A Semantic Framework for Evaluating Topical Search Methods

Rocío L. Cecchini Carlos M. Lorenzetti Ana G. Maguitman
Universidad Nacional del Sur, Dpto. de Ciencias e Ingeniería de la Computación
Bahía Blanca, Argentina, B8000JUQ
{*rlc,cml,agm*}*@cs.uns.edu.ar*

and

Filippo Menczer
Indiana University, School of Informatics and Computing
Bloomington, USA, 47408

Abstract

The absence of reliable and efficient techniques to evaluate information retrieval systems has become a bottleneck in the development of novel retrieval methods. In traditional approaches users or hired evaluators provide manual assessments of relevance. However these approaches are neither efficient nor reliable since they do not scale with the complexity and heterogeneity of available digital information. Automatic approaches, on the other hand, could be efficient but disregard semantic data, which is usually important to assess the actual performance of the evaluated methods. This article proposes to use topic ontologies and semantic similarity data derived from these ontologies to implement an automatic semantic evaluation framework for information retrieval systems. The use of semantic similarity data allows to capture the notion of partial relevance, generalizing traditional evaluation metrics, and giving rise to novel performance measures such as semantic precision and semantic harmonic mean. The validity of the approach is supported by user studies and the application of the proposed framework is illustrated with the evaluation of topical retrieval systems. The evaluated systems include a baseline, a supervised version of the Bo1 query refinement method and two multi-objective evolutionary algorithms for context-based retrieval. Finally, we discuss the advantages of applying evaluation metrics that account for semantic similarity data and partial relevance over existing metrics based on the notion of total relevance.

Keywords: information retrieval, evaluation, topic ontologies, semantic similarity.

1 Introduction

Information retrieval is the science of locating, from a large document collection, those documents that provide information on a given subject. Building test collections is a crucial aspect of information retrieval experimentation. The predominant approach used for the evaluation of information retrieval systems, first introduced in the Cranfield experiments [9], requires a collection of documents, a set of topics or queries, and a set of relevance judgments created by human assessors who mark the documents as relevant or irrelevant to a particular topic or query. However, reading large sets of document collections and judging them is expensive, especially when these documents cover diverse topics. In light of this difficulty a number of frameworks for automatic or semiautomatic evaluation have been proposed.

A common approach that has been applied in automatic evaluations is based on the use of pseudo-relevance judgments automatically computed from the retrieved documents themselves. A simple framework based on these ideas is the one proposed in [14]. In this approach the vector space model is used to represent queries and results. Then, the relevance of each result is estimated based on the similarity between the query vector and the result vector. Another approach for automatic evaluation uses a list of terms that are believed to be relevant to a query (*onTopic* list) and a list of irrelevant terms (*offTopic* list) [3]. This evaluation method scores every result d by considering the appearances of *onTopic* and *offTopic* terms in d. The authors show that their method is highly correlated with official TREC collections [29]. Click-through data have also been exploited to assess the effectiveness of retrieval

systems [16]. However, studies suggest that there is a bias inherent in this data: users tend to click on highly ranked documents regardless of their quality [6].

Editor-driven topic ontologies such as ODP[1] (Open Directory Project) have enabled the design of automatic evaluation frameworks. In [5] the ODP ontology is used to find sets of pseudo-relevant documents assuming that entries are relevant to a given query if their editor-entered titles match the query. Additionally, all entries in a leaf-level taxonomy category are relevant to a given query if the category title matches the query. Haveliwala et al. [12] defines a partial ordering on documents from the ODP ontology based on the ODP hierarchical structure. The inferred ordering is then used as a precompiled test collection to evaluate several strategies for similarity search on the Web. In another attempt to automatically assess the semantic relationship between Web pages, Menczer adapted Lin's information theoretic measure of similarity [15] and computed it over a large number of pairs of pages from ODP [22]. Lin's measure of similarity has several desirable properties and a solid theoretical justification. However, as it was the case for Haveliwala et al.'s ordering, the proposed measure is defined only in terms of the hierarchical component of the ODP ontology and fails to capture many semantic relationships induced by the ontology's non-hierarchical components (symbolic and related links). As a result, according to this measure, the 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.

In light of this limitation Maguitman et al. [20] proposed an information theoretic measure of semantic similarity that generalizes Lin's tree-based similarity to the case of a graph. This measure of similarity can be applied to objects stored in the nodes of arbitrary graphs, in particular topical ontologies that combine hierarchical and non-hierarchical components such as Yahoo!, ODP and their derivatives. Therefore, it can be usefully exploited to derive semantic relationships between millions of Web pages stored in these topical ontologies, giving way to the design of more precise automatic evaluation framework than those that are based only on the hierarchical component of these ontologies.

The goal of this article is to further evaluate this graph-based information theoretic measure of semantic similarity and to illustrate its application in the evaluation of topical search systems.

