Algorithmic Detection of Semantic Similarity

Ana G. Maguitman†‡
anmaguit@cs.indiana.edu

Filippo Menczer†‡
fil@indiana.edu

Heather Roinestad†
hroinest@cs.indiana.edu

Alessandro Vespignani‡
alexv@indiana.edu

† Department of Computer Science
‡ School of Informatics
Indiana University
Bloomington, IN 47408

ABSTRACT

Automatic extraction of semantic information from text and links in Web pages is key to improving the quality of search results. However, the assessment of automatic semantic measures is limited by the coverage of user studies, which do not scale with the size, heterogeneity, and growth of the Web. Here we propose to leverage human-generated metadata — namely topical directories — to measure semantic relationships among massive numbers of pairs of Web pages or topics. The Open Directory Project classifies millions of URLs in a topical ontology, providing a rich source from which semantic relationships between Web pages can be derived. While semantic similarity measures based on taxonomies (trees) are well studied, the design of well-founded similarity measures for objects stored in the nodes of arbitrary ontologies (graphs) is an open problem. This paper defines an information-theoretic measure of semantic similarity that exploits both the hierarchical and non-hierarchical structure of an ontology. An experimental study shows that this measure improves significantly on the traditional taxonomy-based approach. This novel measure allows us to address the general question of how text and link analyses can be combined to derive measures of relevance that are in good agreement with semantic similarity. Surprisingly, the traditional use of text similarity turns out to be ineffective for relevance ranking.

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)*

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.

General Terms

Measurement, Experimentation.

Keywords

Web mining, Web search, semantic similarity, content and link similarity, ranking evaluation.

1. INTRODUCTION

Developing Web search mechanisms depends on addressing two central questions: (1) how to find related Web pages, and (2) given a set of potentially related Web pages, how to rank them according to relevance. To evaluate the effectiveness of a Web search mechanism in finding and ranking results, measures of semantic similarity are needed. In traditional approaches users provide manual assessments of relevance, or semantic similarity. This is difficult and expensive. More importantly, it does not scale with the size, heterogeneity, and growth of the Web — subjects can evaluate sets of queries, but cannot cover exhaustively all topics.

The Open Directory Project[1] (ODP) is a large human-edited directory of the Web, employed by hundreds of portals and search sites including Google. The ODP classifies millions of URLs in a topical ontology. Ontologies help to make sense out of a set of objects. Once the meaning of a set of objects is available, it can be usefully exploited to derive semantic relationships between those objects. Therefore, the ODP provides a rich source from which measurements of semantic similarity between Web pages can be obtained.

An ontology is a special kind of network. The problem of evaluating semantic similarity in a network has a long history in psychological theory [22]. More recently, semantic similarity became fundamental in knowledge representation where special kinds of networks or ontologies are used to describe objects and their relationships [6].

Many proposals estimate semantic similarity in a network representation by computing distance between the nodes. These frameworks are based on the premise that the closer the semantic relationship of two objects, the closer they will be in the network representation. However, as it has

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

been discussed by a number of sources, issues arise when attempting to apply distance-based schemes for measuring object similarities in certain classes of networks where links may not represent uniform distances [19].

In ontologies, certain links connect very dense and general categories while others connect more specific ones. To address this problem, some proposals estimate semantic similarity in a taxonomy based on the notion of information content [19, 12]. In these approaches, the semantic similarity between two objects is related to their commonality and to their differences. Given a set of objects in an "is-a" taxonomy, the commonality of two objects can be estimated by the extent to which they share information, indicated by the most specific class in the hierarchy that subsumes both. The meaning of the individual objects can be measured by looking at the classes rooted at each of the topics.

Ontologies are often equated with "is-a" taxonomies, but ontologies need not be limited to these forms. For example, the ODP ontology is more complex than a simple tree. Some categories have multiple criteria to classify subcategories. The "Business" category, for instance, is subdivided by types of organizations (cooperatives, small businesses, major companies, etc.) as well as by areas (automotive, health care, telecom, etc.). Furthermore, the ODP has various types of cross-reference links between categories, so that a node may have multiple parent nodes, and even cycles are present.

While semantic similarity measures based on trees are well studied [5], the design of well-founded similarity measures for objects stored in the nodes of arbitrary graphs is an open problem. A few empirical measures have been proposed, for example based on minimum cut/maximum flow algorithms [13], but no information-theoretic measure is known. The central question addressed in this paper is how to estimate semantic similarity in generalized ontologies, such as the ODP graph, taking advantage of both their hierarchical ("is-a" links) and non-hierarchical (cross links) components.

1.1 Contributions and Outline

In the next section we introduce a novel graph-based measure of semantic similarity. To the best of our knowledge this is the first information-theoretic measure of similarity that is applicable to objects stored in the nodes of arbitrary graphs, in particular topical ontologies and Web directories that combine hierarchical and non-hierarchical components such as Yahoo!, ODP and their derivatives.

Section 3 compares the graph-based semantic similarity measure to the tree-based one, analyzing the differences between the two measurements and presenting an evaluation against human judgments of Web page similarity. We show that the new measure predicts human responses to a much greater accuracy.

Having validated the proposed semantic similarity measure, in Section 4 we begin to explore the question of applications, namely how text and link analyses can be combined to derive measures of relevance that are in good agreement with semantic similarity. We consider various combinations of text and link similarity and discuss how these correlate with semantic similarity and how well they rank pages. We find that surprisingly, classic text-based content similarity is a very noisy feature, whose value is at best weakly correlated. We discuss the potential applications of this result to the design of semantic similarity estimates from lexical and link similarity, and to the optimization of ranking functions in search engines.

2. SEMANTIC SIMILARITY

2.1 Tree-Based Similarity

Lin [12] has investigated an information theoretic definition of similarity that is applicable as long as the domain has a probabilistic model. This proposal can be used to derive a measure of semantic similarity between topics in an "is-a" taxonomy.

