Getting Lucky in Ontology Search: A Data-Driven Evaluation Framework for Ontology Ranking

Natalya F. Noy, Paul R. Alexander, Rave Harpaz,
Patricia L. Whetzel, Raymond W. Fergerson, and Mark A. Musen

Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA, 94305, USA
{noy,palexander,rharpaz,whetzel,rayferg,musen}@stanford.edu

Abstract. With hundreds, if not thousands, of ontologies available today in many different domains, ontology search and ranking has become an important and timely problem. When a user searches a collection of ontologies for her terms of interest, there are often dozens of ontologies that contain these terms. How does she know which ontology is the most relevant to her search? Our research group hosts BioPortal, a public repository of more than 330 ontologies in the biomedical domain. When a term that a user searches for is available in multiple ontologies, how do we rank the results and how do we measure how well our ranking works? In this paper, we develop an evaluation framework that enables developers to compare and analyze the performance of different ontology-ranking methods. Our framework is based on processing search logs and determining how often users select the top link that the search engine offers. We evaluate our framework by analyzing the data on BioPortal searches. We explore several different ranking algorithms and measure the effectiveness of each ranking by measuring how often users click on the highest ranked ontology. We collected log data from more than 4,800 BioPortal searches. Our results show that regardless of the ranking, in more than half the searches, users select the first link. Thus, it is even more critical to ensure that the ranking is appropriate if we want to have satisfied users. Our further analysis demonstrates that ranking ontologies based on page view data significantly improves the user experience, with an approximately 26% increase in the number of users who select the highest ranked ontology for the search.

1 "I'm Feeling Lucky" in Ontology Search

Consider a user who needs to find an ontology to use as a source of terms to annotate descriptions of clinical trials. She searches a library of ontologies [1], such as BioPortal, a public repository of more than 300 biomedical ontologies and terminologies [2]. She puts in a term "myocardial infarction"—her subject of interest. She receives 149 results in 32 ontologies. Twenty two ontologies contain a class named precisely "myocardial infarction" (with variation only in capitalization); other results have this phrase as synonyms of the class name, or have

H. Alani et al. (Eds.): ISWC 2013, Part I, LNCS 8218, pp. 444–459, 2013.

it in a property value. If our user is not familiar with the ontologies, how does she know which one of the 22 ontologies to use? Which one does everybody else use? Which one has more information about the terms that she is interested in? Naturally, to answer this question perfectly, we must know much more than our user's search term. It would help to know which task she is trying to achieve (e.g., annotation of text), what are her preferred ontologies, whether or not she requires conformance to specific standards, and so on. However, in many cases, we do not have this information; when a user searches an ontology library, the only information that we often have is the user's search term—and we must produce the best ranking of results based only on this information.

Ontology researchers have addressed the problem of ontology selection and ranking over the years. They have proposed a number of algorithms, which take into account the ontologies themselves, the search terms, and the repository as a whole. We review some of these approaches in Section 2. Researchers evaluated these approaches in small-scale user studies with hand-selected users.

In this paper, we propose a framework for evaluating the effectiveness of ontology ranking by using search logs. We analyze the position of the ontologies that the user selects after an ontology-search engine presents her with the search results. We use the position of that selection among the search results as a measure of the effectiveness of a ranking algorithm: the closer the user's selection is to the top-ranked result, the better the algorithm worked for this user. Our goal is to achieve a ranking in which most users feel "lucky" by following the top link, just as many of us do with Web search engines (e.g., Google and Bing). We evaluate our approach by using extensive search logs from the users who perform search on the BioPortal site over a period of several months. Specifically, this paper makes the following contributions:

- We propose a data-driven framework for evaluating ontology ranking based on user search logs.
- We propose several features for ontology ranking based on user behavior in BioPortal, an open community-based ontology repository. These features include pageviews, web service calls, comments left on the site, and others.
- We use our data-driven framework to evaluate the effect of different features on the ontology ranking based on search logs from four months of BioPortal searches (4,859 by users from 969 unique IP addresses).

2 Related Work in Ontology Ranking and Evaluation

The problem of finding the "best" ontology in response to a user's search consists of two main components: (1) *selecting* relevant ontologies from a collection and (2) *ranking* the results to present the most relevant ontologies first.

Over the past decade, researchers have developed many algorithms for selecting ontologies that are relevant to a user query. These algorithms use description logic reasoning [3], corpus analysis [4,5], graph matching [6] and other approaches in order to find the relevant ontologies. When traditional retrieval methods do not return sufficient results, algorithms use *query expansion* based on the hierarchy in the ontology [7], lexical-semantic relations [8], or statistical analyses [9]. In many cases, terms in more than one ontology match the user query, and therefore, we must rank the results in a way that we believe to be most meaningful to the user [10]. Researchers have explored links between ontologies [11], structure-based ranking [12], user ratings [13], and hybrid ranking based on several factors, such as frequency of search terms, where in the metadata the search results appear, and the type of the ontology [14].