2 Topic Ontologies and Semantic Similarity

Web topic ontologies are means of classifying Web pages based on their content. In these ontologies, topics are typically organized in a hierarchical scheme in such a way that more specific topics are part of more general ones. In addition, it is possible to include cross-references to link different topics in a non-hierarchical scheme. The ODP ontology is one of the largest human-edited directory of the Web. It classifies millions of pages into a topical ontology combining a hierarchical and non-hierarchical scheme. This topical directory can be used to measure semantic relationships among massive numbers of pairs of Web pages or topics.

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. [24]). 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 (e.g., [25]).

To illustrate the limitations of the distance-based schemes take the ODP sample shown in Figure 1. While the edge-based distance between the topics JAPANESE GARDENS and COOKING is the same as the one between the topics JAPANESE GARDENS and BONSAI AND SUISEKI, it is clear that the semantic relationship between the second pair is stronger than the semantic relationship between the first pair. The reason for this stronger semantic relationship lays in the fact that the lowest common ancestor of the topics JAPANESE GARDENS and BONSAI AND SUISEKI is the topic GARDENS, a more specific topic than HOME, which is the lowest common ancestor of the topics JAPANESE GARDENS and COOKING. To address the issue of specificity, some proposals estimate semantic similarity in a taxonomy based on the notion of information content [25, 15]. In information theory [10], 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 a taxonomy by counting the fraction of objects stored in the subtree rooted at t (i.e., objects stored in node t and its descendants) out of all the objects in the taxonomy.

According to Lin's proposal [15], the semantic similarity between two topics t_1 and t_2 in a taxonomy is measured as the ratio between the meaning of their lowest common ancestor and their individual meanings. This can be expressed

[1] http://dmoz.org.

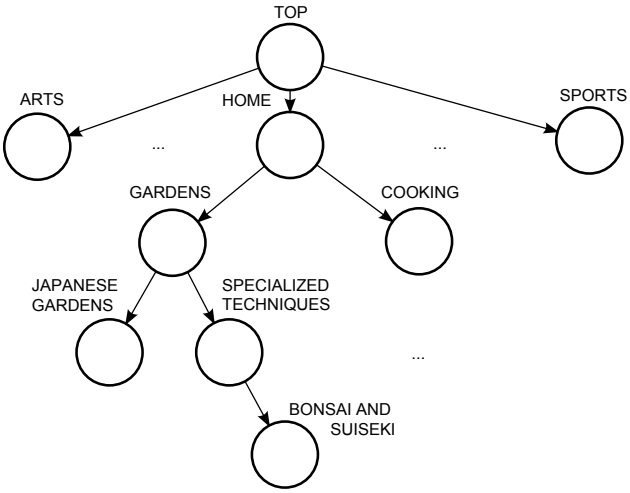

Figure 1: A portion of a topic taxonomy.

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 *topic*(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 $\sigma_s^T(\textit{topic}(d_1), \textit{topic}(d_2))$. To simplify notation, we use $\sigma_s^T(d_1, d_2)$ as a shorthand for $\sigma_s^T(\textit{topic}(d_1), \textit{topic}(d_2))$.

An important distinction between taxonomies and general topic ontologies such as ODP is that edges in a taxonomy are all "is-a" links, while in ODP edges can have diverse types such as "is-a", "symbolic" and "related". The existence of "symbolic" and "related" edges should be given due consideration as they have important implication in the semantic relationships between the topics linked by them. Consider for example the portion of the ODP shown in Figure 2. If only the taxonomy edges are considered, then the semantic similarity between the topics BONSAI AND SUISEKI and BONSAI would be zero, which does not reflect the strong semantic relationship existing between both topics.

To address this limitation Maguitman et al. [20] defined a graph-based semantic similarity measure σ_s^G that generalizes Lin's tree-based similarity σ_s^T to exploit both the hierarchical and non-hierarchical components of an ontology. In the following we recall the definitions that are necessary to characterize σ_s^G.

2.1 Defining and Computing a Graph-Based Semantic Similarity Measure

A topic ontology graph is a graph of nodes representing topics. Each node contains objects representing documents (Web 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 1 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. $|t_3| = 4$).