According to Lin's proposal, the semantic similarity between two topics in a taxonomy is defined as a function of the meaning shared by the topics and the meaning of each of the individual topics. In a taxonomy, the meaning shared by two topics can be recognized by looking at the lowest common ancestor, which corresponds to the most specific common classification of the two topics. Once this common classification is identified, the meaning shared by two topics can be measured by the amount of information needed to state the commonality of the two topics. Likewise, the meaning of each of the individual topics is measured by the amount of information needed to fully describe each of the two topics.

In information theory [3], the information content of a class or topic t is measured by the negative log likelihood, $-\log \Pr[t]$. The semantic similarity between two topics t_1 and t_2 in a taxonomy is then measured as the ratio between their common meaning and their individual meanings as follows:

$$\sigma_s^T(t_1, t_2) = \frac{2 \cdot \log \Pr[t_0(t_1, t_2)]}{\log \Pr[t_1] + \log \Pr[t_2]}$$

where $t_0(t_1, t_2)$ is the lowest common ancestor topic for t_1 and t_2 in the tree, and $\Pr[t]$ represents the prior probability that any page is classified under topic t. Given a document d classified in a topic taxonomy, we use $t(d)$ to refer to the topic node containing d. Given two documents d_1 and d_2 in a topic taxonomy the semantic similarity between them is estimated as $\sigma_s^T(t(d_1), t(d_2))$. To simplify notation, we use $\sigma_s^T(d_1, d_2)$ as a shorthand for $\sigma_s^T(t(d_1), t(d_2))$. From here on, we will refer to measure σ_s^T as the tree-based semantic similarity. The tree-based semantic similarity measure for a simple taxonomy is illustrated in Figure 1. In this example, documents d_1 and d_2 are contained in topics t_1 and t_2 respectively, while topic t_0 is their lowest common ancestor. In practice $\Pr[t]$ can be computed offline for every topic t in the ODP by counting the fraction of pages stored in the subtree rooted at node t ($subtree(t)$), out of all the pages in the tree.

This measure of semantic similarity has several desirable properties and a solid theoretical justification. It is a straightforward extension of the information-theoretic similarity measure [12], 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^T corresponds to the familiar tree distance measure [10].

In prior work [14, 15, 16] we computed the σ_s^T measure for all pairs of pages in a stratified sample of about 150,000 pages from across the ODP. For each of the resulting 3.8×10^9 pairs we also computed text and link similarity measures, and mapped the correlations between these and semantic similarity. An interesting result was that these correlations

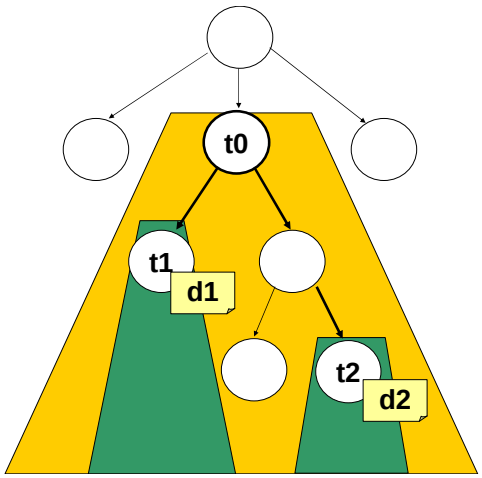

Figure 1: Illustration of tree-based semantic similarity in a taxonomy.

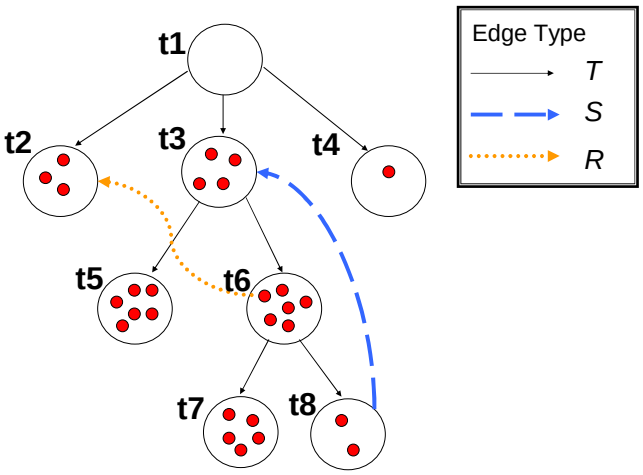

Figure 2: Illustration of a simple ontology.

were quite weak across all pairs, but became significantly stronger for pages within certain top level categories such as "news" and "reference." However, because σ_s^T is defined only in terms of the hierarchical component of the ODP, it fails to capture many semantic relationships induced by the ontology's non-hierarchical components (symbolic and related links). As a result, the tree-based semantic similarity between pages in topics that belong to different top-level categories is zero even if the topics are clearly related. This yielded an unreliable picture when all topics were considered.

2.2 Graph-Based Similarity

Let us now generalize the semantic similarity measure to deal with arbitrary graphs. We wish to define a graph-based semantic similarity measure σ_s^G that generalizes the tree-based similarity σ_s^T to exploit both the hierarchical and non-hierarchical components of an ontology.

A topic ontology graph is a graph of nodes representing topics. Each node contains objects representing documents (pages). An ontology graph has a hierarchical (tree) component made by "is-a" links, and a non-hierarchical component made by cross links of different types.

For example, the ODP ontology is a directed graph $G = (V, E)$ where:

- V is a set of nodes, representing topics containing documents;
- E is a set of edges between nodes in V, partitioned into three subsets T, S and R, such that:
  - T corresponds to the hierarchical component of the ontology,
  - S corresponds to the non-hierarchical component made of "symbolic" cross-links,
  - R corresponds to the non-hierarchical component made of "related" cross-links.

Figure 2 shows a simple example of an ontology graph G. This is defined by the sets $V = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$, $T = \{(t_1, t_2), (t_1, t_3), (t_1, t_4), (t_3, t_5), (t_3, t_6), (t_6, t_7), (t_6, t_8)\}$, $S = \{(t_8, t_3)\}$, and $R = \{(t_6, t_2)\}$. In addition, each node $t \in V$ contains a set of objects. We use $|t|$ to refer to the number of objects stored in node t (e.g, $|t_3| = 4$).

