic similarity can be approximated from content and link cues, let us map the σ_s landscape as a function of σ_c and σ_l. There are two mappings. Averaging within a bin, i.e. normalizing local semantic similarity by the number of pairs in each bin, highlights the expected values of σ_s and is akin to precision. Summing within a bin, i.e. normalizing local semantic similarity by the total number of pairs, captures the relative mass of semantically similar pairs and is akin to recall.

The two mappings correspond to the following formal definitions of *localized* precision and recall:

$$P(s_c, s_l) = \frac{S(s_c, s_l)}{N(s_c, s_l)} \tag{4}$$

$$R(s_c, s_l) = \frac{S(s_c, s_l)}{S_{tot}} \tag{5}$$

where

$$S(s_c, s_l) = \sum_{p,q:\sigma_c(p,q)=s_c,\sigma_l(p,q)=s_l} \sigma_s(p, q) \tag{6}$$

$$N(s_c, s_l) = |p, q : \sigma_c(p, q) = s_c, \sigma_l(p, q) = s_l| \tag{7}$$

$$S_{tot} = \sum_{p,q} \sigma_s(p, q) \tag{8}$$

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

The semantic maps in Figure 1 provide for a rich account of the information about meaning that can be inferred from text and link cues. Although the majority of pairs occur near the origin (high R), all this relevance mass is washed away in a sea of unrelated pairs (low P). This highlights the challenging tradeoff between high precision and high recall; search engines typically emphasize precision because most users only view a handful of results, but the recall map suggests that much relevant information is lost that way.

Focusing on the precision map, for very high content similarity, there is significant noise making it difficult to get a clear signal from link cues. There are many related pages in this region, but they look like needles in a haystack of unrelated pages and are not detected by link analysis. Maybe these are cases where authors do not know of related pages or do not want to point to the competition. So very high content similarity is not a reliable signal.

A surprising basin of low precision can be seen for medium-high content similarity and low link similarity. One explanation originates from recalling that the semantic similarity measure is based on a tree view of the ODP ontology. It is possible that the basin corresponds to pairs of pages that are in reality more semantically related than the measure reveals. For example if one page about email marketing is classified under the "Computers" branch while another is under the "Business" branch, $\sigma_s = 0$ even though the two are semantically related. Thus such a basin may be symptomatic of the limitations of any particular tree ontology.

The highest precision is achieved in the same medium-high content similarity range, but for maximum link similarity. Here there are very few pairs, but those few correspond to highly related pages. This region seems the most ripe for information discovery, yet it requires a nonlinear combination of content and link similarity.

To analyze the generality of the above observations, the brute force analysis was repeated for each top-level ODP topic. Figures 2 and 3 show the 15 topical recall and precision maps, respectively. There is indeed a significant level of heterogeneity in these maps, both qualitatively and quantitatively.

Looking at semantic recall maps, the topics that display higher correlation coefficients are those for which non-zero recall values extend further away from the origin toward high content and link