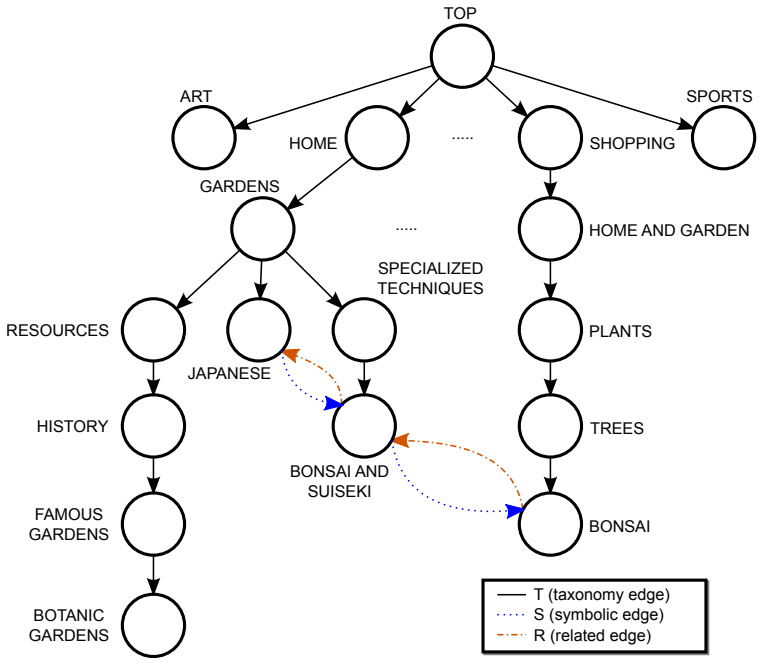

Figure 2: Illustration of a simple topic ontology.

The extension of σ_s^T to an ontology graph raises several questions: (1) how to deal with edges of diverse type in an ontology, (2) how to find the most specific common ancestor of a pair of topics, and (3) how to extend the definition of subtree rooted at a topic for the ontology case.

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 SHOPPING/HOME AND GARDEN/PLANTS/TREES/BONSAI also belong to topic HOME/GARDENS/SPECIALIZED TECHNIQUES/BONSAI AND SUISEKI. 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 there is a direct path in the graph G from t_i to t_j, where at most one edge from S or R participates in the path. 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 and is defined so that $\mathbf{T}_{ij} = \alpha$ if $(i, j) \in T$ and $\mathbf{T}_{ij} = 0$ otherwise. We use $\mathbf{T}$ with 1s on the diagonal (i.e., $\mathbf{T}_{ii} = 1$ for all i). Additional adjacency matrices are used 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 $\odot$ [13] defined on matrices as follows:[2]

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

Let $\mathbf{T}^{(1)} = \mathbf{T}$ and $\mathbf{T}^{(r+1)} = \mathbf{T}^{(r)} \odot \mathbf{T}$. 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 $t_j \in subtree(t_i)$, and $\mathbf{T}^+_{ij} = 0$ otherwise.

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.

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

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

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^G_s is a generalization of σ^T_s. 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 $\sigma^G_s(t_1, t_2)$ corresponds to the lowest common ancestor of t_1 and t_2.

The proposed semantic similarity measure σ^G_s as well as Lin's similarity measure σ^T_s was applied to the ODP ontology and computed for more than half million topic nodes. As a result, we obtained the semantic similarity values σ^G_s and σ^T_s for more than 1.26×10^{12} pairs of pages.[3] We found out that σ^G_s and σ^T_s are moderately correlated (Pearson coefficient $r_P = 0.51$). Further analysis indicated that the two measures give us estimates of semantic similarity that are quantitatively and qualitatively different (see [20] for details).

3 Validation

In [20] we reported a human-subject experiment to compare the proposed semantic similarity measure σ^G_s against Lin's measure σ^T_s. The goal of that experiment was to contrast the predictions of the two semantic similarity measures against human judgments of Web pages relatedness. 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. Measure σ^G_s was a better estimate of human-predictions in 84.65% of the cases while σ^T_s was a better predictor in 5.70% of the cases (the remaining 9.65% of the cases were undecided).

Although σ^G_s significantly improves the predictions made by σ^T_s, the study outlined above focuses on cases where σ^G_s and σ^T_s disagree. Thus it tells us that σ^G_s is more accurate than σ^T_s but is too biased to satisfactorily answer the broader question of how well σ^G_s predicts assessments of semantic similarity by human subjects in general.

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

[3] This required a computational effort of approximately 5,000 CPU hours using 20 nodes of a IU's AVIDD super-computing facility, resulting in 1 TB of data.