The extension of σ_s^T to an ontology graph raises two questions. First, how to find the most specific common ancestor of a pair of topics in a graph; second, how to extend the definition of subtree rooted at a topic for the graph case.

An important distinction between taxonomies and ontologies such as the ODP graph is that edges in a taxonomy are all of the same type ("is-a" links), while in the ODP graph edges can have diverse types (e.g., "is-a", "symbolic", "related"). Different types of edges have different meanings and should be used accordingly. One way to distinguish the role of different edges is to assign them weights, and to vary these weights according to the edge's type. The weight $w_{ij} \in [0, 1]$ for an edge between topic t_i and t_j can be interpreted as an explicit measure of the degree of membership of t_j in the family of topics rooted at t_i. The weight setting we have adopted for the edges in the ODP graph is as follows: $w_{ij} = \alpha$ for $(i, j) \in T$, $w_{ij} = \beta$ for $(i, j) \in S$, and $w_{ij} = \gamma$ for $(i, j) \in R$. We set $\alpha = \beta = 1$ because symbolic links seem to be treated as first-class taxonomy ("is-a") links in the ODP Web interface. Since duplication of URLs is disallowed, symbolic links are a way to represent multiple memberships, for example the fact that the pages in topic "Society/Issues/Fraud/Internet" also belong to topic "Computers/Internet/Fraud." On the other hand, we set $\gamma = 0.5$ because related links are treated differently in the ODP Web interface, labeled as "see also" topics. Intuitively the semantic relationship is weaker. Different weighting schemes could be explored.

As a starting point, let $w_{ij} > 0$ if and only if there is an edge of some type between topics t_i and t_j. However, to estimate topic membership, transitive relations between edges should also be considered. Let $t_i\downarrow$ be the family of topics t_j such that either $i = j$ or there is a path $(e_1, \ldots, e_n)$ satisfying:

1. $e_1 = (t_i, t_k)$ for some $t_k \in V$,
2. $e_n = (t_k, t_j)$ for some $t_k \in V$,
3. $e_k \in T \cup S \cup R$ for $k = 1 \ldots n$,
4. $e_k \in S \cup R$ for at most one k.

The above conditions express that $t_j \in t_i\downarrow$ if there is a directed path in the graph G from t_i to t_j, where at most one edge from S or R participates in the path. The motivation for disregarding multiple non-hierarchical links in the transitive relations that determine topic membership is both practical and conceptual. From a computational perspective, allowing multiple cross links is infeasible because it leads to a dense topic membership, i.e., every topic belongs to almost every other topic. This is also not robust because a few unreliable cross links make significant global changes to the membership functions. More importantly, considering multiple cross links in each path would make the classification meaningless by mixing all topics together. Considering at most one cross link in each membership path allows us to capture the non-hierarchical components of the ontology while preserving feasibility, robustness, and meaning. We refer to $t_i\downarrow$ as the *cone* of topic t_i. Because edges may be associated with different weights, different topics t_j can have different degree of membership in $t_i\downarrow$.

In order to make the implicit membership relations explicit, we represent the graph structure by means of adjacency matrices and apply a number of operations to them. A matrix $\mathbf{T}$ is used to represent the hierarchical structure of an ontology. Matrix $\mathbf{T}$ codifies edges in T, augmented with 1s on the diagonal:

$$\mathbf{T}_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \alpha & \text{if } i \neq j \text{ and } (i,j) \in T, \\ 0 & \text{otherwise.} \end{cases}$$

We use additional adjacency matrices to represent the non-hierarchical components of an ontology. For the case of the ODP graph, a matrix $\mathbf{S}$ is defined so that $\mathbf{S}_{ij} = \beta$ if $(i,j) \in S$ and $\mathbf{S}_{ij} = 0$ otherwise. A matrix $\mathbf{R}$ is defined analogously, as $\mathbf{R}_{ij} = \gamma$ if $(i,j) \in R$ and $\mathbf{R}_{ij} = 0$ otherwise. Consider the operation $\vee$ on matrices, defined as $[A \vee B]_{ij} = \max(A_{ij}, B_{ij})$, and let $\mathbf{G} = \mathbf{T} \vee \mathbf{S} \vee \mathbf{R}$. Matrix $\mathbf{G}$ is the adjacency matrix of graph G augmented with 1s on the diagonal.

We will use the MaxProduct fuzzy composition function $\odot$ [8] defined on matrices as follows:[2]

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = \max_k (\mathbf{A}_{ik} \cdot \mathbf{B}_{kj}).$$

Let $\mathbf{T}^{(0)} = \mathbf{T}$ and $\mathbf{T}^{(r+1)} = \mathbf{T}^{(0)} \odot \mathbf{T}^{(r)}$. We define the closure of $\mathbf{T}$, denoted $\mathbf{T}^+$ as follows:

$$\mathbf{T}^+ = \lim_{r \to \infty} \mathbf{T}^{(r)}.$$

In this matrix, $\mathbf{T}^+_{ij} = 1$ if $t_j \in \mathit{subtree}(t_i)$, and $\mathbf{T}^+_{ij} = 0$ otherwise. Note that the computation of the closure $\mathbf{T}^+$ converges in a number of steps which is bounded by the maximum depth of the tree $\mathbf{T}$, is independent of the weight α, and does not involve the weights β and γ.

Finally, we compute the matrix $\mathbf{W}$ as follows:

$$\mathbf{W} = \mathbf{T}^+ \odot \mathbf{G} \odot \mathbf{T}^+.$$

The element $\mathbf{W}_{ij}$ can be interpreted as a fuzzy membership value of topic t_j in the cone $t_i\downarrow$, therefore we refer to $\mathbf{W}$ as the *fuzzy membership matrix* of G.

[2]With our choice of weights, MaxProduct composition is equivalent to MaxMin composition.