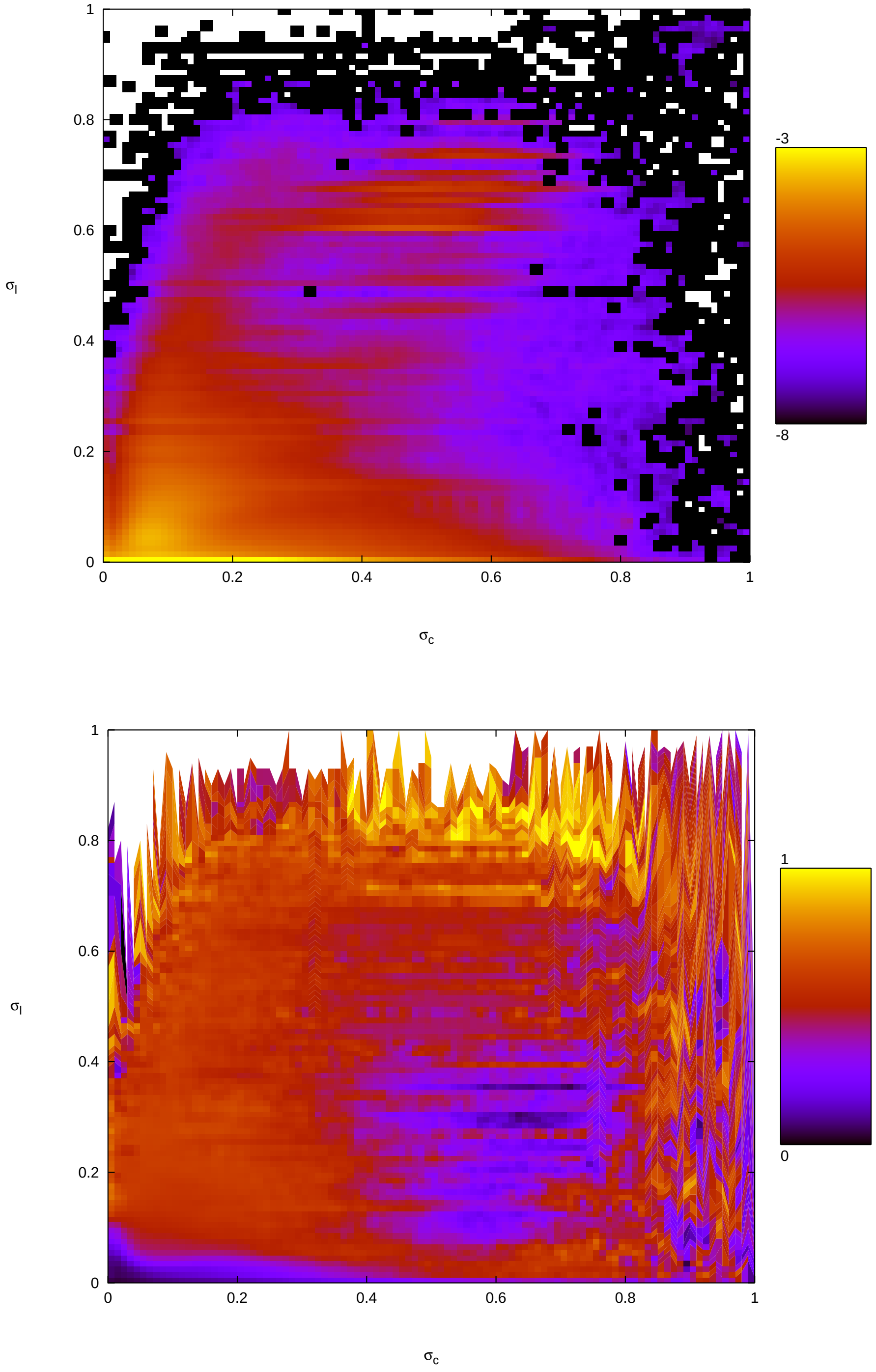

Figure 1: Semantic maps of localized recall (top) and precision (bottom) for all pairs of sample Web pages. Colors represent the number of pairs in each bin. White represents missing data (no pairs). R values span 8 orders of magnitude in the unit interval and for readability is visualized on a logarithmic scale between 10^{-8} and 10^{-3} or above.