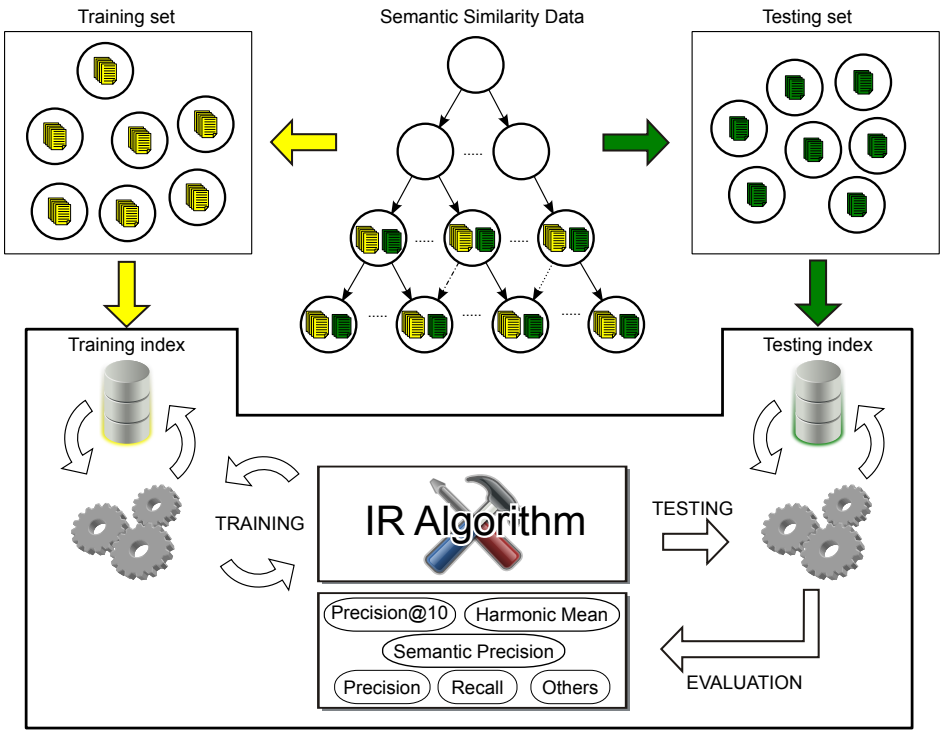

Figure 3: Framework for evaluating topical search.

3.1 Validation of σ_s^G as a Ranking Function

To provide stronger evidence supporting the effectiveness of σ_s^G as a predictor of human assessments of similarity, we conducted a new experiment. The goal of this new experiment was to determine if the rankings induced by σ_s^G were in accordance with rankings produced by humans.

Twenty volunteer subjects were recruited to answer questions about similarity rankings for Web pages. For each question, they were presented with a target Web page and three candidate Web pages that had to be ranked based to their similarity to the target page. The subjects had to answer by sorting the three candidate pages. 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 triplets of candidate Web pages. The candidate pages were selected with controlled differences in their semantic similarity to the target page, ensuring that there was a difference in σ_s^G of at least 0.1 among them. To ensure that the participants made their choice independently of the questions already answered, we randomized the order of the options. The result of the experiment was an average Spearman rank correlation coefficient $\rho = 0.73$.

4 Evaluation Framework based on ODP and Semantic Similarity

The general evaluation framework proposed in this article is depicted in Figure 3. The semantic similarity data, as well as the training and testing sets are shown at the top of the figure. The bottom of the figure illustrates the implemented framework and its components, which consist of a training index, a testing index, a set of evaluation metrics and a set of IR algorithms to be evaluated using the framework. The components of the framework will be described in this section.

The IR algorithms evaluated in this work are topical search algorithms. We define topical search as a process which goal is to retrieve resources relevant to a thematic context (e.g., [18]). The thematic context can consist of a document that is being editing or a Web page that is being visited. The availability of powerful search interfaces makes it possible to develop efficient topical search systems. Access to relevant material through these interfaces requires the submission of queries. As a consequence, learning to automatically formulate effective topical queries is an important research problem in the area of topical search.

In order to determine if a topical search system is effective we need to identify the set of relevant documents for a given topic. The classification of Web pages into topics as well as their semantic similarity derived from topical ontologies can be usefully exploited to build a test collection. In particular, these topical ontologies serve as a means to identify relevant (and partially relevant) documents for each topic. Once these relevance assessments are available, appropriate performance metrics that reflect different aspects of the effectiveness of topical search systems can be computed.

Consider the ODP topic ontology. Let R_t be the set containing all the documents associated with the subtree rooted at topic t (i.e., all documents associated with topic t and its subtopics). In addition, other topics in the ODP ontology could be semantically similar to t and hence the documents associated with these topics are partially relevant to t. We use $\sigma_s^G(t, \mathit{topic}(d))$ to refer to the semantic similarity between topic t and the topic assigned to document d. Additionally, we use A_q to refer to the set of documents returned by a search system using q as a query, while A_{q10} is the set of top-10 ranked documents returned for query q.

4.1 Evaluation Metrics

Information retrieval systems efficiency is measured by comparing its performance based on a common set of queries and a repository of documents. A document which answers a question or a topic is referred as a 'relevant' document. Effectiveness is a measure of the system's ability to satisfy the user needs in terms of the amount of relevant documents retrieved. The repository can be divided into two sets, the set of relevant documents for a given topic t, named R_t, and the set of non relevant documents. On the other hand, for a given query q, any information retrieval system recovers a set of documents, named the answer set A_q. Several classical information retrieval performance evaluation metrics have been proposed based on these two sets or its complements [28].

In order to evaluate the performance of a topical search system using the ODP ontology we could use the following metrics which are taken directly or are adapted from those traditional metrics.