A number of the studies of the methods for ontology search and ranking conducted some user evaluations. However, to the best of our knowledge, none of these works used the log analysis of user searches to evaluate the ranking. Furthermore, when researchers conducted user studies to evaluate how well the ranking worked (e.g., AKTiveRank [12]), these studies were based on the results from a small number of users. The high number of visitors to BioPortal (more than 100,000 page views and more than 60,000 unique visitors each month) allowed us for the first time to perform an analysis that used thousands of user searches. Thus, both the approach and the scale make our analysis unique.

3 The Framework for Data-Driven Evaluation of Ontology Ranking

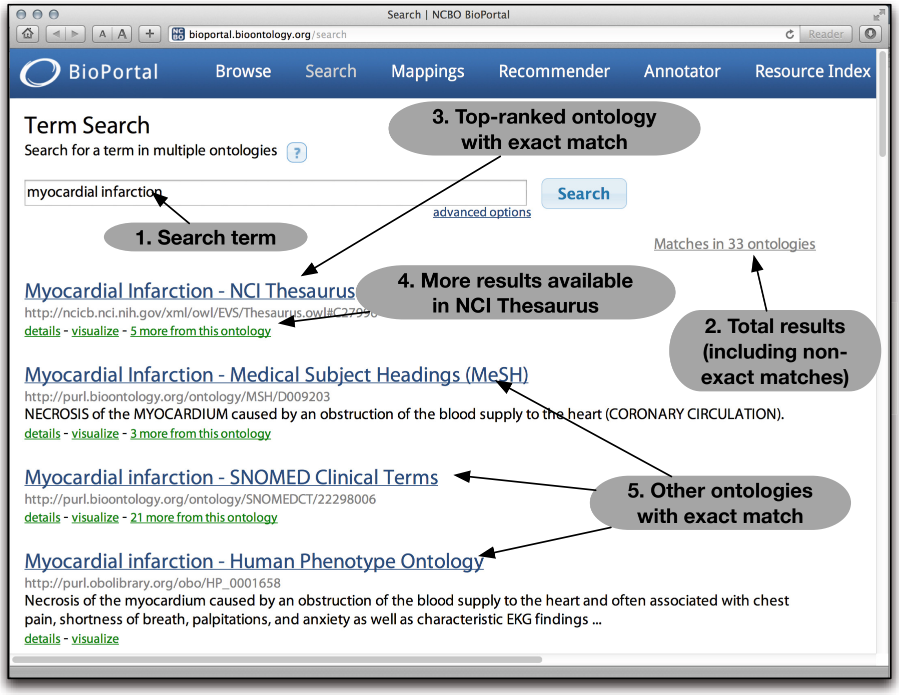
The basic idea in our framework is rather simple: when users search a collection of ontologies, our goal is for the user to find what she is looking for in the first result on the page. We use ontology ranking to order the search results and we record in the search log the position of the ontology that the user selected. The more users click on the first result, or the higher the average position that the users click on, the better the ontology ranking that we used to order the results. We explain our framework using the search in BioPortal as an example.

3.1 Ontology Search in BioPortal

BioPortal is a community-based repository of biomedical ontologies [15].[1] At the time of this writing, it contains more than 330 public ontologies with almost six million terms in them. Search across all ontologies is one of the key features of BioPortal. The system indexes all preferred names, synonyms, and property values for all classes across all ontologies. Users search against this index. The users can limit the search only to preferred names or ids of the terms, or choose to include property values. The users can search across all ontologies or in a group of ontologies of interest, or in a single ontology; they can choose to include or to exclude obsolete terms from the search, and so on.

For instance, Figure 1 shows the search results in BioPortal after the user has searched for “myocardial infarction” across all ontologies. The first 22 results correspond to the ontologies that have the exact term “myocardial infarction.” We group the result by ontologies. If an ontology has more than one class that

[1] `http://bioportal.bioontology.org`


Search | NCBO BioPortal
bioportal.bioontology.org/search
Reader
BioPortal
Browse
Search
Mappings
Recommender
Annotator
Resource Index
Search
advanced options
Matches in 33 ontologies
details - visualize - 5 more from this ontology
details - visualize - 3 more from this ontology
details - visualize - 21 more from this ontology
details - visualize

Term Search

Search for a term in multiple ontologies

myocardial infarction

1. Search term

2. Total results (including non-exact matches)

3. Top-ranked ontology with exact match

4. More results available in NCI Thesaurus

5. Other ontologies with exact match

Myocardial Infarction - NCI Thesaurus
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C2799

Myocardial Infarction - Medical Subject Headings (MeSH)
http://purl.bioontology.org/ontology/MSH/D009203
NECROSIS of the MYOCARDIUM caused by an obstruction of the blood supply to the heart (CORONARY CIRCULATION).