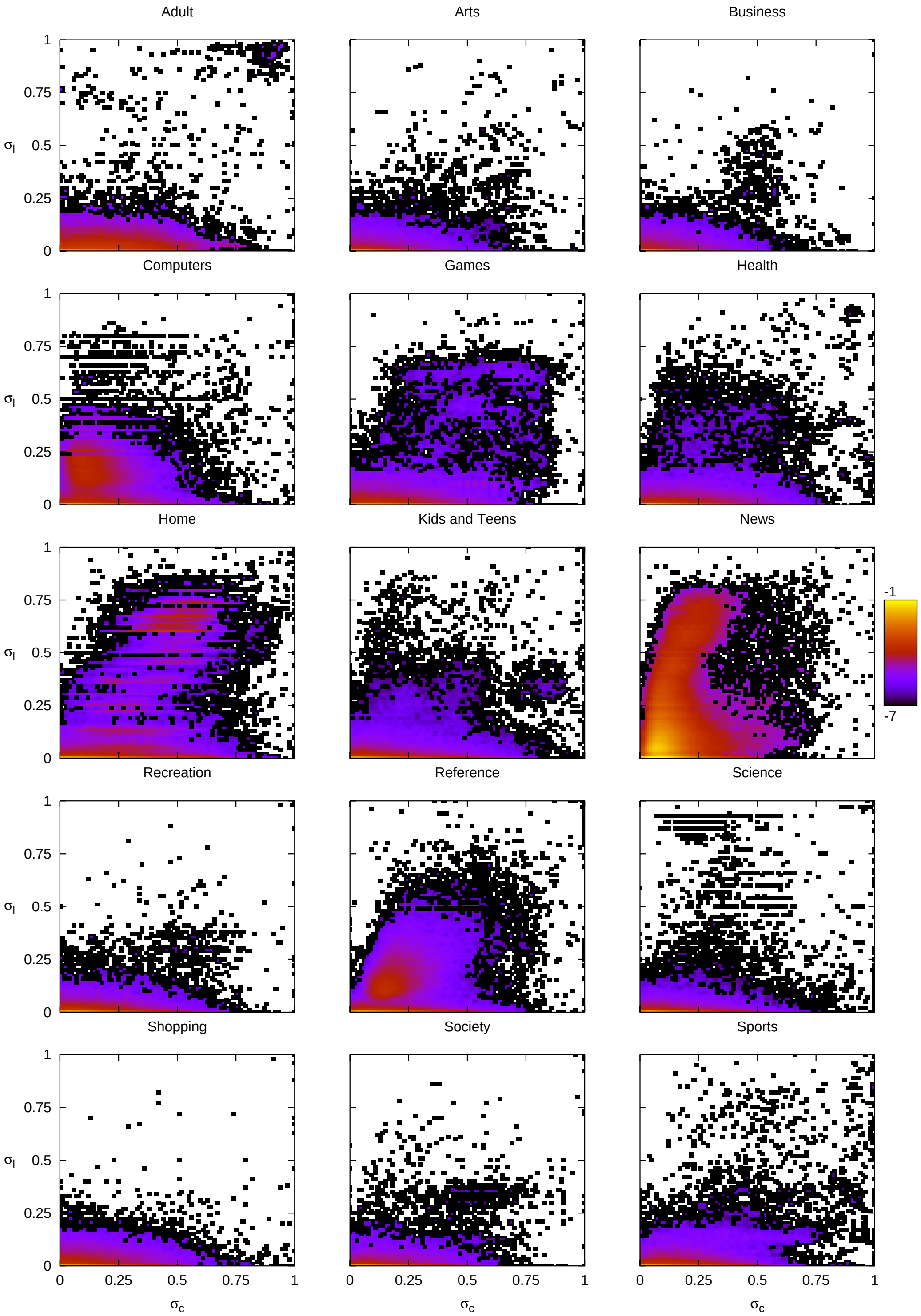

Figure 2: Semantic recall maps for each top-level topic. Recall is visualized on a log color scale between 10^{-7} and 10^{-1}.