4.1.1 Precision.

This well-known performance evaluation metric is computed as the fraction of retrieved documents which are known to be relevant to topic t:

$$\mathit{Precision}(q, t) = |A_q \cap R_t|/|A_q|.$$

4.1.2 Semantic Precision.

As mentioned above, other topics in the ontology could be semantically similar (and therefore partially relevant) to topic t. Therefore, we propose a measure of semantic precision defined as follows:

$$\mathit{Precision}_S(q, t) = \sum_{d \in A_q} \sigma_s^G(t, \mathit{topic}(d))/|A_q|.$$

Note that for all $d \in t$ we have that $\sigma_s^G(t, \mathit{topic}(d)) = 1$. Consequently $\mathit{Precision}_S$ can be seen as a generalization of *Precision*, where $\mathit{Precision}_S$ takes into account not only relevant but also partially relevant documents.

4.1.3 Precision at rank 10.

Since topical retrieval typically results in a large number of matches, sorted according to some criteria, rather than looking at precision, we can take precision at rank 10, which is computed as the fraction of the top 10 retrieved documents which are known to be relevant:

$$\mathit{Precision@10}(q, t) = |A_{q10} \cap R_t|/|A_{q10}|.$$

4.1.4 Semantic Precision at rank 10.

We compute semantic precision at rank 10 as a generalization of *Precision@10* by considering the fraction of the top ten retrieved documents which are known to be relevant or partially relevant to t:

$$\textit{Precision}_S@10(q,t) = \sum_{d \in A_{q10}} \sigma_s^G(t, \textit{topic}(d))/|A_{q10}|.$$

4.1.5 Recall.

We adopt the traditional performance measure of recall [4] as another criterion for evaluating query effectiveness. For a query q and a topic t, recall is defined as the fraction of relevant documents R_t that are in the answer set A_q:

$$\textit{Recall}(q,t) = \frac{|A_q \cap R_t|}{|R_t|}.$$

4.1.6 Harmonic Mean.

Finally, we use the function *F-score*, which is the weighted harmonic mean of precision and recall [4] that allows an absolute way to compare systems:

$$\textit{F-score}(q,t) = \frac{2 \cdot \textit{Precision}(q,t) \cdot \textit{Recall}(q,t)}{\textit{Precision}(q,t) + \textit{Recall}(q,t)}.$$

Other metrics could be used to compute a weighted harmonic mean. For example, we can compute *F-score@10* as follows:

$$\textit{F-score@10}(q,t) = \frac{2 \cdot \textit{Precision@10}(q,t) \cdot \textit{Recall}(q,t)}{\textit{Precision@10}(q,t) + \textit{Recall}(q,t)}.$$

In addition, we propose a weighted harmonic mean that takes into consideration partially relevant material by aggregating $\textit{Precision}_S$ and *Recall* as follows:

$$\textit{F-score}_S(q,t) = \frac{2 \cdot \textit{Precision}_S(q,t) \cdot \textit{Recall}(q,t)}{\textit{Precision}_S(q,t) + \textit{Recall}(q,t)}.$$

Analogously, we can define $\textit{F-score}_S@10(q,t)$ as follows:

$$\textit{F-score}_S@10(q,t) = \frac{2 \cdot \textit{Precision}_S@10(q,t) \cdot \textit{Recall}(q,t)}{\textit{Precision}_S@10(q,t) + \textit{Recall}(q,t)}.$$

4.2 A Short Description of the Evaluated Systems

In illustrating the application of the proposed evaluation framework we will focus on assessing the performance of supervised topical search systems. Supervised systems require explicit relevance feedback, which is typically obtained from users who indicate the relevance of each of the retrieved documents. The best-known algorithm for relevance feedback has been proposed by Rocchio [26]. Given an initial query vector $\overrightarrow{q}$ a modified query $\overrightarrow{q_m}$ is computed as follows:

$$\overrightarrow{q_m} = \alpha \overrightarrow{q} + \beta \sum_{\overrightarrow{d_j} \in R_t} \overrightarrow{d_j} - \gamma \sum_{\overrightarrow{d_j} \in I_t} \overrightarrow{d_j}.$$

where R_t and I_t are the sets of relevant and irrelevant documents respectively and α, β and γ are tuning parameters. A common strategy is to set α and β to a value greater than 0 and γ to 0, which yields a positive feedback strategy. When user relevance judgments are unavailable, the set R_t is initialized with the top k retrieved documents and I_t is set to $\emptyset$. This yields an unsupervised relevance feedback method.

4.2.1 The Bo1 and Bo1* Methods.

A successful generalization of Rocchio's method is the Divergence from Randomness mechanism with Bose-Einstein statistics (Bo1) [2]. To apply this model, we first need to assign weights to terms based on their informativeness. This

is estimated by the divergence between the term distribution in the top-ranked documents and a random distribution as follows:

$$w(t) = tf_x.log_2\frac{1+P_n}{P_n} + log_2(1+P_n)$$

where tf_x is the frequency of the query term in the top-ranked documents and P_n is the proportion of documents in the collection that contains t. Finally, the query is expanded by merging the most informative terms with the original query terms.