As an illustration, consider the example ontology in Figure 2. In this case the matrices $\mathbf{T}$, $\mathbf{G}$, $\mathbf{T}^+$ and $\mathbf{W}$ are defined as follows:

$$\mathbf{T} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\mathbf{G} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\mathbf{T}^+ = \begin{array}{c} \\ \mathit{subtree}(t_1) \\ \mathit{subtree}(t_2) \\ \mathit{subtree}(t_3) \\ \mathit{subtree}(t_4) \\ \mathit{subtree}(t_5) \\ \mathit{subtree}(t_6) \\ \mathit{subtree}(t_7) \\ \mathit{subtree}(t_8) \end{array} \begin{array}{c} \begin{array}{cccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$\mathbf{W} = \begin{array}{c} \\ t_1\downarrow \\ t_2\downarrow \\ t_3\downarrow \\ t_4\downarrow \\ t_5\downarrow \\ t_6\downarrow \\ t_7\downarrow \\ t_8\downarrow \end{array} \begin{array}{c} \begin{array}{cccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .5 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & .5 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

The semantic similarity between two topics t_1 and t_2 in an ontology graph can now be estimated as follows:

$$\sigma_s^G(t_1, t_2) = \max_k \frac{2 \cdot \min(\mathbf{W}_{k1}, \mathbf{W}_{k2}) \cdot \log \Pr[t_k]}{\log(\Pr[t_1|t_k] \cdot \Pr[t_k]) + \log(\Pr[t_2|t_k] \cdot \Pr[t_k])}.$$

The probability $\Pr[t_k]$ represents the prior probability that any document is classified under topic t_k and is computed as:

$$\Pr[t_k] = \frac{\sum_{t_j \in V} (\mathbf{W}_{kj} \cdot |t_j|)}{|U|},$$

where $|U|$ is the number of documents in the ontology. The posterior probability $\Pr[t_i|t_k]$ represents the probability that any document will be classified under topic t_i given that it is classified under t_k, and is computed as follows:

$$\Pr[t_i|t_k] = \frac{\sum_{t_j \in V} (\min(\mathbf{W}_{ij}, \mathbf{W}_{kj}) \cdot |t_j|)}{\sum_{t_j \in V} (\mathbf{W}_{kj} \cdot |t_j|)}.$$

The proposed definition of σ_s^G is a generalization of σ_s^T. In the special case when G is a tree (i.e., $S = R = \emptyset$), then $t_i\downarrow$ is equal to $subtree(t_i)$, the topic subtree rooted at t_i, and all topics $t \in subtree(t_i)$ belong to $t_i\downarrow$ with a degree of membership equal to 1. If t_k is an ancestor of t_1 and t_2 in a taxonomy, then $\min(\mathbf{W}_{k1}, \mathbf{W}_{k2}) = 1$ and $\Pr[t_i|t_k] \cdot \Pr[t_k] = \Pr[t_i]$ for $i = 1, 2$. In addition, if there are no cross-links in G, the topic t_k whose index k maximizes $\sigma_s^G(t_1, t_2)$ corresponds to the lowest common ancestor of t_1 and t_2.

3. EVALUATION

The proposed graph-based semantic similarity measure was applied to the ODP ontology. The portion of the ODP graph we have used for our analysis consists of more than half million topic nodes (only *World* and *Regional* categories were discarded). Computing semantic similarity for each pair of nodes in such a huge graph required more than 5,000 CPU hours on IU's Analysis and Visualization of Instrument-Driven Data (AVIDD) supercomputer facility. The computational component of AVIDD consists of two clusters, each with 208 Prestonia 2.4-GHz processors. The computed graph-based semantic similarity measurements in compressed format occupies more than 1 TB of IU's Massive Data Storage System. After computing the graph-based semantic similarity, we dynamically computed the less computationally expensive tree-based semantic similarity on the same ODP topic pairs.

3.1 Analysis of Differences

The first question to ask of the newly proposed graph-based semantic similarity definition is whether it produces different measurements from the traditional tree-based similarity. The two measures are moderately correlated (Pearson coefficient $r_P = 0.51$). To dig deeper, we map in Figure 3 the distributions of similarities. Each (σ_s^T, σ_s^G) coordinate encodes how many pairs of pages in the ODP have semantic similarities falling in the corresponding bin. By definition σ_s^T is a lower bound for σ_s^G. Significant numbers of pairs yield $\sigma_s^G > \sigma_s^T$, indicating that the graph-based measure indeed captures semantic relationships that are missed by the tree-based measure. The largest difference is hard to observe in the map because it occurs in the $\sigma_s^T = 0$ bins. Here there are many pairs in different top-level categories of the ODP, which are related according to non-hierarchical links.

To better quantify the differences between σ_s^T and σ_s^G, Figure 3 also shows the average graph-based similarity $\langle\sigma_s^G\rangle$ as a function of σ_s^T. The relative difference is as large as 20% around $\sigma_s^T = 0.32$. The inset highlights the largest difference, which occurs for $\sigma_s^T = 0$.

3.2 Validation by User Study

Knowing that tree-based and graph-based measures give us quantitatively different estimates of semantic similarity, we conducted a human-subjects experiment to evaluate the proposed graph-based measure σ_s^G. As a baseline for comparison we used Lin's tree-based measure σ_s^T. The goal of this experiment was to contrast the predictions of the two semantic similarity measures against human judgments of Web pages relatedness.

Thirty-eight volunteer subjects were recruited for a 30 minute experiment conducted online. Subjects answered 30 questions about similarity between Web pages. For each

