Algorithmic Computation and Approximation of Semantic Similarity

Ana G. Maguitman[†‡], Filippo Menczer[†‡], Fulya Erdinc[†],
Heather Roinestad[†], Alessandro Vespignani[‡]
† Department of Computer Science
‡ School of Informatics
Indiana University
Bloomington, IN 47408
{*anmaguit,fil,ferdinc,hroinest,alexv*}*@indiana.edu*

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.

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 [32]. More recently, semantic similarity became fundamental in knowledge representation where special kinds of networks or ontologies are used to describe objects and their relationships [8].

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 [7], 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 [20], 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 briefly review some of the existing information-theoretic proposals to estimate semantic similarity, in particular, we focus on a tree-based notion of semantic similarity proposed by Lin [17]. In section 3 we propose a semantic similarity measure that generalizes the tree-based similarity to the case of a graph. 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. We close the section by addressing the question of how

[1] http://dmoz.org

to generalize our definition of graph-similarity and by proposing a family of measures that can be used to compute semantic similarity on other kinds of ontologies.

Section 4 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 5 we begin to explore the question of applications, namely how text and link analyses can be used to derive measures of relevance that are in good agreement with semantic similarity. We consider various extensions and combinations of basic text and link similarity and discuss how these correlate with semantic similarity. We find that surprisingly, classic text-based content similarity is a very noisy feature, whose value is at best weakly correlated with semantic similarity. 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 Information-Theoretic Measures of Semantic Similarity

Many measures have been developed to estimate semantic similarity in a network representation. Early proposals have used path distances between the nodes in the network (e.g. [28]). These frameworks are based on the premise that the stronger the semantic relationship of two objects, the closer they will be in the network representation. However, as it has 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 [29, 10, 11].

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 [29, 17]. 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. Once this common classification is identified, the meaning shared by two objects can be measured by the amount of information needed to state the commonality of the two objects.

In information theory [3], the information content of a class or topic t is measured by the negative log likelihood, $-\log \Pr[t]$, where $\Pr[t]$ represents the prior probability that any object is classified under topic t. In practice $\Pr[t]$ can be computed for every topic t in an "is-a" taxonomy by counting the fraction of objects stored in node t and its descendants out of all the objects in the taxonomy.

Based on this quantitative characterization of object commonality Resnik [29] introduced 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 two topics t_1 and t_2 in an "is-a" taxonomy:

$$\sigma(t_1, t_2) = \max_{t_s \in S(t_1,t_2)} (-\log \Pr[t_s])$$

where $S(t_1, t_2)$ is the set of topics that subsume both t_1 and t_2. Resnik's measure has been applied with some degree of success to diverse scenarios, including concept relatedness in WordNet [25] and protein similarity based on their Gene Ontology (GO) annotations [19]. A limitation of Resnik's measure is that the similarities between all the children of a topic t are identical, independently of their information content.

Lin [17] has investigated an information theoretic definition of semantic similarity closely related to Resnik's measure. In Lin's proposal, not only the common meaning of the two topics but also their individual meaning is taken into account. Indeed, according to Lin's proposal, the semantic similarity between two topics t_1 and t_2 in a taxonomy is defined as a function of the meaning shared by the topics (represented by the most specific topic that subsumes t_1 and t_2) and the meaning of each of the individual topics:

$$\sigma(t_1, t_2) = \max_{t_s \in S(t_1,t_2)} \frac{2 \cdot \log \Pr[t_s]}{\log \Pr[t_1] + \log \Pr[t_2]}$$

Assuming the taxonomy is a tree, the semantic similarity between two topics t_1 and t_2 is then measured as the ratio between the meaning of their lowest common ancestor and their individual meanings. This can be expressed 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. 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 tree 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.

This measure of semantic similarity has several desirable properties and a solid theoretical justification. It 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^T corresponds to the familiar tree distance measure [15].

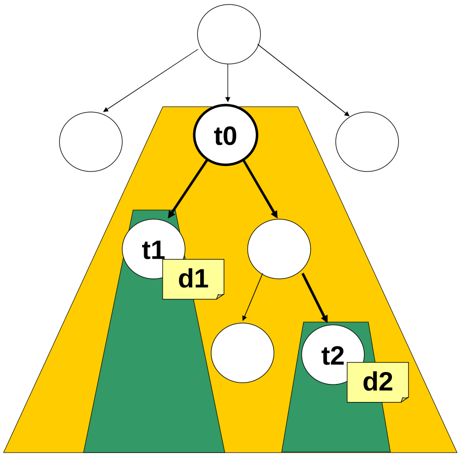


Figure 1: Illustration of tree-based semantic similarity in a taxonomy.

In prior work [21, 22, 23] 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 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. For instance, according to the tree-based semantic similarity the pages stored under the topic "Business/E-Commerce" are unrelated to the ones stored under the topic "Computers/Software/Business/E-Commerce." This yielded an unreliable picture when all topics were considered.

3 Graph-Based Semantic 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, $|t_3| = 4$).

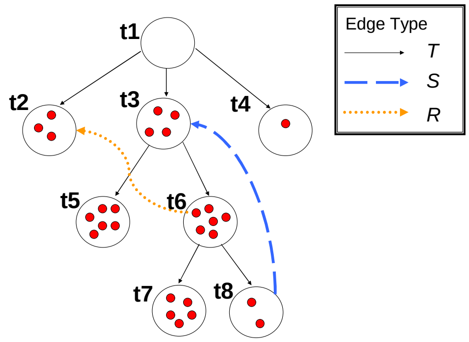


Figure 2: Illustration of a simple ontology.