The main problem of the Bo1 query refinement method is that its effectiveness is correlated with the quality of the top-ranked documents returned by the first-pass retrieval. If relevance feedback is available, it is possible to implement a supervised version of the Bo1 method, which we will refer to as Bo1*. This new method is identical to the Bo1 method except that rather than considering the top-ranked documents to assign weights to terms, we look only at the top-ranked documents which are known to be relevant. Once the initial queries have been refined by applying the Bo1* method on the training set, they can be used on a different set. The Bo1* method can be regarded as a supervised version of the Bo1.

4.2.2 Multi-Objective Evolutionary Algorithms for Topical Search.

In [8] we presented a novel approach to learn topical queries that simultaneously satisfy multiple retrieval objectives. The proposed methods consist in training a Multi-Objective Evolutionary Algorithm (MOEA) that incrementally moves a population of queries towards the proposed objectives.

In order to run a MOEA for evolving topical queries we need to generate an initial population of queries. Each chromosome represents a query and each term corresponds to a gene that can be manipulated by the genetic operators. The vector-space model is used in this approach [4] and therefore each query is represented as a vector in term space.

In our tests, we used a portion of the ODP ontology to train the MOEAs and a different portion to test it. The initial queries were formed with a fixed number of terms extracted from the topic description available from the ODP. Documents from the training portion of ODP were used to build a training index, which was used to implement a search interface. Following the classical steps of evolutionary algorithms, the best queries have higher chances of being selected for subsequent generations and therefore as generations pass, queries associated with improved search results will predominate. Furthermore, the mating process continually combines these queries in new ways, generating ever more sophisticated solutions. Although all terms used to form the initial population of queries are part of the topic description, novel terms extracted from relevant documents can be included in the queries after mutation takes place. Mutation consists in replacing a randomly selected query term by another term obtained from a *mutation pool*. This pool initially contains terms extracted from the topic description and is incrementally updated with new terms from the relevant documents recovered by the system.

Although we have analyzed different evolutionary algorithms techniques following the above general schema, we will limit the evaluation reported here to two strategies:

- NSGA-II: Multiple objectives are simultaneously optimized with a different fitness function used for each objective. For this purpose we used NSGA-II (Nondominated Sorting Genetic Algorithm – II) [11], a MOEA based on the Pareto dominance concept (dominance is a partial order that could be established among vectors defined over an n-dimensional space). Some of the key aspects of NSGA-II are its diversity mechanism based on crowding distance, the application of elitism and its fast convergence. In our tests, NSGA-II attempted to maximize *Precision@10* and *Recall*.

- Aggregative MOEA: A single fitness function that aggregates multiple objectives as a scalar value is used. For that purpose, we have used the *F-score@10* measure introduced earlier.

Due to space limitations we refer the reader to [8] for details on the implementation of these MOEA strategies for topical search and focus here on how their performance was assessed using the proposed evaluation framework.

4.3 Evaluation Settings

Our evaluations were run on 448 topics from the third level of the ODP hierarchy. For each topic we collected all of its URLs as well as those in its subtopics. The language of the topics used for the evaluation was restricted to English and only topics with at least 100 URLs were considered. The total number of collected pages was more than 350,000.

We divided each topic in such a way that 2/3 of its pages were used to create a training index and 1/3 to create a testing index. The Terrier framework [23] was used to index these pages and to create a search engine. We used the stopword list provided by Terrier and Porter stemming was performed on all terms. In addition we took advantage of the ODP ontology structure to associate a semantic similarity measure to each pair of topics. In our evaluations we compared the performance of four topical search strategies that consisted in (1) queries generated directly from the initial topic description (baseline); (2) queries generated using the Bo1* query-refinement technique reviewed earlier in this article; (3) queries evolved using NSGA-II; and (4) queries evolved using the aggregative MOEA strategy.

Out of the 448 topics used to populate the indices, a subset of 110 randomly selected topics was used to evaluate the supervised topical search systems discussed in the previous section. For the training stage we run the MOEAs with a population of 250 queries, a crossover probability of 0.7 and a mutation probability of 0.03. The selection of values for these parameters was guided by previous studies [7]. For each analyzed topic the population of queries was randomly initialized using its ODP description. The size of each query was a random number between 1 and 32.

4.4 Evaluation Results

The charts presented on Figure 4 depict the query performance for the individual topics using *F-score$_S$@10*. Each of the 110 topics corresponds to a trial and is represented by a point. The point's vertical coordinate (z) corresponds to the performance of NSGA-II (chart on the left-hand side of the figure) or the aggregative MOEA (chart on the right-hand side of the figure), while the point's other two coordinates (x and y) correspond to the baseline and the Bo1* method. Note that different markers are used to illustrate the cases in which each of the tested methods performs better than the other two. In addition we can observe the projection of each point on the x-y, x-z and y-z planes. These charts show us that NSGA-II is superior to the baseline and the Bo1* method for 101 topics while the aggregative MOEA is the best method for 105 of the tested topics.