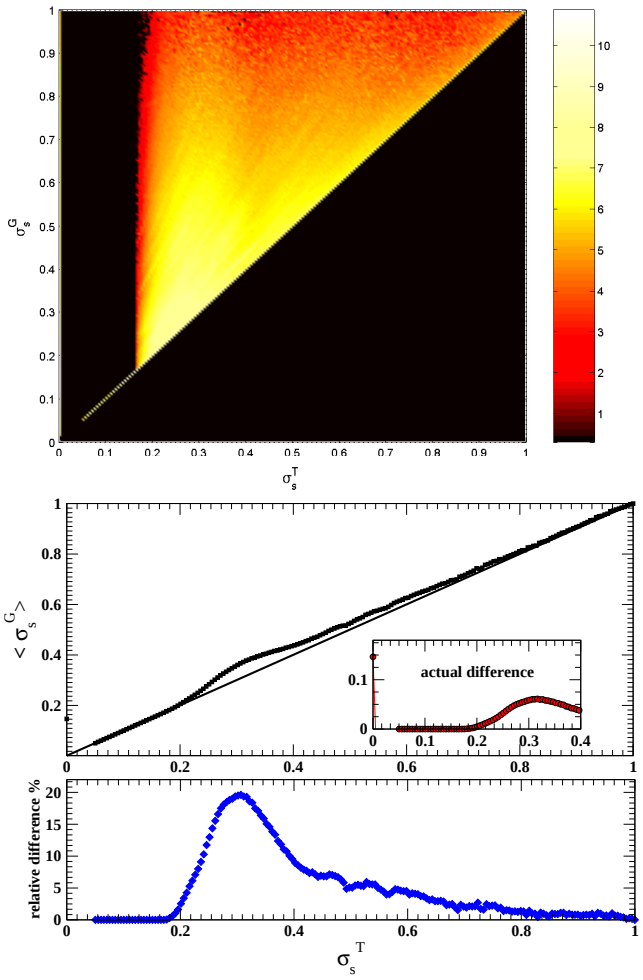

Figure 3: Top: 200×200 bin histogram showing the distributions of 1.26×10^{12} pairs of pages according to tree-based vs. graph-based semantic similarity. Colors encode numbers of pairs on a log scale. Bottom: Averaging of σ_s^G for each σ_s^T bin highlights the difference between the two similarity measurements.

question, they were presented with a target Web page and two candidate Web pages (see Figure 4). The subjects had to answer by selecting from the two candidate pages the one that was more related to the target Web page or by indicating that neither of the candidate pages was related to the target. A total of 6 target Web pages randomly selected from the ODP directory were used for the evaluation. For each target Web page we presented a series of 5 pairs of candidate Web pages. To investigate which of the two methods was a better predictor of human assessments of Web page similarity, the candidate pages were selected with controlled differences in their semantic similarity to the target page. Given a target Web page p^T, each pair of candidate pages p_1^C and p_2^C used in our study satisfied the following two conditions:

$$\text{Condition 1:} \quad \sigma_s^T(p_1^C, p^T) \geq \sigma_s^T(p_2^C, p^T)$$
$$\text{Condition 2:} \quad \sigma_s^G(p_1^C, p^T) < \sigma_s^G(p_2^C, p^T)$$

Figure 4: A snapshot of the experiment setup for our user study. The pages displayed are those of Table 1.

The use of the above conditions guarantees that for each question the two models disagreed on their prediction of which of the two candidate pages is more related to the target page. The pages in the 30 triplets were chosen at random among all the cases satisfying the above conditions. To ensure that the participants made their choice independently of the questions already answered, we randomized the order of the options. Table 1 shows an example of a triplet of pages used in our study, corresponding to the question in the snapshot of Figure 4. The users were presented with the target and candidate pages only — no information related to the topics of the pages was shown to the users.

The semantic similarity between the target page and each of the candidate pages in our example, according to the two measurements is as follows:

$$\sigma_s^T(p_1^C, p^T) = 0.24 \quad \sigma_s^T(p_2^C, p^T) = 0.50$$
$$\sigma_s^G(p_1^C, p^T) = 0.91 \quad \sigma_s^G(p_2^C, p^T) = 0.70$$

For this triplet of pages, the tree-based method predicts that p_2^C is more similar to the target than p_1^C ($\sigma_s^T(p_2^C, p^T) > \sigma_s^T(p_1^C, p^T)$). On the other hand, according to the prediction made by the graph-based method p_1^C should be preferred over p_2^C ($\sigma_s^G(p_1^C, p^T) > \sigma_s^G(p_2^C, p^T)$).

To test which of the two methods was a better predictor of subjects' judgments of Web page similarity we considered the selections made by each of the human-subjects and computed the percentage of correct predictions made by the two methods. Table 2 summarizes the statistical results. This comparison table shows that the graph-based semantic similarity measure results in statistically significant improvements over the tree-based one.[3]

[3]This made it unnecessary to recruit a larger subject pool.

Table 2: Mean, standard deviation, and standard error of the percentage of correct predictions by tree-based vs. graph-based semantic similarity, as determined from the assessments by the N subjects. The fact that the confidence intervals do not overlap is equivalent to using a t-test to determine that the difference in average accuracy is statistically significant at the 95% confidence level.

	N	MEAN	STDEV	SE	95% C.I.
σ_s^T	38	5.70%	4.71%	0.76%	**(4.2%, 7.2%)**
σ_s^G	38	84.65%	11.19%	1.82%	**(81.1%, 88.2%)**

4. APPLICATIONS

Having validated our semantic similarity measure σ_s^G, let us now begin to explore its applications to performance evaluation. Using σ_s^G as a surrogate for user assessments of semantic similarity, we can address the general question of how text and link analyses can be combined to derive measures of relevance that are in good agreement with semantic similarity. An analogous approach has been used in the past to evaluate similarity search, but relying on only the hierarchical ODP structure as a proxy for semantic similarity [7, 16].

Let us start by introducing two representative similarity measures σ_c and σ_ℓ based on textual content and hyperlinks, respectively. Each is based on the TF-IDF vector representation and "cosine similarity" function traditionally used in information retrieval [20]. For *content similarity* we use:

$$\sigma_c(p_1, p_2) = \frac{\vec{p}_1^{\,c} \cdot \vec{p}_2^{\,c}}{\|\vec{p}_1^{\,c}\| \cdot \|\vec{p}_2^{\,c}\|}$$

where (p_1, p_2) is a pair of Web pages and $\vec{p}_i^{\,c}$ is the TF-IDF vector representation of p_i, based on the terms in the page. Noise words are eliminated [4] and other words are conflated using the standard Porter stemmer [18].