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 fuzzy union operation $\cup$ on matrices representing relations, defined as $[A \cup B]_{ij} = \max(A_{ij}, B_{ij})$, and let $\mathbf{G} = \mathbf{T} \cup \mathbf{S} \cup \mathbf{R}$. Matrix $\mathbf{G}$ is the adjacency matrix of graph G augmented with 1s on the diagonal.

We will use the MaxProduct fuzzy composition function $\odot$ [12] 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.

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:

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

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

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

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

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

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

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

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 $\mathit{subtree}(t_i)$, the topic subtree rooted at t_i, and all topics $t \in \mathit{subtree}(t_i)$ 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.

3.1 Towards a More General Definition of Graph-Similarity

A natural question that arises is how to generalize the proposed measure of graph similarity in such a way that it can be applied to other ontologies. Different ontologies have different kinds of edges with diverse semantics. As we have seen earlier, the ODP ontology has three types of edges "is-a", "symbolic" and "related." Our choice of a particular weighting scheme ($\alpha = 1$, $\beta = 1$, and $\gamma = 0.5$) as well as the selection of a specific composition operator (MaxProduct) and transitive membership relations reflect our interpretation of the semantics for the different types of edges.

Applying graph-based semantic similarity to other ontologies requires appropriate mechanisms to model different kinds of ontology components and their interactions. For example, the Gene Ontology[3] has two kinds of hierarchical edges ("is-a" and "part-of"). On the other hand, the WordNet ontology[4] has a much richer typology of relations. This includes semantic relations between synsets (synonym sets) such as hypernym, hyponym, meronym and holonym as well as lexical relations between senses of words (members of synsets) such as antonym, "also see", derived forms and participle.

Two main aspects of the proposed graph-based semantic similarity measure can be generalized. One is the definition of σ_s^G and the other is the way the information content of a class or topic, characterized by the notion of topic cone, is estimated. The definition of σ_s^G as discussed in this work is sufficiently general to suit a variety of graph ontologies. What seems particularly sensitive to the specific ontologies is how a topic cone (matrix $\mathbf{W}$) is calculated as this depends directly on the semantics attached to the edges.

Let matrix $\mathbf{E_{ij}}$ codify a class of edges, where $\mathbf{E_{ij}}$ is augmented with 1s on the diagonal. A family of characterizations of the notion of topic cone can be expressed as instances of the following general formula:

$$\mathbf{W} = \bigcup_i \bigodot_j \mathbf{E_{ij}}^{(k_{ij})}$$

where $\bigcup$ is the fuzzy Union operator, $\bigodot$ is the MaxProduct fuzzy composition operator (or some other suitable fuzzy composition operator such as MaxMin) and $k_{ij} \in \mathbb{N} \cup \{+\}$.

The above formula is expressive enough to model the notion of topic cone in different classes of ontologies. For example, in the case of a taxonomy it is sufficient to set $\mathbf{E_{11}} = \mathbf{T}$ and $k_{11} = +$. For the case of the ODP graph it is easy to see that our formulation of matrix $\mathbf{W}$ can also be expressed as a special case of the above general formula as follows:

$$\mathbf{W} = (\mathbf{T}^+ \odot \mathbf{S} \odot \mathbf{T}^+) \cup (\mathbf{T}^+ \odot \mathbf{R} \odot \mathbf{T}^+).$$

Figure 3(a) illustrates how a path is computed according to this definition of topic cone.

We are exploring other promising formulations of matrix $\mathbf{W}$, including the following ones:

$$\mathbf{W} = (\mathbf{T}^+ \odot \mathbf{S} \odot \mathbf{T}^+) \cup (\mathbf{T}^+ \odot \mathbf{R}),$$

[3] http://www.geneontology.org/

[4] http://wordnet.princeton.edu/

and

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

According to the first formulation, illustrated in Figure 3(b), a path in a topic cone can contain any number of hierarchical edges (T) but at most one cross-link (S or R). In addition, it must satisfy that if a cross-link of type "related" (R) occurs in a path, it must be the last in the path.

The second formulation of topic cone, illustrated in Figure 3(c), also allows any number of hierarchical edges in a path. Cross-links of type "symbolic" and "related" can occur at most once each in a path, with links of type "related" only occurring at the end of a path.

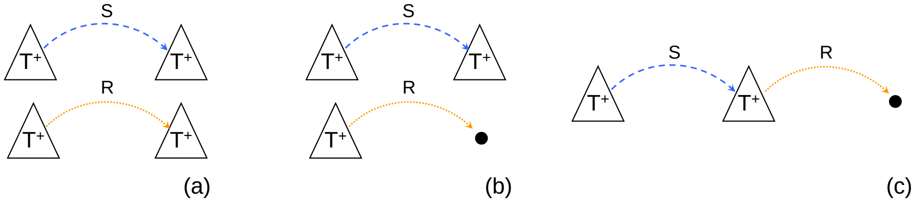


Figure 3: Illustration of three ways of identifying paths in a topic cone.