The systems were also evaluated using the *Precision@10* metric observing that NSGA-II outperforms both the baseline and Bo1* for 100 of the tested topics while the aggregative MOEA is the best method for 105 topics. Using the *Recall* metric, NSGA-II outperforms both the baseline and Bo1* for 96 of the tested topics while the aggregative MOEA is the best method for 91 topics. Due to space limitations we do not include charts for these measures.

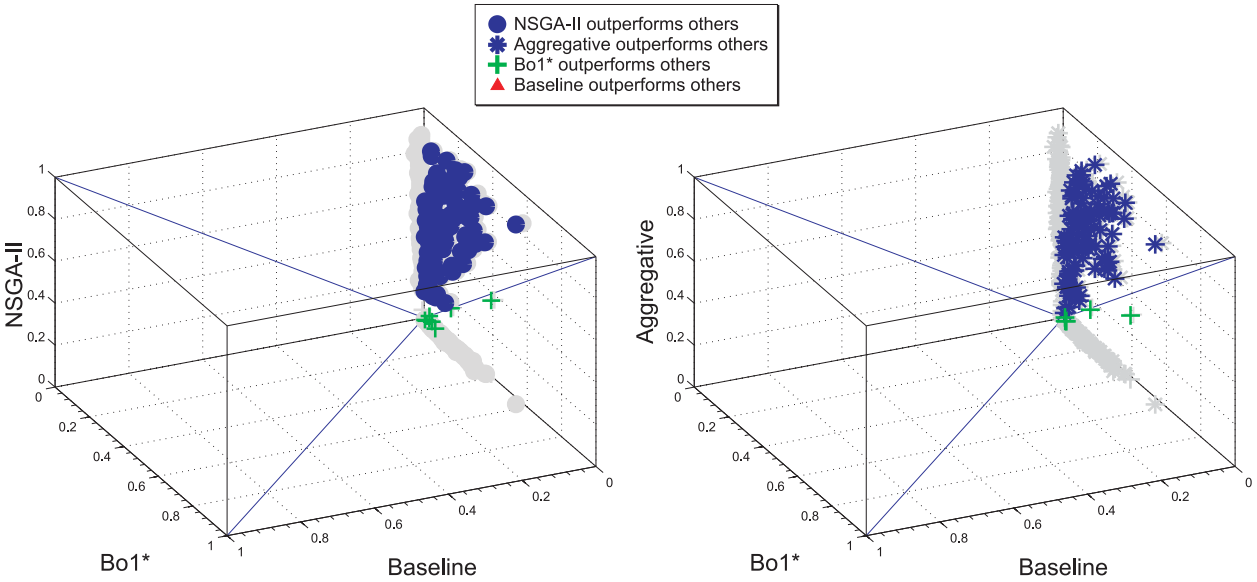

Figure 4: A comparison of the baseline, Bo1* and NSGA-II (left) and a comparison of the baseline, Bo1* and the aggregative MOEA (right) for 110 topics based on the *F-score$_S$@10* measure.

Table 1 presents the statistics comparing the performance of the baseline queries against the performance of the other strategies. From Table 1 we observe that the measures that have been extended with semantic similarity data appear to provide a more realistic account of the advantage of the various techniques over the baseline. The "soft" extended measures give more credit to all techniques, but relatively more to the baseline, so that the relative improve-

Average *Precision@10*			
	mean	95% C.I.	improvement
Baseline	0.015	[0.013,0.017]	–
Bo1*	0.138	[0.117,0.159]	802%
NSGA-II	0.536	[0.479,0.593]	3405%
Aggregative MOEA	0.496	[0.442,0.549]	3142%
Average *Precision*$_S$*@10*			
	mean	95% C.I.	improvement
Baseline	0.293	[0.283,0.303]	–
Bo1*	0.480	[0.452,0.508]	64%
NSGA-II	0.714	[0.658,0.770]	144%
Aggregative MOEA	0.759	[0.699,0.819]	160%
Average *Recall*			
	mean	95% C.I.	improvement
Baseline	0.051	[0.048,0.055]	–
Bo1*	0.440	[0.411,0.470]	768%
NSGA-II	0.586	[0.558,0.614]	1049%
Aggregative MOEA	0.559	[0.530,0.588]	990%
Average *F-score@10*			
	mean	95% C.I.	improvement
Baseline	0.008	[0.007,0.009]	–
Bo1*	0.141	[0.121,0.161]	1753%
NSGA-II	0.504	[0.460,0.547]	6519%
Aggregative MOEA	0.477	[0.436,0.518]	6170%
Average *F-score*$_S$*@10*			
	mean	95% C.I.	improvement
Baseline	0.074	[0.069,0.080]	–
Bo1*	0.459	[0.431,0.488]	520%
NSGA-II	0.622	[0.584,0.660]	736%
Aggregative MOEA	0.644	[0.601,0.687]	762%

Table 1: Baseline vs. queries refined with the Bo1* method, queries evolved with NSGA-II and queries evolved with the aggregative MOEA: mean, confidence intervals and improvement over the baseline for average query quality based on 110 topics.

ment appears smaller. This is indicated by the fact that the observed improvement of 160% in $Precision_S@10$ is more believable than one of 3142%. The same observation holds for an improvement in $F\text{-}score_S@10$ of 762% versus 6170%.