For *link similarity* measure we define:

$$\sigma_\ell(p_1, p_2) = \frac{\vec{p}_1^{\,\ell} \cdot \vec{p}_2^{\,\ell}}{\|\vec{p}_1^{\,\ell}\| \cdot \|\vec{p}_2^{\,\ell}\|}$$

where $\vec{p}_i^{\,\ell}$ is the *link frequency–inverse document frequency* (LF-IDF) vector representation of page p_i. LF-IDF is analogous to TF-IDF, except that hyperlinks (URLs) are used in place of words (terms). A page link vector is composed of its outlinks, inlinks, and the pages's own URL. Link similarity is a measure of the local undirected clustering coefficient between two pages. A high value of σ_ℓ indicates that the two pages belong to a clique of pages. Related measures are often used in link analysis to identify a community around a topic. This measure generalizes co-citation [21] and bibliographic coupling [9], but also considers directed paths of length $L \leq 2$ links between pages. Such directed paths are important because they could be navigated by a user or crawler. Outlinks were obtained from the pages themselves, while inlinks were obtained from a search engine.[4]

One could of course explore alternative content and link similarity measures, however our preliminary experiments indicate that other commonly used measures such as TF-

[4]We used the Google Web API (`www.google.com/apis/`) with special permission.

Table 1: Example of a triplet used in the evaluation

Page	URL	Topic
p^T	`http://www.muppetsonline.com/`	Arts Performing_Arts Puppetry Muppets
p_1^C	`http://www.theentertainmentbusiness.com/sesame.htm`	Arts Television Programs Children's Sesame_Street Characters
p_2^C	`http://www.yale.edu/yags/`	Arts Performing_Arts Circus Juggling Clubs_and_Organizations College_Juggling_Clubs

based cosine similarity and the Jaccard coefficient do not qualitatively alter the observations that follow.

Once text and links were extracted from the 1.12×10^6 Web pages of the ODP ontology, $\sigma_c \in [0,1]$ and $\sigma_\ell \in [0,1]$ were computed for each of 1.26×10^{12} pairs of pages. Semantic similarities σ_s^T and σ_s^G were measured as well. Two $200 \times 200 \times 200$ histograms with coordinates $(\sigma_c, \sigma_\ell, \sigma_s^T)$ and $(\sigma_c, \sigma_\ell, \sigma_s^G)$ were generated to analyze the relationships between the various similarity measures. We focus on the latter, graph-based semantic similarity in the following analysis. The computation of these histograms (and the one for (σ_s^T, σ_s^G), cf. Section 3.1) required approximately 4,000 additional CPU hours on the AVIDD facility.

4.1 Combining Content and Link Similarity

The massive data thus collected allows us to study how well different automatic similarity measures based on observable features (content and links) approximate semantic similarity. We considered a number of simple functions $f(\sigma_c, \sigma_\ell)$ including:

- various linear combinations $f = \lambda\sigma_c + (1-\lambda)\sigma_\ell$ for $0 \le \lambda \le 1$, of which we report the cases $\lambda = 0$ ($f = \sigma_\ell$), $\lambda = 0.2$, $\lambda = 0.8$, and $\lambda = 1$ ($f = \sigma_c$);
- the product $f = \sigma_c\sigma_\ell$;
- the step-linear function $f = \sigma_c H(\sigma_\ell)$, where $H(\sigma_\ell) = 1$ for $\sigma_\ell > 0$ and 0 otherwise;

and other functions omitted for space considerations. Figure 5 plots the Pearson and Spearman correlations between σ_s^G and these functions, versus a threshold on σ_c.

The Pearson correlation coefficient r_P tells us the degree to which the values of each function $f(\sigma_c, \sigma_\ell)$ agree with σ_s^G. We can see that the correlations are rather weak, $0 < r_P < 0.2$, for all f in the plot when we consider all page pairs. If we restrict the analysis to pairs that have content similarity σ_c above a minimum threshold, the correlations can become much stronger. It is meaningful to use a σ_c threshold because in applications such as search engines, the pages to be ranked are those that are retrieved from an index based on a match, typically between pages and a user query or some other model page. It is interesting to observe that the functions that rely heavily on content similarity ($f = \lambda\sigma_c + (1-\lambda)\sigma_\ell$ for high λ) perform particularly poorly at predicting semantic similarity. They are at best weakly correlated with σ_s^G unless one applies a very high σ_c threshold. This is rather surprising because prior to the introduction of link based importance measures such as PageRank [1] content was the sole source of evidence for ranking pages, and content similarity is still widely seen as a central component of any ranking algorithm.

The Pearson correlation assumes normally distributed values. Since the similarity functions defined above have mostly exponential distributions, it is worth to validate the above results using the Spearman rank order correlation coefficient r_S, which is high if two functions agree on the rankings they produce irrespective of the actual values. This is reasonable in our setting because from a search engine user perspective, what matters is the order of the hit pages and not the values used by the ranking function. The Spearman correlation data in Figure 5 confirms the above observations, with even more striking evidence of the noisy nature of content similarity. One can see a clear separation between the poor rankings produced by functions that depend linearly on σ_c and the relatively good rankings produced by functions that either do not consider σ_c or that scale σ_c by σ_ℓ.

The above analysis highlights an extremely low discrimination power of lexical similarity. This might suggest a filtering role for lexical similarity, in which all pages below a small threshold would not be considered while above the threshold only link-based measures would be used for the sake of ranking. While such a bold strategy must be scrutinized carefully, it could lead to a significant simplification of ranking algorithms.