These and other characterizations of topic cone will be studied in detail in future work. The results presented in the rest of this article are based on our original characterization of topic cone.

4 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.

4.1 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 4 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 4 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 $\sigma_s^T = 0$.

4.2 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 questions about similarity between Web pages. For each question, they were presented with a target Web page and two candidate Web pages (see Figure 5). 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. Given the constraint on the duration of an experiment, there is a trade-off between diversity and number of examples. One could allocate each question to a different triplet, or have a smaller number of target pages with several different pairs of candidate pages for each target. Preliminary tests indicated that the former approach imposed a higher cognitive load on the subjects, requiring more time per question and decreasing the total number of questions they could answer in the allotted time. To increase the number of questions and the precision of the results, we settled on the latter approach. 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, for a total of 30 questions. 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:

$$\begin{array}{ll} \text{Condition 1:} & \sigma_s^T(p_1^C, p^T) \geq \sigma_s^T(p_2^C, p^T) \\ \text{Condition 2:} & \sigma_s^G(p_1^C, p^T) < \sigma_s^G(p_2^C, p^T) \end{array}$$

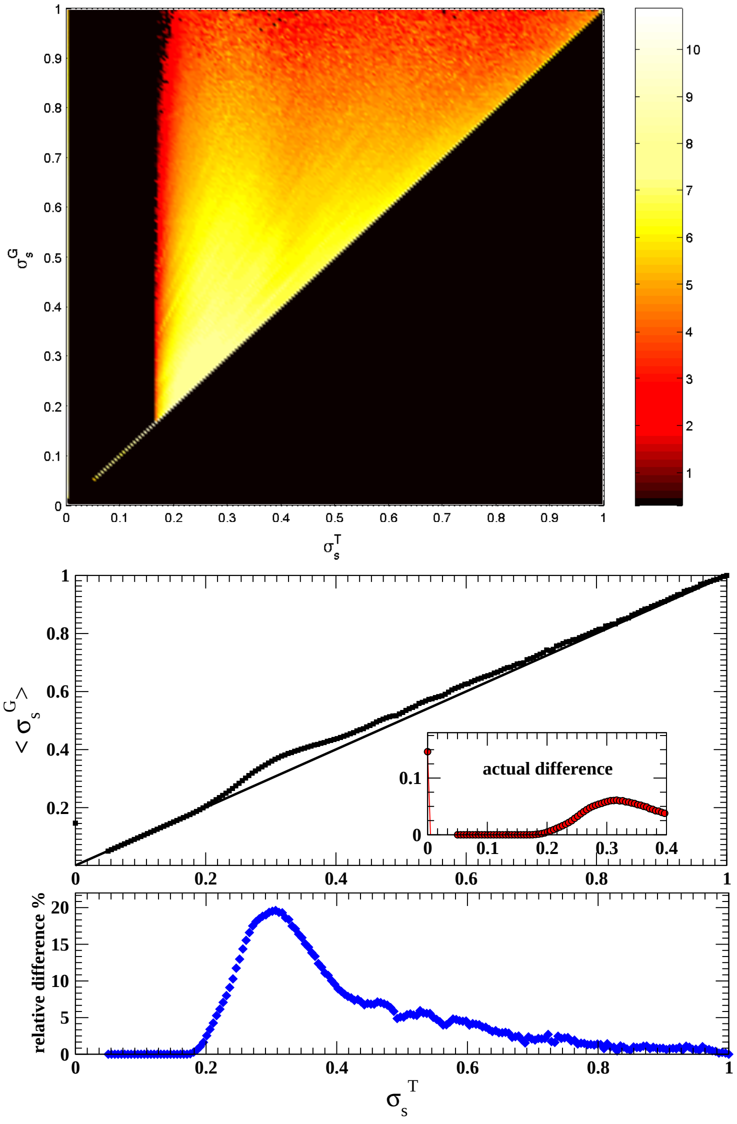


Figure 4: Top: 200 × 200 bin histogram showing the distributions of 1.26×10^{12} pairs of pages according to tree-based vs. graph-based semantic similarity. Colors encode numbers of pairs on a log scale. Middle: Average graph similarity $\langle \sigma_s^G \rangle$ for each σ_s^T bin, contrasted with the baseline $\langle \sigma_s^G \rangle = \sigma_s^T$ (thin line). The inset highlights the difference between the two similarity measurements. Bottom: Relative difference $(\langle \sigma_s^G \rangle - \sigma_s^T)/\sigma_s^T$ versus σ_s^T.