5 Conclusions

This paper addresses the problem of automatically evaluating topical retrieval systems using topical ontologies and semantic similarity data.

Evaluation has proved to play a crucial role in the development of search techniques, and heavily relies on telling apart relevant from irrelevant material, which is hard and expensive when performed manually.

After reviewing a definition of semantic similarity for topical ontologies and providing experimental evidence supporting its effectiveness, we have proposed an evaluation framework that includes classical and adapted performance metrics derived from semantic similarity data.

Semantic measures provide a better understanding of existing relationships between webpages and allow to find highly related documents that are not possible to discover with other techniques.

Metrics that rely on semantic similarity data have also been used in the evaluation of semi-supervised topical search systems [17]. However, the use of semantic similarity data does not need to be limited to the evaluation of topical retrieval system. In [19] semantic data is used to evaluate mechanisms for integrating and combining text and link analysis to derive measures of relevance that are in good agreement with semantic similarity. Phenomena such as the emergence of semantic network topologies have also been studied in the light of the proposed semantic similarity measure. For instance, it has been used 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 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 [1]. Semantic similarity data was also used for grounding the evaluation of similarity measures for social bookmarking and tagging systems [27, 21]. In the future, we expect to adapt the proposed framework to evaluate other information retrieval applications, such as classification and clustering algorithms.

Acknowledgement

This research work is supported by Universidad Nacional del Sur (PGI 24/ZN13) and Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 11220090100863).

References

[1] R. Akavipat, L.-S. Wu, F. Menczer, and A. G. Maguitman. Emerging semantic communities in peer web search. In *Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, P2PIR '06, pages 1–8, New York, NY, USA, 2006. ACM.

[2] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, October 2002.

[3] E. Amitay, D. Carmel, R. Lempel, and A. Soffer. Scaling ir-system evaluation using term relevance sets. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 10–17, New York, NY, USA, 2004. ACM.

[4] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co. Inc., Reading, MA, USA, May 1999.

[5] S. M. Beitzel, E. C. Jensen, A. Chowdhury, and D. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 17–23, New York, NY, USA, 2003. ACM.

[6] J. Boyan, D. Freitag, and T. Joachims. A machine learning architecture for optimizing web search engines. In *AAAI Workshop on Internet-based Information Systems*, 1996.

[7] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Using genetic algorithms to evolve a population of topical queries. *Information Processing and Management*, 44(6):1863–1878, 2008.

[8] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Multi-objective Evolutionary Algorithms for Context-based Search. *Journal of the American Society for Information Science and Technology*, 61(6):1258–1274, June 2010.

[9] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 3–12, New York, NY, USA, 1991. ACM.

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

[11] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, April 2002.

[12] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 432–442, New York, NY, USA, 2002. ACM.

[13] A. Kandel. *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley, Boston, MA, USA, 1986.

[14] L. Li and Y. Shang. A new method for automatic performance comparison of search engines. *World Wide Web*, 3(4):241–247, 2000.

[15] D. Lin. An Information-theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Franciso, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[16] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru. Automatic search engine performance evaluation with click-through data analysis. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1133–1134, New York, NY, USA, 2007. ACM.

[17] C. M. Lorenzetti and A. G. Maguitman. A semi-supervised incremental algorithm to automatically formulate topical queries. *Information Sciences*, 179(12):1881–1892, 2009. Including Special Issue on Web Search.

[18] A. G. Maguitman, D. B. Leake, and T. Reichherzer. Suggesting novel but related topics: towards context-based support for knowledge model extension. In *IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 207–214, New York, NY, USA, 2005. ACM Press.

[19] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani. Algorithmic computation and approximation of semantic similarity. *World Wide Web*, 9(4):431–456, 2006.

[20] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 107–116, New York, NY, USA, 2005. ACM.

[21] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 641–650, New York, NY, USA, 2009. ACM.

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

[23] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of the ACM Workshop on Open Source Information Retrieval (OSIR 2006)*, SIGIR '06, pages 18–24, New York, NY, USA, August 2006. ACM.

[24] R. 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.

[25] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[26] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[27] L. Stoilova, T. Holloway, B. Markines, A. G. Maguitman, and F. Menczer. Givealink: mining a semantic network of bookmarks for web search and recommendation. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 66–73, New York, NY, USA, 2005. ACM.

[28] C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979.

[29] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiments and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 2005.