Myocardial infarction - SNOMED Clinical Terms
http://purl.bioontology.org/ontology/SNOMEDCT/22298006

Myocardial infarction - Human Phenotype Ontology
http://purl.obolibrary.org/obo/HP_0001658
Necrosis of the myocardium caused by an obstruction of the blood supply to the heart and often associated with chest pain, shortness of breath, palpitations, and anxiety as well as characteristic EKG findings ...

Fig. 1. Search results for "myocardial infarction" in BioPortal: 1. User searches for "myocardial infarction." 2. There are 33 ontologies that contain classes with names, term URIs, or property values that match the search term exactly or partially; of these, 22 ontologies have the exact match. 3. Among the ontologies with the exact match, the NCI Thesaurus has the highest ranking and BioPortal presents it first in the search results. 4. The NCI Thesaurus has 5 more results, which are not necessarily exact matches. 5. The order of other ontologies with exact matches (MeSH, SNOMED CT, etc.) corresponds to their ranking (Table 1, column *Pageviews*).

is relevant to the query, users can access these results by expanding the link for "more from this ontology." For instance, the top result, the NCI Thesaurus, has 12 more search results. The search result shows the pertinent information for the term that matched the user query exactly: the term label, the term URI, and a snippet of a textual definition of the term if the ontology has such definition. The user can also click on a link to have additional details about the term or to have a graph visualizing the neighborhood of the term to appear in a pop-up window. After the user examines the search results, she clicks on the result that seems most relevant to access the term in the ontology browser in BioPortal.

In the example in Figure 1, our search returned 22 ontologies that contain a class with preferred name matching the search string precisely. BioPortal has an ordered ranked list of all its ontologies, which we update regularly. Section 4 discusses the specific ranking approaches that we tested. For instance, the *API+Projects* column in Table 1 shows the top 10 ontologies in the ranking that BioPortal used when we took the screenshot for Figure 1. In this ranking, among the ontologies that had an exact match for the term "myocardial infarction," the highest rank belonged to NCI Thesaurus. The two ontologies that are ranked higher than NCI Thesaurus (column *API+Projects* in Table 1) do not contain the search term and hence do not appear in the search results.

The rest of the columns in Table 1 present the top 10 ontologies in other ranking orders that we evaluated (Section 4). In order to determine which ranking works better for our users, we recorded user actions in the search logs. Each time a user selects an ontology in the search results to open this ontology in the browser, we record the following data: the search term, the position that the user clicked, whether or not the result was an exact match or an approximate match, the ontologies that were ranked higher than the one that the user selected, the user IP address and other provenance information.

We use the position of the ontology that the user selected as a measure of how effective our ranking was for this particular search. If the user selects the first link and later finds out that this link is not what she was looking for, she will come back to the search results and follow a different link. We record both actions as two different searches.

In order to analyze the effectiveness of a specific ranking relative to another ranking, we compare the collection of positions of ontologies that the users select. We can compare the median and the mean of the position in a set of user search logs. The closer both numbers are to 1 (the user selecting only the highest ranked result), the closer our ranking is to a perfect one.

This framework provides a data-driven evaluation approach to ontology ranking. By varying the internal ranking R, we can compare the effect of various features in composing the ranking: given two rankings, R_i and R_j, the one with the lower mean and median of the positions of selected ontologies is the closer one to a perfect ranking.

3.2 Defining the Data-Driven Evaluation Framework

More formally, consider an ontology collection C and a set of ontologies $\{O_1, O_2, ...O_n\}$ in the collection C. We define a ranking R as a complete order on the set $\{O_1, O_2, ...O_n\}$. When a user searches the collection C for a term t (e.g., "myocardial infarction"), let the set C_t be the subset of ontologies from C that is returned as the result of the search for the term t. In the search results presented to the user, the ontologies in the set C_t are ordered according to the ranking R.

We define the **effectiveness of the ranking** R based on the user behavior after the search engine presents the ontologies in the set C_t ranked according to R. The ranking R is a **perfect ranking** if every user selects the first choice presented by the search engine. The closer the user behavior is to the perfect ranking, the more effective the ranking R is.