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

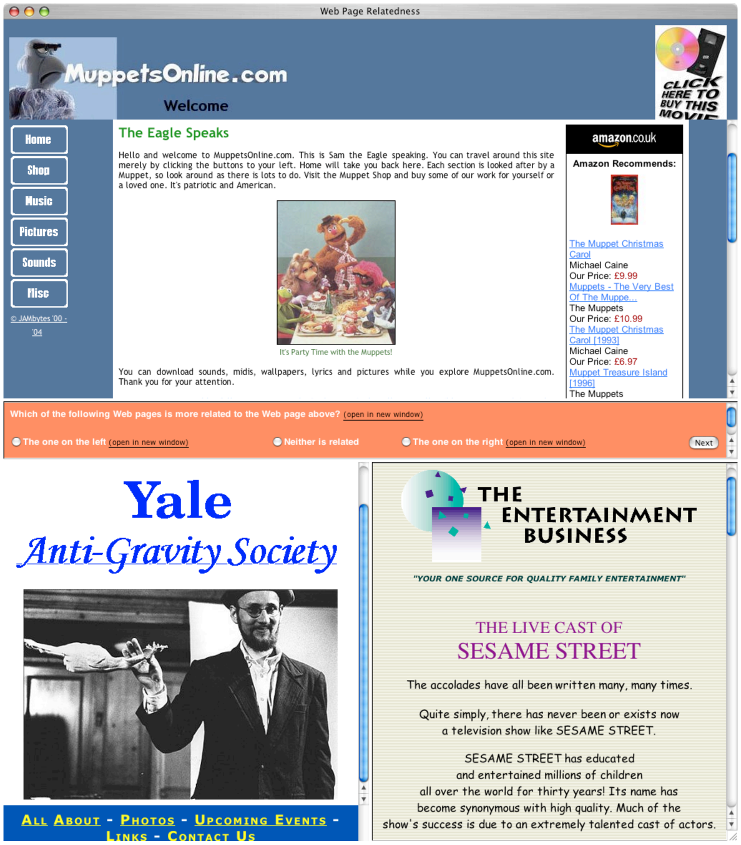
The use of the above conditions guarantees that for each question the two models disagreed on their prediction of which of the two candidate pages is more related to the target page. The pages in the 30 triplets were chosen at random among all the cases satisfying the above conditions. To ensure that 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 5. 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:

$$\begin{array}{ll} \sigma_s^T(p_1^C, p^T) = 0.24 & \sigma_s^T(p_2^C, p^T) = 0.50 \\ \sigma_s^G(p_1^C, p^T) = 0.91 & \sigma_s^G(p_2^C, p^T) = 0.70 \end{array}$$

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 subjects and computed the percentage of correct predictions made by the two methods. Table 2 summarizes the statistical results. This comparison shows that the graph-based semantic similarity measure results in statistically significant improvements over the tree-based one.



The Eagle Speaks

Hello and welcome to MuppetsOnline.com. This is Sam the Eagle speaking. You can travel around this site merely by clicking the buttons to your left. Home will take you back here. Each section is looked after by a Muppet, so look around as there is lots to do. Visit the Muppet Shop and buy some of our work for yourself or a loved one. It's patriotic and American.

You can download sounds, midis, wallpapers, lyrics and pictures while you explore MuppetsOnline.com. Thank you for your attention.

The Muppet Christmas Carol
Michael Caine
Our Price: £9.99
Muppets - The Very Best Of The Muppe...
The Muppets
Our Price: £10.99
The Muppet Christmas Carol [1993]
Michael Caine
Our Price: £6.97
Muppet Treasure Island [1996]
The Muppets

Which of the following Web pages is more related to the Web page above? (open in new window)

○ The one on the left (open in new window) ○ Neither is related ○ The one on the right (open in new window)

"YOUR ONE SOURCE FOR QUALITY FAMILY ENTERTAINMENT"

THE LIVE CAST OF
SESAME STREET

The accolades have all been written many, many times.

Quite simply, there has never been or exists now
a television show like SESAME STREET.

SESAME STREET has educated
and entertained millions of children
all over the world for thirty years! Its name has
become synonymous with high quality. Much of the
show's success is due to an extremely talented cast of actors.

Figure 5: A snapshot of the experiment setup for our user study. The pages displayed are those of Table 1.

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%)**

5 Case Studies

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 [9, 23].

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 [30]. 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 [6] and other words are conflated using the standard Porter stemmer [27].

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 [31] and bibliographic coupling [14], 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.[5]

One could of course explore alternative link and content representations and similarity measures, such as those based on conceptual graphs [24]. However our preliminary experiments indicate that other commonly used measures such as TF-based cosine similarity and the Jaccard coefficient do not qualitatively alter the observations that follow.

5.1 Combining Content and Link Similarity

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. A $200 \times 200 \times 200$ histogram with coordinates $(\sigma_c, \sigma_\ell, \sigma_s^G)$ was generated to analyze the relationships between the various similarity measures.

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

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 $0 \leq \lambda \leq 1$, 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 6 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 6 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