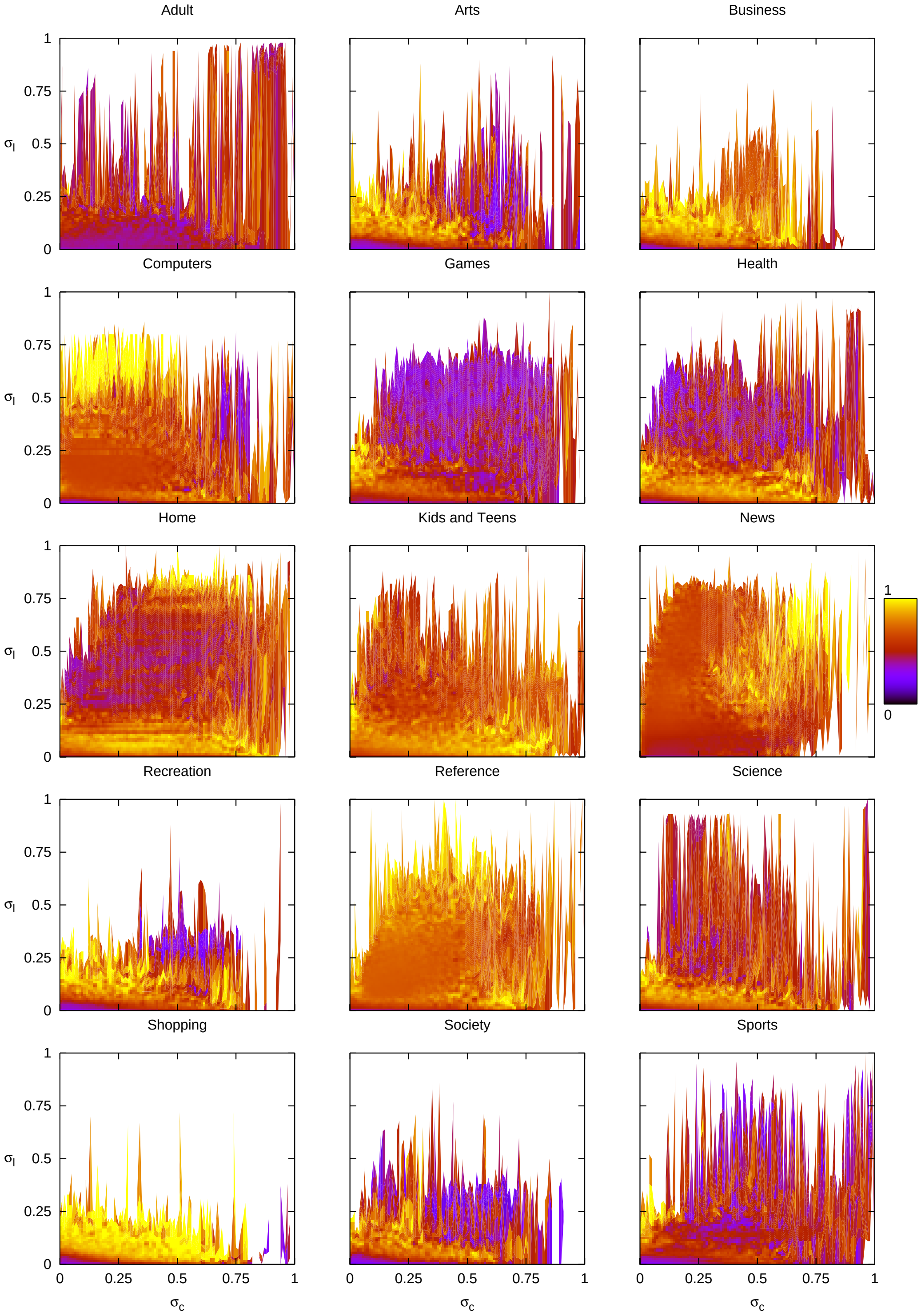

Figure 3: Semantic precision maps for each top-level topic.

similarity. One exception is the lonely peak in the top right corner of the "Adult" recall map. This is explained as a clique of *spammer* sites designed to fool the ranking algorithms of search engines by boosting content similarity and PageRank. But in topics such as "Home," "News," and "Reference" it is clear that semantic similarity is correlated with both content and link similarity, and therefore text and links are informative cues about the meaning of pages. No topic displays the negative basin of the general recall map, reinforcing the hypothesis that the basin is due to cross-topic ($\sigma_s = 0$) pairs.