Table 1. The top 10 ontologies in each of the four ontology rankings that we used in the study. This ranking dictates the order of search results. The table presents four rankings: The first group are the top 10 ontologies based on pageviews in BioPortal; the second group presents the ranking based on combination of pageviews in BioPortal and API calls; the third group is the ranking based on API calls and use in projects submitted by users; the final group presents the ranking based on combination of all features. See Section 4 for details of the ranking features in.

Pageviews	Pageviews + API
1. National Drug File	1. SNOMED Clinical Terms
2. SNOMED Clinical Terms	2. NCI Thesaurus
3. MedDRA	3. Human disease ontology
4. International Classification of Diseases	4. MedDRA
5. NCI Thesaurus	5. International Classification of Diseases
6. Mouse adult gross anatomy	6. National Drug File
7. RadLex	7. Ontology for Biomedical Investigations
8. Bioinformatics operations... (EDAM)	8. Human Phenotype Ontology
9. Human disease ontology	9. Experimental Factor Ontology
10. RxNORM	10. Medical Subject Headings (MeSH)

API + Projects	All
1. Gene Ontology	1. NCI Thesaurus
2. Gene Ontology Extension	2. SNOMED Clinical Terms
3. NCI Thesaurus	3. Ontology for Biomedical Investigations
4. Medical Subject Headings (MeSH)	4. Human disease ontology
5. Ontology for Biomedical Investigations	5. RadLex
6. Foundational Model of Anatomy	6. Experimental Factor Ontology
7. SNOMED Clinical Terms	7. Medical Subject Headings (MeSH)
8. NCBI organismal classification	8. Foundational Model of Anatomy
9. Chemical entities of biological interest	9. NCBI organismal classification
10. Cell type	10. NIF Standard Ontology

3.3 Analyzing and Comparing Rankings

We use the search-log data to analyze the effectiveness of a specific ontology ranking and to compare the effectiveness of different rankings to one another. For our analysis, we use only the results that had the exact match for the search term—these results constitute the first batch of search results that BioPortal presents to users and it orders this set based on its current internal ontology ranking R. For each result, we take the position of the ontology that the user selected. For example, consider five entries in our search log for a period of time when a ranking R_i was active: Suppose one entry indicates that the user selected the ontology in position 2, another entry has the user selecting the ontology in position 10 for her search, and the three remaining entries have the users select the top link. Then, $P_{R_i} = \{2, 10, 1, 1, 1\}$. Thus, we get a set P_R of all positions of ontologies that users have selected over a period of time when the ranking R was active. We analyze the set P_{R_i} for each ranking R_i that we want to evaluate.

In order to analyze each individual ranking R_i, we compute the following metrics for the corresponding set P_{R_i}:

Median Selected Position: the median position that the user selects;

Mean Selected Position: the average value for the position of the ontology that users select; the closer this value is to 1, the closer our ranking is to a perfect ranking for ontology search.

Percentage of Selections in the Top Position: the fraction of users that have selected the top link among the results that the search engine presented.

We use a randomly generated ranking of ontologies R_{random} as a baseline. Presenting ontologies in a random order for several days allowed us to obtain the baseline for user behavior. We created this baseline in order to answer the question of how much the users tend to select the first result that we present, regardless of the ontology rank.

To compare rankings among one another, we performed a series of pair-wise statistical tests based on the Wilcoxon rank-sum test, followed by a Bonferroni correction to reduce the chance of type-I errors due to multiple comparisons. We first perform a one-sided Wilcoxon rank-sum test to determine whether each of the rankings R_i provides a statistically significant improvement over the randomly generated ranking R_{random} as determined by the two corresponding sets of selected ontology positions P_{R_i} and $P_{R_{random}}$ (Test 1). Here, the null hypothesis (H_0) is that the distributions of P_{R_i} and $P_{R_{random}}$ are identical. The alternative hypothesis (H_a) is that the distribution of $P_{R_{random}}$ is shifted to the right of P_{R_i}; in other words, ranking R_i is more effective than R_{random}. A small p-value in this case is an indicator that the location shift (i.e, ranking improvement) is unlikely to due to chance. We then compare each pair of rankings R_i , R_j to each other using a two-sided Wilcoxon rank-sum test to determine whether they are statistically different (Test 2). In this test, H_a is the hypothesis that the distributions of P_{R_i}, P_{R_j} are not identical (location shift is not equal to zero); or in other words, the distributions P_{R_i}, P_{R_j} are statistically different.

In the rest of this paper, we describe the application of this framework to analyze a number of ontology ranking features in BioPortal.