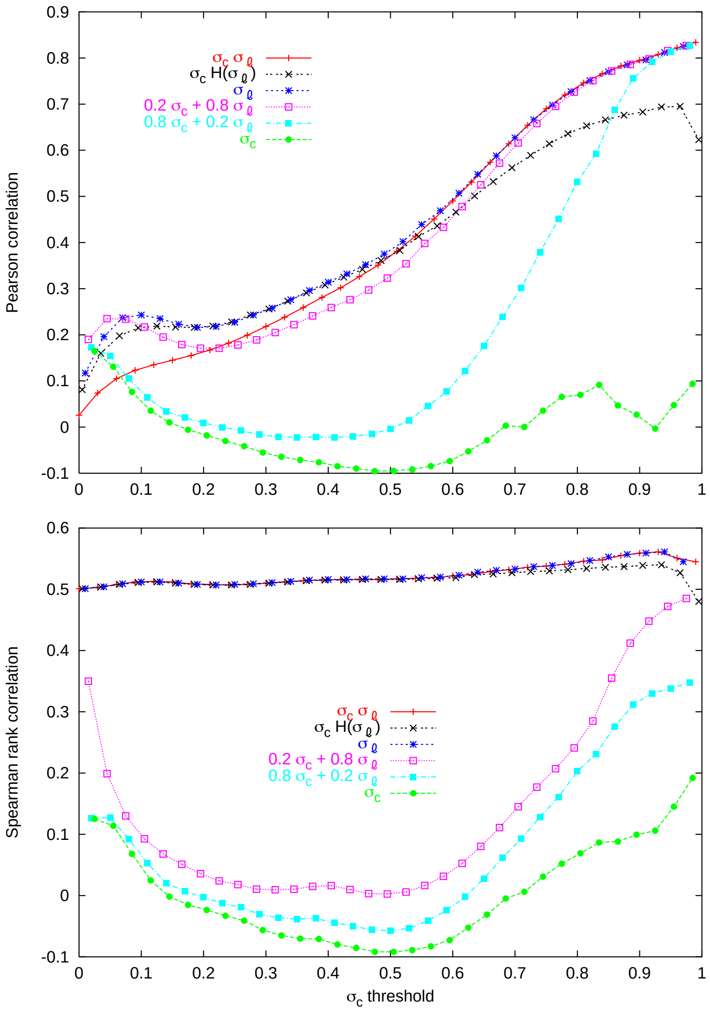


Figure 6: Pearson (top) and Spearman (bottom) correlations between graph-based semantic similarity σ_s^G and different functional combinations of content and link similarity, applying increasing thresholds on content similarity.

such a bold strategy must be scrutinized carefully, it could lead to a significant simplification of ranking algorithms.

5.2 Exploiting Term Co-occurrence to Improve Content Similarity

The observed poor performance of the traditional measure of content similarity drove us to explore extended forms of document similarity that exploit latent semantic relationships coming from term co-occurrence. When computing cosine similarity based on the vector space model, terms are represented as pairwise orthogonal vectors and document vectors are represented as linear combinations of these term vectors. Representing terms as orthogonal vectors presupposes semantic independence among them, which is clearly an unrealistic assumption.

The negative effects of this simplifying assumption have been addressed by previous studies and many extensions of the basic vector space model have been proposed based on the idea that terms that tend to co-occur are semantically related. For example, techniques such as Latent Semantic Indexing (LSI) apply singular value decomposition (SVD) to reduce the dimensions of the term-document space, harvesting the latent relations existing between documents and between terms [5]. The LSI method is computationally expensive and therefore many methods have been proposed to approximate LSI with reduced cost. An example of such methods is based on mapping documents to a kernel space where documents that do not share any term can still be close to each other [4]. A similar idea has also been investigated in [13] and more recently in [18]. In these proposals term-similarity is computed based on document similarity, and vice versa. This gives rise to a series of recursive equations that converges to a "more semantic" form of content similarity than the traditional one.

In order to investigate the effect of term co-occurrence on document similarity we implemented an extended form of content similarity. As in previous proposals, the underlying assumption for this new measure of similarity is that document similarity affects term similarity and vice versa, but instead of repeatedly computing one form of similarity in terms of the other, we only looked at a single step in this recursive process.

As a starting point, we computed term co-occurrence as follows:

$$\kappa(t_1, t_2) = \frac{\vec{t_1} \cdot \vec{t_2}}{\|\vec{t_1}\| \cdot \|\vec{t_2}\|}$$

where (t_1, t_2) is a pair of terms and $\vec{t_i}$ is the vector representation of t_i, based on the documents in which t_i occurs.

Finally, given a pair of Web pages (p_1, p_2), our extended form of document similarity was computed as follows:

$$\sigma_\kappa(p_1, p_2) = \frac{(\vec{p_1}^c \times \mathbf{K}) \cdot (\vec{p_2}^c \times \mathbf{K})}{\|\vec{p_1}^c \times \mathbf{K}\| \cdot \|\vec{p_2}^c \times \mathbf{K}\|}$$

where $\vec{p_i}^c \times \mathbf{K}$ is the TF-IDF vector representation of p_i projected into a non-orthogonal term space defined by the term-term matrix $\mathbf{K}$, where $[\mathbf{K}]_{ij} = \kappa(t_i, t_j)$.

In order to investigate if σ_κ is a good approximation of σ_s^G we used a subset of the data discussed in section 5.1. This reduced set consists of 150,000 URLs from 47,174 topics. The sample was obtained by extracting 10,000 URLs from each of the 15 top-level branches of the ODP ontology. Terms occurring in a single document were eliminated. After this term cleaning process, those documents containing no terms were also removed, resulting on a final corpus of