Semantic precision maps display even greater heterogeneity, and differ significantly from the general precision map. With the exception of "Adult," all topics have visible regions of high precision (in yellow) with various sizes, shapes, and locations. A couple of topics — "Computers" and "News" — have large, well localized high precision regions. These highlight the diverse semantic inferences that can be drawn from text and link cues depending on topical context.

6 Combining content and link similarity

In information retrieval the effectiveness of a document ranking system can be assessed, if the relevant set is known, using 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.[3] However, in "query by example" retrieval systems, pages are used as queries. Let us follow this approach, considering each page in our sample as an example and ranking all other pages, then using semantic similarity estimates to assess the ranking.

Let us define *functional* precision and recall as follows:

$$P(f, \beta) = \frac{\displaystyle\sum_{p,q: f(\sigma_c(p,q), \sigma_l(p,q)) \geq \beta} \sigma_s(p, q)}{|p, q : f(\sigma_c(p, q), \sigma_l(p, q)) \geq \beta|} \tag{9}$$

$$R(f, \beta) = \frac{\displaystyle\sum_{p,q: f(\sigma_c(p,q), \sigma_l(p,q)) \geq \beta} \sigma_s(p, q)}{\displaystyle\sum_{p,q} \sigma_s(p, q)} \tag{10}$$

where f is some function of σ_c and σ_l that expresses how content and link similarity are to be combined in order to estimate semantic similarity. Pairs are then ranked by f. The threshold β is used as an independent rank parameter, which allows to compute P and R for an increasing number of pairs (as β decreases).

The simplest way to combine content and link similarity is a linear combination as expressed by the function $f(\sigma_c, \sigma_l) = \alpha\sigma_l + (1 - \alpha)\sigma_c$. The special cases $\alpha = 0$ and $\alpha = 1$ correspond to pure content-based and link-based ranking, respectively. Content and link similarity can also be combined in non-linear ways, for example by the product $f(\sigma_c, \sigma_l) = \sigma_c\sigma_l$ where pairs are favored if content and link similarity are both high.

[3] Even large user studies can only identify subsets of relevant pages; there are simply too many pages to consider.