4 Features in BioPortal Ontology Ranking

We have actively solicited suggestions from our user community on what features to use in ranking BioPortal ontologies. As the result of these discussion, we selected the following list of features that could affect the ranking of ontologies:

Pageviews (PV): We use Google Analytics to measure the number of pageviews that each ontology in BioPortal receives. Because BioPortal allows users to browse multiple versions of the same ontology, we aggregate browsing history across versions: whichever version of an ontology O_{V_i} a user browses, those pageviews contribute to the browsing activity for the ontology O. We use an interval of one month each time to create a new ranking of BioPortal ontologies based on pageviews. This feature measures how frequently users browse an ontology in BioPortal: the more frequently the users browse a particular ontology, the higher its rank.

API Activity (API): Many developers use the NCBO Web services API [15] to access the ontologies from within their applications. Web service calls allow the caller to specify which ontology to use. For example, a group focusing on diseases may use all disease ontologies or specify only the ontologies that they consider to be the "best." The more frequently an ontology is explicitly specified in the Web service API calls, the higher its ranking along this feature. Specifically, we count the number of unique API keys (users) that access each ontology through the API.

Projects (Pr): BioPortal users can describe their ontology-related projects on the BioPortal site. The users can then link these project descriptions to the ontologies that they use in the projects. The more projects use an ontology, the higher its rank based on this feature.

Notes and Reviews (NR): BioPortal users can also provide reviews of ontologies and attach comments (notes) and new term requests to individual classes in an ontology. This activity is another indicator that we take into account to determine the ontology rank.

We ranked the ontologies based on each feature and then combined the *ranks* to create the ranking that relied on more than one feature. We could also add a weight to any of the features if we want to emphasize any one of them. In our experiments to date, we assigned each feature the same weight. We discuss additional features that we can include in ontology ranking in Section 6.

In our experiment, we evaluated the following ontology rankings, with each ranking being active for a period of time. For rankings that use multiple features, we added the ranks for each feature and based the combined ranking on this sum.

Random (R_{random}): provides a baseline for the user search behavior

Browsing Activity only (R_{PV}): reflects the interaction with BioPortal ontologies through the browser

Browsing Activity and API Activity (R_{PV+API}): reflects the general use of an ontology, through the pageviews or through API calls

API Activity and Number of Projects (R_{API+Pr}): reflects the use of the ontology in projects through measuring the explicit links between ontologies and projects as specified by the users on the BioPortal site and the use of the ontology in the API calls that developers make.

All of the Above (R_{All}): reflects a combination of all the measures that we studied. Specifically, it combines PV, API, and Pr, all with equal weights.

Using projects (R_{Pr}) or Notes and reviews (R_{NR}) alone did not differentiate the ontologies significantly, with 87% of the ontologies having at most one note or review. For R_{Pr}, 78% of ontologies had 4 or fewer projects. Thus, we did not yet use this feature by itself for the ranking in the live system. In future work, we plan to consider additional combination of features that take into account the features with low degree of differentiation, such as R_{NR}. Because there was some variability in the number of projects, with 22 different ranks, we used R_{Pr} in combination with R_{API}.

5 Results

We collected the data on user activity in BioPortal between January 1, 2013 and April 10, 2013 (Table 2).[2] The number of searches for each ranking ranged between 500 and 694. We considered only the searches where the user clicked on one of the ontologies with the exact match. This search behavior was affected the most by the rankings.

We describe the analysis of the features that we used for ranking (Section 5.1), search-log data in Section 5.2 and we compare the effects of features that we described in Section 4 on the effectiveness of ranking in Section 5.3.

5.1 Analysis of the Features

Figure 2 presents the ranges for the features that we considered for the ranking. Recall that when computing combined rank, we used the rank of ontologies for each feature rather than the absolute values for the features. The graphs show that the notes provided too little differentiation between ontologies and thus we did not use them in these experiments.

5.2 Analysis of the Search Data

In the period that we studied, the users performed the total of 4,859 searches. Of these searches, we analyzed the 3,029 searches (62%) where the user selected one of the ontologies with an exact match for the search term. These searches came from 969 unique IP address.

The users searched for 2,276 unique terms. In other words, more than 75% of the search terms appeared only once in searches over a period of 81 days.

The average number of ontologies that BioPortal returned for the searches in our analysis was 11 ontologies with exact matches for the user's search term.

BioPortal users can create an account on the site and log in to the site as they browse. Being logged in allows users, for example, to custom-tailor the set of ontologies that they see (e.g., by limiting this set only to the ontologies that they are interested in), to add reviews and comments on the ontologies, and to describe their projects. We found that only 3% of the searches were performed by users who were logged in to BioPortal during the search.

[2] The exact date when we pushed each new ranking to the BioPortal depended on the release schedule and other operational requirements, resulting in the slight variation in the number of days for each ranking. We decided to keep all the data rather than to truncate each period to 15 days in order to analyze as much data as possible.