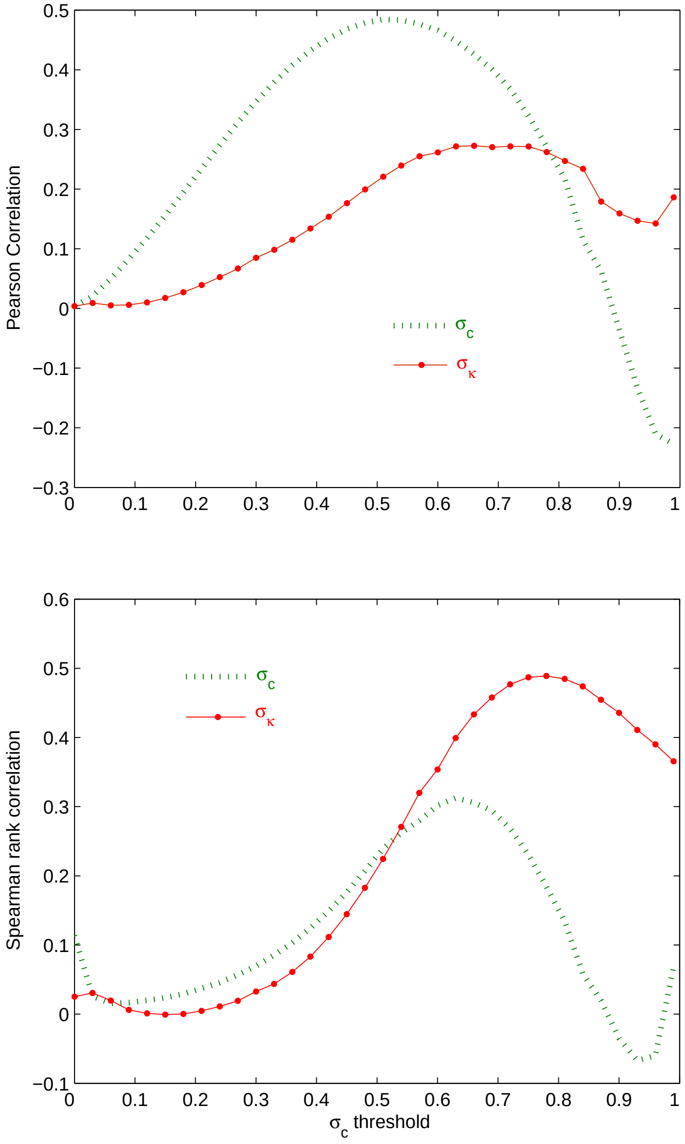


Figure 7: Pearson (top) and Spearman (bottom) correlations between graph-based semantic similarity σ_s^G and the two forms of content similarity, applying increasing thresholds on content similarity.

124,172 documents and 94,859 terms. The use of a sample considerably smaller than the one described in section 5.1 was necessary due to the higher time and memory resources required to compute σ_κ. We have compared the results obtained for the different functional combinations of content and link similarity described in section 5.1 using the original sample and the reduced one. In this analysis we have observed that the relative performance of the different functional combinations of content and link similarity remains essentially unaffected despite the reduction on the number of pages used in the evaluation. This justifies our use of a smaller corpus for this and future studies.

Based on this corpus, we generated a $200 \times 200 \times 200$ histogram with coordinates $(\sigma_c, \sigma_\kappa, \sigma_s^G)$ for 1.54×10^{10} pairs of pages. Figure 7 plots Pearson and Spearman correlations between σ_s^G and the two forms of content similarity, σ_c and σ_κ, versus a threshold on σ_c. For Pearson correlation we observe that σ_s^G is better correlated to σ_κ than to σ_c for $\sigma_c \geq 0.8$, while for Spearman correlation the improvement of σ_κ over σ_c can be observed when $\sigma_c \geq 0.5$. This result indicates that after filtering unrelated Web pages, the new measure of content similarity produces a better ranking of pages than the traditional measure of content similarity. This provides new supporting evidence for the usefulness of exploiting term co-occurrence to approximate semantic similarity. A subsequent analysis showed us that the product $f = \sigma_\kappa \sigma_\ell$ does not outperform $f = \sigma_c \sigma_\ell$ or $f = \sigma_\ell$, which once again highlights the superiority of link similarity as an approximation of semantic similarity.

5.3 Integrating Content and Link Similarity

An alternative way to approximate semantic similarity is based on *integrating* (rather than combining) content and link similarity. We have implemented a measure of similarity based on paths of length $L \leq 2$ links between pages, where the importance of a link in a path is adjusted by two weighting factors. First, we used the link IDF value to discount similarity if the link pointed to a page with many inlinks. Second, we used lexical similarity between pages to weaken the importance of those links connecting pages with low content similarity.

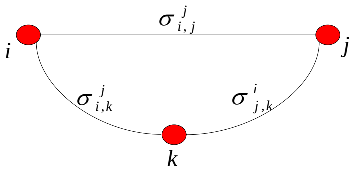


Figure 8: Diagram illustrating how two pages p_i and p_j can be connected to each other by a path of length $L \leq 2$.