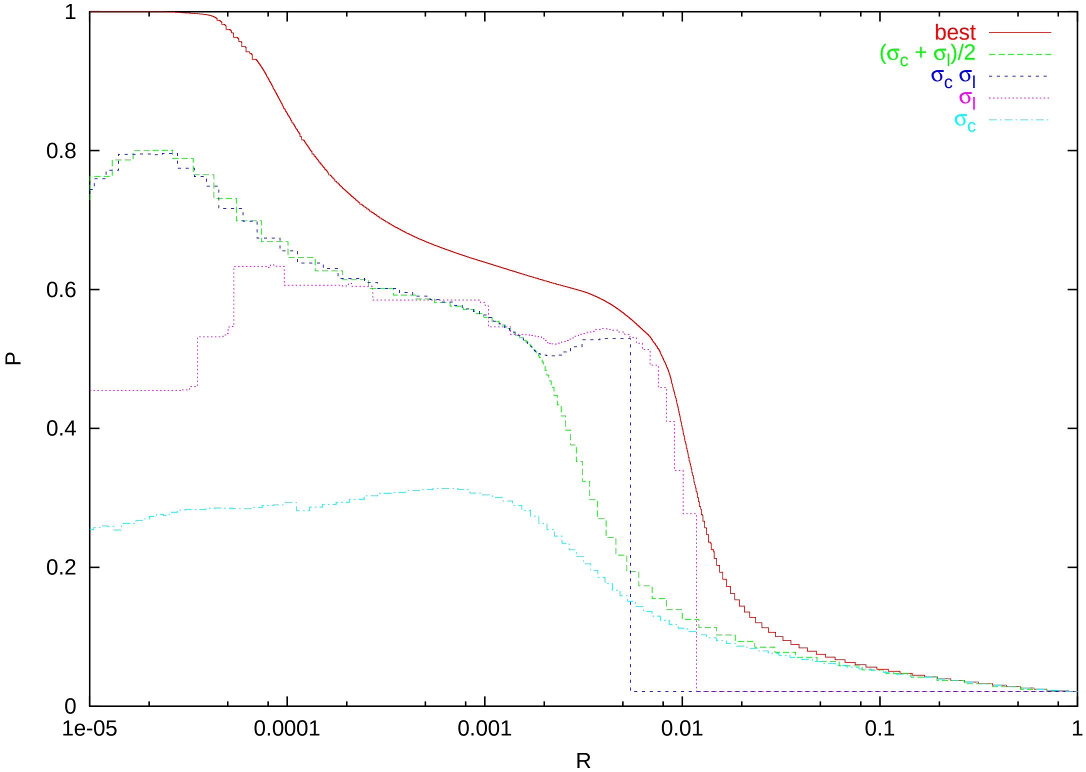

Figure 4: Functional precision-recall plots for rankings based on various combinations of content and link similarity. For comparison the best performance achievable is obtained by ranking (σ_c, σ_l) bins by their *localized* precision.

The functional precision-recall plots in Figure 4 are based on a few sample combinations of link and content similarity. Since most of the σ_s "mass" occurs near the origin ($\sigma_c = \sigma_l = 0$) and the most interesting region is that of low recall (the "top hits"), recall is better visualized on a logarithmic scale. The plots based on just one cue show that ranking by link similarity produces better precision at low recall levels, while ranking by content similarity produces better precision at high recall levels. Any combination of content and link similarity yields better compromise rankings, even though the type of combination (linear combination or product) makes little difference at low recall. This suggests that if a search engine could efficiently rank pages based on contextual link analysis (rather than query-independent link analysis, e.g. PageRank), they could achieve significantly higher precision at low recall and thus place more relevant hits in the first page.

To study how well different combinations of content and link similarity rank pages within different topical contexts, the analysis was repeated for each top-level ODP topic. The plots[4] confirm that the topics with highest similarity correlations are both easier (in that the optimal precision-recall curves are very high) and more amenable to information discovery (in that content and link analysis best approximate optimal performance). Yet different topics require different strategies; e.g., link similarity is best for "Computers" while combinations work best for "News" and all functions perform equally poorly for "Sports."

[4] See extended tech. rep.: `http://informatics.indiana.edu/research/publications/publications.asp?id=17`

7 Discussion

Understanding how semantic information can be mined from the content and links of Web pages is key for advancing search technology. The framework outlined here represents the first large-scale, brute-force effort to map semantic association in a topological space generated from content and link based metrics. A large amount of data collected from billions of pairs of Web pages has been used to build a number of semantic maps that visualize the relationship between content and link cues and the inferences about page meaning that they allow. While this paper focuses on retrieval and ranking, the framework can also provide us with useful insight for crawling and clustering applications.

This empirical study underlines the limits of "one-size-fits-all" solutions in Web information retrieval. No single content/link analysis technique will work best in every user's information context. While quantitatively a novel observation, this is not surprising and underscores the importance of moving from universal search to specialized search tools; Google has started testing personalized search and site-flavored search in recent months. The context of individual users must be harnessed to build user-centric models of semantic relationships between pages. For example, users who use bookmark managers to store important URLs have a semantic hierarchy readily available for local analysis and personalization of search results. This is one of the directions that we are beginning to explore within a collaborative peer crawling and searching framework.