4.2 Evaluating Ranking Functions

Let us finally illustrate how the proposed semantic similarity function can be used to automatically evaluate alternative ranking functions. This makes it possible to mine through a large number of alternative functions automatically and cheaply, reserving user studies for the most promising candidates. We want to compare the quality of a ranking

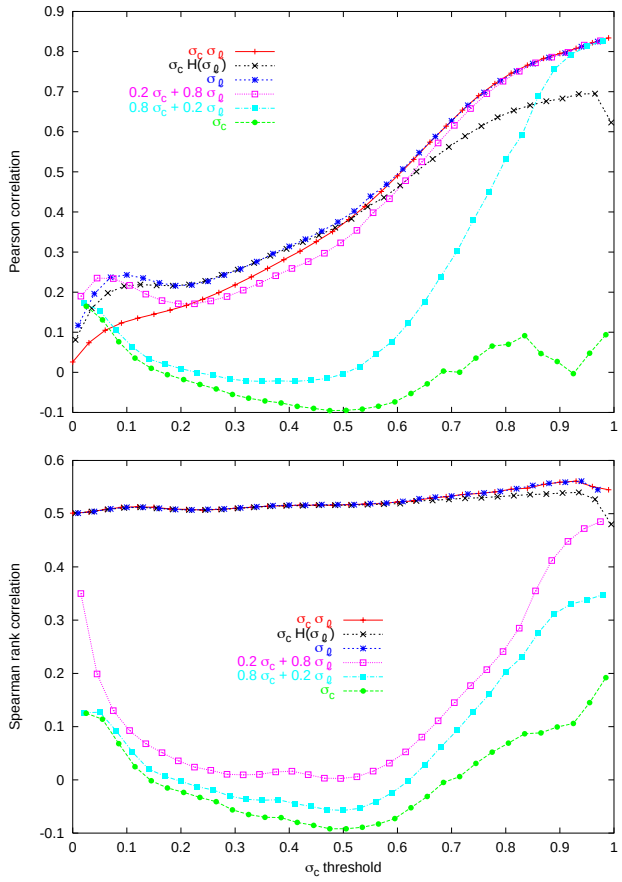

Figure 5: Pearson (top) and Spearman (bottom) correlations between graph-based semantic similarity σ_s^G and different functional combinations of content and link similarity, applying increasing thresholds on content similarity.

function to the baseline ranking obtained by the use of semantic similarity. The *sliding ratio* score [17, 11] compares two rankings when graded quality assessments are available.[5] This measure is defined as the ratio between the cumulative quality scores of the top-ranked pages according to two ranking functions. We can generalize the sliding ratio in the following ways:

- use a page as a target rather than an arbitrary query, as is done in "query by example" systems;
- use σ_s^G as a reference ranking function;
- sum over all pages in an ontology such as the ODP, each used in turn as a target, thus covering the entire topical space and eliminating the dependence on a single target.

[5]In the common case when just binary relevance assessments are available, one resorts to precision and recall; the sliding ratio score is a more sophisticated measure enabled by more refined semantic similarity data.

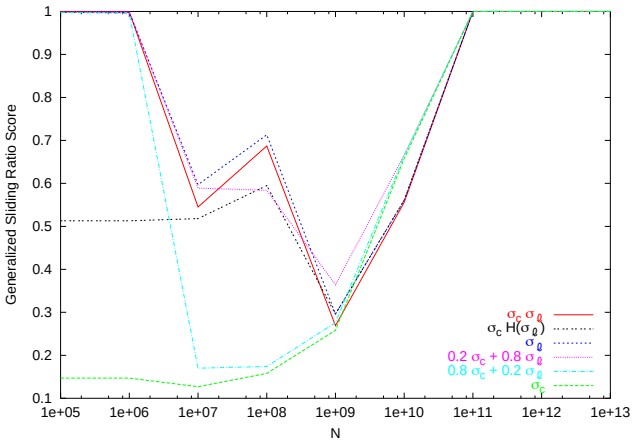

Figure 6: Generalized sliding ratio score plots for different functional combinations of content and link similarity. We omit the region $N < 10^5$ where GSR is constant for all f up to the resolution of our histogram bins.

Let us thus define a *generalized sliding ratio* score as follows:

$$GSR(f,N) = \frac{\sum_{(i,j):rank_f(i,j)=1}^{N} \sigma_s^G(i,j)}{\sum_{(i,j):rank_{\sigma_s^G}(i,j)=1}^{N} \sigma_s^G(i,j)}$$

where (i,j) is a pair of pages, f is a ranking function to be tested, and N is the number of top-ranked pairs considered. Note that for any f, $GSR(f,N) \to 1$ as N tends to the total number of pairs. The ideal ranking function is one such that $GSR(f,N) \approx 1$ for low N as well. In simplistic terms, $GSR(f,N)$ tells us how well a function f ranks the top N pairs of pages.

The generalized sliding ratio score can be readily measured on our ODP data for any $f(\sigma_c, \sigma_\ell)$. Only pairs with $\sigma_c > 0$ are considered, since typically in a search engine only pages matching the query are retrieved. In Figure 6 we plot $GSR(f,N)$ versus N for the simple combination functions $f(\sigma_c, \sigma_\ell)$ introduced in Section 4.1. Consistently with the correlation results, the functions that depend heavily on content similarity rank poorly. Again this is only an illustration of how the σ_s^G measure can be applied to the evaluation of arbitrary ranking functions.

5. DISCUSSION

In this paper we introduced a novel measure of semantic similarity for Web pages that generalizes the well-founded information-theoretic tree-based semantic similarity measure to the general case in which pages are classified in the nodes of an arbitrary graph ontology with both hierarchical and non-hierarchical components. This measure can be readily applied to mine semantic data from topical ontologies and Web directories such as Yahoo!, the ODP and their derivatives.

Similarity is commonly viewed as an example of relation satisfying the following three conditions:

- Maximality: $\sigma(a,b) \leq \sigma(a,a) = 1$.
- Symmetry: $\sigma(a,b) = \sigma(b,a)$.
- Triangular Inequality: $\sigma(a,b) \cdot \sigma(b,c) \leq \sigma(a,c)$.

These conditions are adaptations of the *minimality*, *symmetry* and *triangle inequality* axioms of metric distance functions. The definition of σ_s^G proposed in this paper satisfies maximality and symmetry but not the triangular inequality condition. With sufficient computational resources, a new measure of semantic similarity satisfying the triangular inequality principle can be computed by applying an adaptation of Dijkstra's shortest path algorithm [2] to σ_s^G:

$$\begin{aligned} \sigma^{(0)}(i,j) &= \sigma_s^G(i,j) \\ \sigma^{(r+1)}(i,j) &= \max\left(\sigma^{(r)}(i,j), \max_k\left(\sigma^{(0)}(i,k) \cdot \sigma^{(r)}(k,j)\right)\right) \\ \sigma(i,j) &= \lim_{r\to\infty} \sigma^{(r)}(i,j) \end{aligned}$$

While in many cases the lower limit imposed by the triangular inequality appears to be intuitive, many authors have argued against it. Tversky [22] illustrates this position with an example about the similarity between countries: *"Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of their political affinity); but Jamaica and Russia are not similar at all."* This example fits the case of Web pages and their topics, suggesting that the triangular inequality should not be accepted as a cornerstone of similarity models.