In order to compute the similarity for a pair of pages (p_i, p_j) we represent p_i and p_j as vectors with elements of the form $\sigma^j_{i,k}$ and $\sigma^i_{j,k}$ respectively. As illustrated in Figure 8, the elements $\sigma^j_{i,k}$ and $\sigma^i_{j,k}$ are obtained by considering the pages p_k connected to both p_i and p_j by means of a link. Formally, consider undirected paths of length $L \leq 2$ between p_i and p_j, where for some k there exist links $p_i \rightarrow p_k$ or $p_k \rightarrow p_i$ and $p_j \rightarrow p_k$ or $p_k \rightarrow p_j$. Then, the values of the elements $\sigma^j_{i,k}$ are defined as follows:

$$\sigma^j_{i,k} = \begin{cases} 1 & \text{if } i = k, \\ \sigma_c(p_i, p_k) \cdot \text{IDF}(p_k)/2 & \text{if } p_i \rightarrow p_k, \\ \sigma_c(p_i, p_k) \cdot \text{IDF}(p_i)/2 & \text{if } p_k \rightarrow p_i, \\ \sigma_c(p_i, p_k) \cdot (\text{IDF}(p_i) + \text{IDF}(p_k))/2 & \text{if } p_i \rightarrow p_k \text{ and } p_k \rightarrow p_i. \end{cases}$$

Let $\vec{p}_1^{\,\sigma_1^2}$ and $\vec{p}_2^{\,\sigma_2^1}$ be the vector representation for a pair of pages (p_1, p_2). The integrated content and link similarity measure for these pages is computed as follows:

$$\sigma_{c_\ell}(p_1, p_2) = \frac{\vec{p}_1^{\,\sigma_1^2} \cdot \vec{p}_2^{\,\sigma_2^1}}{\|\vec{p}_1^{\,\sigma_1^2}\| \cdot \|\vec{p}_2^{\,\sigma_2^1}\|}$$

Once again, our measure of graph similarity σ^G_s was used to investigate if σ_{c_ℓ} is a good approximation of semantic similarity. To perform this analysis we used a subset of the ODP data discussed in section 5.1 consisting of 9.27×10^5 URLs and their corresponding outlinks and inlinks. Figure 9 illustrates the three possible cases that can occur when computing $\sigma_{c_\ell}(p_i, p_j)$: (1) there is a link from p_i to p_j (or viceversa), (2) p_i and p_j are connected to $p_k \in$ ODP, and (3) p_i and p_j are connected to $p_k \notin$ ODP. Because we only collected lexical information for pages inside the ODP, the measure $\sigma_c(p_i, p_k)$ required for the computation $\sigma^j_{i,k}$ was not available for pages p_k outside the ODP. In such cases, we used $\sigma_c(p_i, p_j)$ as a surrogate for $\sigma_c(p_i, p_k)$.

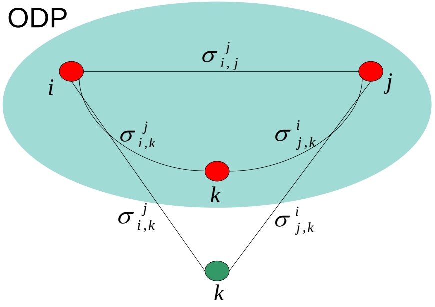


Figure 9: Diagram illustrating three cases of how two pages p_i and p_j in the ODP can be connected to each other by a path of length $L \leq 2$.

To complete our analysis we generated two $200 \times 200 \times 200$ histograms with coordinates $(\sigma_c, \sigma_{c_\ell}, \sigma^G_s)$ and $(\sigma_c, \sigma_\ell, \sigma^G_s)$ for 8.59×10^{11} pairs of pages. Figure 10 shows Pearson and Spearman correlations between σ^G_s and σ_{c_ℓ}, versus a threshold on σ_c. The correlations between σ^G_s and $\sigma_c\sigma_\ell$ are also shown for comparison. This preliminary evaluation suggests that σ_{c_ℓ} is not appreciably superior to the simpler and less computationally expensive $f = \sigma_c\sigma_\ell$.