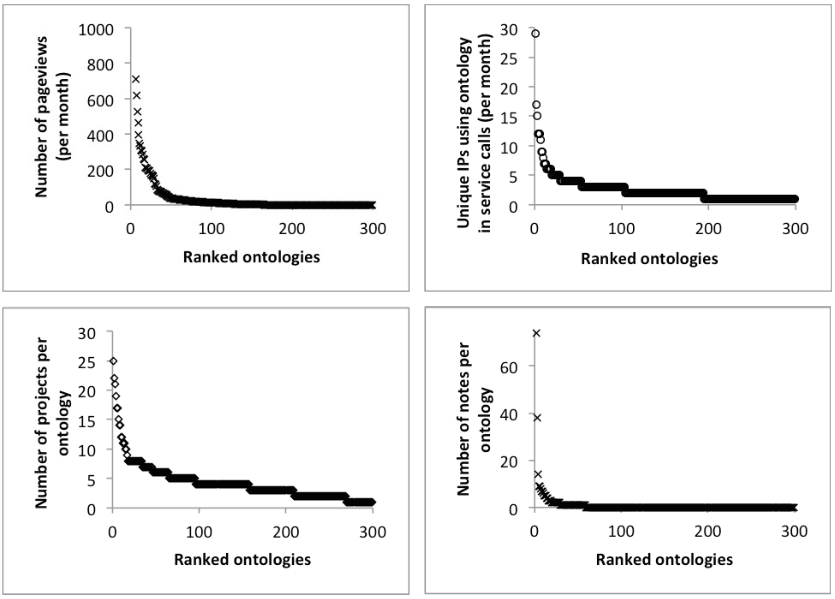


Fig. 2. The distribution of absolute values for the features. The pageviews provide the most discrimination among ontologies, whereas notes and reviews provide essentially none, with most ontologies having fewer than 2 notes. We used the relative rank of an ontology based on the specific feature rather than absolute values. The pageview plot excludes the top 5 ontologies; the monthly pageviews for these ontologies ranged from 1,000 to 10,000.

5.3 Comparing the Rankings

In each ranking that we considered, including the case when we ranked the ontologies randomly, the median position of the selected ontology was 1. In other words, more than half the time, users click on the first search result.

Table 3 displays p-values for Test 1 (Section 3.3), which we used to determine whether each of the four rankings (R_{PV}, R_{PV+API}, R_{API+Pr}, and R_{All}) provides a statistically significant improvement over a randomly generated ranking (R_{random}). According to the information in Table 3, there is strong statistical evidence (extremely small p-values) that the ranking improvement provided by the R_{PV} and R_{PV+API} ranking algorithms is unlikely due to chance (non-random). Furthermore, the p-values support the finding that the R_{PV} and R_{PV+API} algorithms provide performance that is superior to the other ranking algorithms. In other words, using pageviews or pageviews in combination with the API calls as the basis for ranking provides greater improvement compared to using API and projects (R_{API+Pr}) or the combination of all the features (R_{All}), which do not provide performance that is drastically different from the randomly generated ranking. Indeed, as Table 2 shows, the number of searches where the user select the ontology in the top position is 27% and 26% higher than random for R_{PV} and R_{PV+API}, respectively. For R_{PV}, almost 75% of searches result in the

Table 2. The summary information about the ranking algorithms used in the study

	Random	Pageviews	Pageviews +API	API + Projects	All
Period (all dates in 2013)	1/15-1/30	1/1-1/15	3/6-3/21	3/21-4/10	2/4-2/19
Number of days	16	15	15	20	15
Number of searches	500	589	694	639	607
Unique IP addresses	190	168	213	218	180
Searches by logged in users	4	13	11	29	43
Unique search terms	380	455	556	491	490
Unique search terms (%)	76.0%	77.2%	80.1%	76.8%	80.7%
Mean position selected	2.44	1.72	1.78	2.1	2.25
Users selecting top ontology	57.6%	74.4%	72.9%	63.9%	60.8%
Median position selected	1	1	1	1	1

Table 3. Comparing rankings to the random ranking. The p-values to test if improvement in ranking is due to chance (Test 1). The rankings that use Pageviews (R_{PV}) and Pageviews with API (R_{PV+API}) provide performance that is statistically significant.

	Pageviews	Pageviews + API	Projects+API	All
Random	1.26E-09	1.64E-09	0.01192	0.3306

selection of the top link. Notwithstanding, when comparing R_{PV} and R_{PV+API} to each other (Test 2, Section 3.3) we find that the two rankings are statistically indistinguishable from each other (p-value=0.67). This data suggests that combining the *API* feature with the *PV* feature does not provide a significant performance improvement over using the *PV* feature by itself.