Computing the graph-based semantic similarity measure is a computationally expensive task, both in terms of space and time. While matrices $\mathbf{T}$, $\mathbf{G}$, $\mathbf{T}^+$ and $\mathbf{W}$ are sparse and easy to store, codifying the graph-based semantic similarity measure σ_s^G for the ODP topics required the use of 5,712 dense matrices, each one of size $571,148 \times 100$. The time complexity for computing the semantic similarity for n topics is $O(n^3)$ in the worst case; the actual complexity depends on the density of the $\mathbf{W}$ matrix. Some of the techniques adopted to deal with the time complexity of the problem include indexing the sparse structure of the matrices for fast access and using a *software vector register* to compute the MaxProduct fuzzy composition function efficiently. Our approach may not scale easily to ontologies much larger than the ODP graph as it is today. However, approximations of σ_s^G may be computed in reasonable time if appropriate heuristics are applied (e.g., via use of thresholds).

We have shown that the proposed semantic similarity measure predicts human judgments of relatedness with significantly greater accuracy than the tree-based measure. Finally we have undertaken a massive data mining effort on ODP data in order to begin to explore how text and link analyses can be combined to derive measures of relevance in agreement with semantic similarity.

The methodology described here to evaluate ranking algorithms based on semantic similarity can be applied to arbitrary combinations of ranking functions stemming from text analysis (e.g. LSA, query expansion, tag weighting, etc.), link analysis (e.g. authority, PageRank, SiteRank, etc.), and any other features available to a search engine (e.g. freshness, click-through rate, etc.). Yet the applications of the proposed semantic similarity measure are broader than just Web search. Classification, clustering and resource discovery also rely on semantic mining of features that can be extracted automatically.

The main, surprising result of our initial analysis with the graph-based semantic similarity is that the classic text-based TF-IDF cosine similarity is an extremely noisy feature, unfit for ranking Web pages. While it seems helpful to filter out pages with very low lexical similarity ($\sigma_\ell < 0.05$), text-based measures do not seem to help in ranking the remaining pages. On the contrary they are very poorly correlated with semantic similarity, possibly reflecting the extent to which ambiguous terms mislead the search process. While this result helps to explain why early search engines did so poorly and validates the use of link-based measures such as PageRank, the seemingly unredeemed quality of content similarity is unexpected. The implication must be a revisitation of the role of content similarity in ranking Web results.

We are currently exploring alternative ways to approximate semantic similarity by *integrating* (rather than combining) content and link similarity. The correlation plots in Figure 5 suggest that content may play a positive role in filtering hits, if not in ranking them.

In future work the semantic similarity measure must be further validated through user studies. The study presented here focuses on cases where σ_s^G and σ_s^T disagree, and thus it tells us that σ_s^G is more accurate than σ_s^T but is too biased to satisfactorily answer the broader question of how well σ_s^G predicts assessments of semantic similarity by human subjects in general. It is possible that alternative weighting schemes for the different types of links in the ODP ontology may lead to measures with improved accuracy.

The evaluations outlined here have focused on purely local text and link analysis. For example, we have not looked at the role of more global link and text analysis techniques such as PageRank and latent semantic analysis (LSA) in improving the quality of ranking by favoring authoritative pages or improving content similarity. These are also directions for future work.

Due to the growing number of emerging Web search techniques and the scale of the Web, automatic evaluation mechanisms are crucial. In the light of the availability of rich semantic information sources, like the ODP ontology, we have proposed a reliable method for the algorithmic detection of semantic similarity between Web pages. The proposed approach will provide insight for better understanding the limitations of existing search techniques and inspire the development of new and more powerful Web search tools.

6. ACKNOWLEDGMENTS

We are grateful to E. Milios, S. Chakrabarti, J. Kleinberg, L. Adamic, P. Srinivasan, and N. Street for many helpful comments; to R. Bramley for sharing his expertise in scientific computing; to the ODP for making their data publicly available; to Google for their permission to use the Web API extensively; and to IU's Research and Technical Services (especially S. Simms) for technical support. Nihar Sanghvi carried out some of the early data collection.

This work was funded in part by NSF Career Grant IIS-0348940 to FM. The AVIDD Linux Clusters used in our data analysis have been funded in part by NSF Grant CDA-9601632. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

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

[2] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] C. Fox. Lexical analysis and stop lists. In *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.

[5] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.*, 21(1):64–93, 2003.

[6] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[7] 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*, New York, NY, 2002. ACM Press.

[8] A. Kandel. *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley, 1986.

[9] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[10] J. M. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *IEEE Symposium on Foundations of Computer Science*, pages 14–23, 1999.

[11] R. Korfhage. *Information Storage and Retrieval*. John Wiley and Sons, New York, NY, 1997.

[12] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann Publishers Inc., 1998.

[13] W. Lu, J. Janssen, E. Milios, and N. Japkowicz. Node similarity in networked information spaces. In *Proceedings of the Conference of the IBM Centre for Advanced Studies on Collaborative Research (CASCON'01)*. IBM Press, 2001.

[14] F. Menczer. Combining link and content analysis to estimate semantic similarity. In *Alt. Track Papers and Posters Proc. 13th International World Wide Web Conference*, pages 452–453, 2004.

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

[16] F. Menczer. Finding semantic needles in haystacks of web text and links. *IEEE Internet Computing*, 2005. Forthcoming.

[17] S. Polalck. Measures for the comparison of information retrieval systems. *American Documentation*, 19(4):387–397, 1968.

[18] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[19] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[20] G. Salton and M. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.

[21] H. Small. Co-Citation in the scientific literature: A new measure of the relationship between documents. *Journal of the American Society for Information Science*, 42:676–684, 1973.

[22] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.