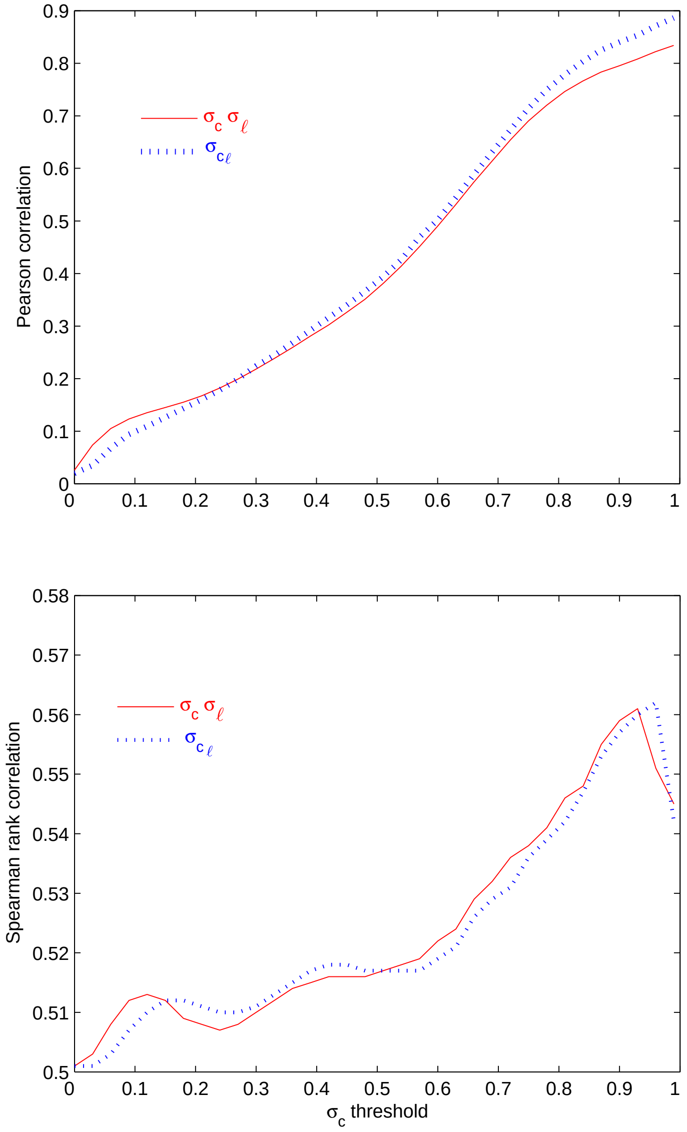


Figure 10: Pearson (top) and Spearman (bottom) correlations between σ_s^G and σ_{c_ℓ}, applying increasing thresholds on content similarity. The correlations between σ_s^G and $\sigma_c \sigma_\ell$ are plotted for comparison.

5.4 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 function to the baseline ranking obtained by the use of semantic similarity. The *sliding ratio* score [26, 16] compares two rankings when graded quality assessments are available.[6] 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.

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

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

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, $GSR(f, N) \rightarrow 1$ 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 11 we plot $GSR(f, N)$ versus N for the simple combination functions $f(\sigma_c, \sigma_\ell)$ introduced in Section 5.1. 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.

6 Discussion

In this paper we introduced a novel measure of semantic similarity for Web pages that generalizes the well-founded information-theoretic tree-based seman-

[6] 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.

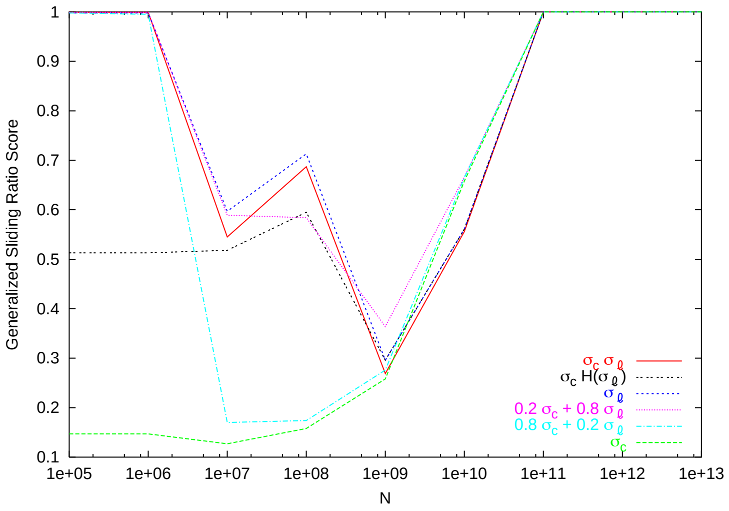


Figure 11: Generalized sliding ratio score plots for different functional combinations of content and link similarity. We omit the region $N < 10^5$ where *GSR* is constant for all f up to the resolution of our histogram bins.

tic similarity measure to the 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 Floyd-Warshall transitive closure 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 [32] 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 a dense matrices of size $571,148 \times 571,148$. 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 the 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 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 ($\sigma_c < 0.05$), 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.

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. LSI, 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. Phenomena such as the emergence of semantic network topologies may also be studied in the light of the proposed semantic similarity measure. For instance, we are currently using semantic similarity to evaluate adaptive peer based distributed search systems. In this evaluation framework queries and peers are associated with topics from the ODP ontology. This allows us to monitor the quality of a peer's neighbors over time by looking at whether a peer chooses "semantically" appropriate neighbors to route its queries.

In future work the semantic similarity measure should 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 indexing 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 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.

7 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 and G. Turner) 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.

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] N. Cristianini, J. Shawe-Taylor, and H. Lodhi. Latent semantic kernels. In C. Brodley and A. Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 66–73, Williams College, US, 2001. Morgan Kaufmann Publishers, San Francisco, US.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

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

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

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

[9] 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.

[10] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, 1998.

[11] C. Joslyn and W. J. Bruno. Weighted pseudo-distances for categorization in semantic hierarchies. In *International Conference on Conceptual Structures. Lecture Notes in Computer Science 3956*, pages 381–395, 2005.

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

[13] J. Kandola, J. Shawe-Taylor, and N. Cristianini. Learning semantic similarity. In *Neural Information Processing Systems 15*, pages 657–664, 2002.

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

[15] 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.

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

[17] 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.

[18] N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W.-Y. Ma. Learning similarity measures in non-orthogonal space. In *CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management*, pages 334–341, New York, NY, USA, 2004. ACM Press.

[19] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[20] 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.

[21] 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.

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

[23] F. Menczer. Mapping the semantics of web text and links. *IEEE Internet Computing*, 9(3):27–36, May/June 2005.

[24] M. Montes-y-Gómez, A. Gelbukh, A. López-López, and R. Baeza-Yates. Flexible comparison of conceptual graphs. In *Proc. 12th International Conference and Workshop on Database and Expert Systems Applications (DEXA)*, Lecture Notes in Computer Science. Springer-Verlag, 2001.

[25] T. Pedersen, S. Patwardhan, and J. Michelizzo. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 1024–1025, 2004.

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

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

[28] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

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

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

[31] 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.

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