Another contribution is the quantification of the intrinsic difficulty of search problems in different topical contexts. Achieving strong search/clustering performance in topics such as "Games" is significantly harder than in topics such as "News." It should therefore come as no surprise that news are the first topical domain successfully tackled by large search engines, first Google[5] and more recently Yahoo![6] — the easy problems are attacked first. The research presented here points to future targets of opportunity for research in Web IR. Cooking recipes might be next!

We have only scratched the surface with respect to the Web regularities that can be discovered with the semantic map approach. One can obtain different maps by considering different definitions of content, link or semantic similarity, different hierarchical classifications or more general ontologies, or different cues altogether.

A massive data mining effort is under way, in which we are considering all pairs of ODP pages rather than just a sample. Managing the resulting terabytes of similarity data ($\sim 10^6$ pages, $\sim 10^{13}$ pairs) imposes significant infrastructure demands. More importantly, we are generalizing the semantic similarity measure of Equation 3 to deal with general (non-tree) graphs; a few empirical measures have been proposed but no information-theoretic measures are found in the literature. Applying this new measure to the full ODP ontology is a non-trivial computational endeavor requiring weeks of runtime on IU's Linux clusters (208 2.4 GHz processors). The aim is to better approximate the golden standard of semantic similarity, i.e. user assessments. Preliminary experiments with human subjects validate the new measure [10]. We will then be in a position to evaluate arbitrary measures that integrate (rather than combine) content and link similarity analysis in a context-sensitive manner, and apply them to Web information discovery tasks such as the TREC Terabyte

[5] http://news.google.com/

[6] http://news.yahoo.com/

track.

Acknowledgements

A. Vespignani, A. Maguitman, and H. Roinestad are collaborating on the next phase of this project; N. Sanghvi has helped with data collection. I am grateful to S. Chakrabarti, J. Kleinberg, L. Adamic, P. Srinivasan, N. Street, and A. Segre for helpful comments on earlier drafts of this paper. Thanks to the ODP for its public RDF data and to Google for extensive use of its Web API. This work is funded in part by NSF CAREER Grant No. IIS-0348940.

References

[1] K Börner, C Chen, and KW Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37:179–255, 2003.

[2] S Brin and L Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[3] S Chakrabarti, MM Joshi, K Punera, and DM Pennock. The structure of broad topics on the Web. In David Lassner, Dave De Roure, and Arun Iyengar, editors, *Proc. 11th International World Wide Web Conference*, pages 251–262, New York, NY, 2002. ACM Press.

[4] BD Davison. Topical locality in the Web. In *Proc. 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–279, 2000.

[5] GW Flake, S Lawrence, CL Giles, and FM Coetzee. Self-organization of the Web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[6] P Ganesan, H Garcia-Molina, and J Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.

[7] TH Haveliwala, A Gionis, D Klein, and P Indyk. Evaluating strategies for similarity search on the Web. In David Lassner, Dave De Roure, and Arun Iyengar, editors, *Proc. 11th International World Wide Web Conference*, New York, NY, 2002. ACM Press.

[8] J Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[9] D Lin. An information-theoretic definition of similarity. In Jude Shavlik, editor, *Proc. 15th Intl. Conference on Machine Learning*, pages 296–304, San Francisco, CA, 1998. Morgan Kaufmann.

[10] Ana Gabriela Maguitman, Filippo Menczer, Heather Roinestad, and Alessandro Vespignani. Algorithmic detection of semantic similarity. In *Proc. 14th International World Wide Web Conference*, 2005. Forthcoming.

[11] F Menczer. Growing and navigating the small world Web by local content. *Proc. Natl. Acad. Sci. USA*, 99(22):14014–14019, 2002.

[12] F Menczer. Correlated topologies in citation networks and the web. *European Physical Journal B*, 38(2):211–221, 2004.

[13] F Menczer. The evolution of document networks. *Proc. Natl. Acad. Sci. USA*, 101:5261–5265, 2004.

[14] F Menczer. Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269, 2004.

[15] F Menczer, G Pant, and P Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 4(4):378–419, 2004.

[16] CJ van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979. Second edition.