6 Discussion

In this paper, we have developed a framework that enables us to evaluate ontology ranking algorithms in a data-driven way. Indeed, we need only to swap out one ranking for another and to continue to collect the data in order to compare different ranking. Because of the relatively high volume of searches on BioPortal, we get sufficient data to determine whether or not a ranking algorithm is working in a matter of a couple of weeks.

6.1 Changes in Ontology Ranking

We start our discussion by providing a sense of how much movement we observed in the four rankings of BioPortal ontologies that we presented in this study. There are more than 330 ontologies in BioPortal and their order differed significantly

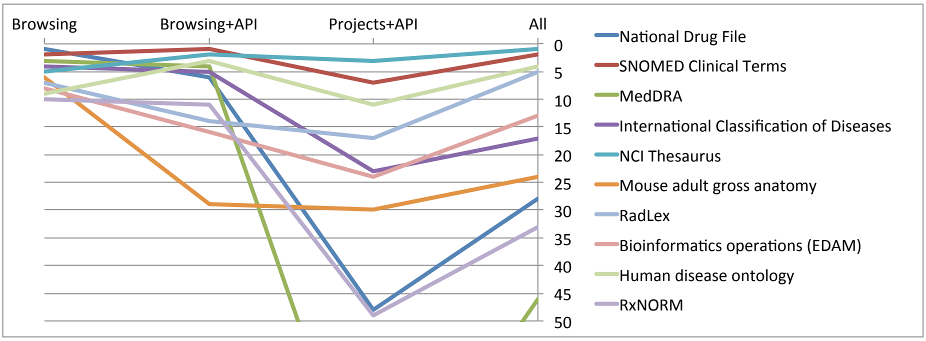


Fig. 3. Rank changes for the top 10 ontologies in the R_{PV} ranking (the ranking based on pageviews). Each line indicates the rank of the ontology based on the corresponding ranking algorithm. The graph captures the ranks between 1 and 50. The line for MedDRA (green) drops off the chart for the ranking based on Projects and API because MedDRA was ranked 89 in that ranking.

from one ranking to another. We compare the movement of ontologies in the rankings relative to the R_{PV} ranking, the ranking that performed the best in our evaluation. Consider the graph in Figure 3, which tracks the ranks of the top ten ontologies in the R_{PV} ranking. Each line represents the rank for a single ontology among these top ten, when we use the corresponding features for the ranking. The ranks for these ontologies ranged from 1 to 89 in the other rankings. We observed the biggest shift from the R_{PV} ranking in the R_{All} ranking, a ranking based on combination of all features. Indeed, the MedDRA terminology, which is ranked first based on page views, was ranked 89th in the ranking based on projects and APIs—an indication that while users often browse MedDRA in BioPortal, they do not use it in their ontology-related projects or access it through the BioPortal API.

Table 4 shows the average number of positions that the ontologies moved up or down relative to the R_{PV} ranking, for the top 100 ontologies. On average, each ontology that moved higher in the ranking, compared to R_{PV}, moved by 17.3 spots in the ranking. Each ontology that moved down in the ranking, moved by 60.7 spots, with the largest average movement between the ranking based on projects and API, R_{Pr+API}, and the ranking based on pageviews, R_{PV}. This result is not surprising because R_{Pr+API} is the only ranking among the ones that we considered that does not take pageviews into account.

6.2 Comparing the Rankings

Our analysis of the four ranking approaches for BioPortal ontologies demonstrated several trends. First, the majority of users select the top link, regardless of the ontology that it comes from. This observation is similar to the results that Joachims and colleagues [16] reported for regular Web search and what they referred to as "Trust bias." The fact that the user behavior changes as the

Table 4. The average movement distance (in the position change) for ontologies relative to the R_{PV} ranking. The data is for the top 100 ontologies in the R_{PV} ranking.

	Pageviews + API	Projects + API	All
Moving *higher* in the ranking	12	24	16
Moving *lower* in the ranking	-52	-73	-57

ranking changes confirms the "quality bias" reported by Joachims and colleagues: the quality of the ranking does affect the clicking behavior of the users. The trust bias appears to be more pronounced in ontology search than in regular web search, possibly because it is harder for users to assess the quality of the result from the snippets that BioPortal provides. For example, not all terms in ontologies have textual definitions, and therefore, the only information that the user might see is the term name and id. This information may not be enough to make informed decision.

Therefore, the better we are at putting the most relevant ontology at the top of the list, the more satisfied the users will be. Second, the rankings that performed the best in our experiments, R_{PV} and R_{PV+API}, were the ones that reflected the activity of users in the BioPortal user interface. In both rankings, the analysis of pageviews for an ontology played the key (or the only) role. This result is not surprising: indeed, the users who interact with the BioPortal search interface—the ones whose logs we used in the analysis—are exactly the users who browse BioPortal. The other rankings had a stronger component from the developers and users who already know which ontologies they need and thus were less helpful in ranking the ontologies in the user interface. These rankings did not improve the effectiveness of the search.

6.3 Other Condiserations

In our study, we focused on the users who perform ontology search. On the one hand, such filtering allowed us to rely on a smaller number of users who perform the same task [17]. At the same time, this decision led to several limitations.

First, if a user selected the top ontology, was not satisfied and then came back and selected a lower ranked one, we will record both selections in the log. This analysis is equivalent to the "click > skip above" strategy described Joachims and colleagues [16]. That work demonstrated that this strategy of assuming that the user finds any clicked result more relevant than the results above it, provide to be one of the most accurate strategies.

In reality, the user did not find what she was looking for in the ontologies that she selected first. Indeed, many users may not have precise or explicit criteria to select the ontology that will satisfy their needs and many of the searches might be exploratory. In order to be more precise about the satisfaction of the user, we may want to count only the last of the positions in a batch of selections from the same IP address with the same search term. Our initial analysis of the data indicates that this change will not have a significant effect on the results because the search logs are dominated by unique search terms. However, we plan to perform the detailed analysis that takes into account the history of consecutive selections from the same user.

Second, we are of course unlikely to have a ranking where every user will select the first search result because users have different requirements and might be interested in different ontologies. The best we can do is get the best result as the top result for as many users as possible. We could also use the user personal preferences and search history to custom-tailor the order. For instance, we can monitor the user's behavior and the ontologies that the specific user browses more frequently, and rank those ontologies higher for the specific user. Recall, however, that only 3% of the searches in our study came from the users who were logged in and "known" to the system.

Furthermore, we currently do not take the search results within the ontology into account: whether an ontology has several non-exact hits on the search term or only one does not effect its ranking for the specific search result. In the future, we can add this information to the ranking for a specific search.

We do not normalize pageviews–the key indicator in the ranking–by the ontology size, a decision that maybe counter-intuitive at first glance. However, it generally takes as much time on behalf of the user to perform X pageviews in a large ontology as it does in a small ontology. Because each page view corresponds to an explicit action by a user, this metric does not privilege large ontologies. However, because large ontologies have broader coverage and are more likely to appear in search results, uses might visit them more often for that reason. Large ontologies (e.g., SNOMED CT, ICD) also usually have some institutional support behind them and thus users are more likely to use those ontologies.

Finally, the ranking that we produce is only as good as the information that we use as input to the ranking. For instance, we believe that the project information is incomplete as many BioPortal users have not entered information for their projects. We are involved in an active outreach effort to expand the coverage of project descriptions. As these descriptions become more comprehensive, the effect of this feature on the ranking may change as well. Similarly, we we get more notes and reviews on the ontologies, that feature will differentiate the projects more and will have a different effect on the ranking. We plan to use our framework to re-evaluate the effects of these features continuously.

6.4 Future Work

Our analysis points to several future directions in improving ontology ranking methods—methods that we can continue testing in our framework. First, we can consider different weights on the features that go into the ranking. For example, we can weigh the rank based on pageview more, but still include other features. Second, we can use our framework to investigate a number of other features that can contribute to ontology ranking, in addition to the features that we have described in this paper. For example, we can consider the following features:

- the percentage of ontology terms that have textual definitions: if ontology developers took care of providing natural-language description for all, or most, of the terms, it might indicate that ontology is more useful for users;
- the number of other ontologies that import an ontology or reuse its terms: if an ontology is frequently reused, it might be ranked higher than others;
- coverage of a document corpus: we use ontologies to index records in many public datasets; an ontology where higher percentage of the terms that are reflected in large teal-life corpora may be more useful.

Our framework enables us to evaluate the effectiveness of ontology ranking for the purposes of ontology search. These result do not necessarily translate to a more general solution to ontology-evaluation. Indeed, as many researchers have pointed out, the best way to approach ontology evaluation is through task-specific evaluation [18]. While there is likely a correlation between the ranking for the purposes of improving the user search experience and more general ontology evaluation, we need to investigate this link in further research.

Note that these and other features and their positive or negative effect on ontology ranking are the hypotheses that we can test in our framework. Our results so far have demonstrated that some "common-sense" hypotheses do not necessarily hold if we analyze search data.

In our experiments, we focused exclusively on the search task. Analyzing the user behavior throughout the system, including their browsing of ontologies, will give us a more complete picture of user satisfaction. For example, the usage logs can reveal whether users explore multiple ontologies before settling on a single one. We can analyze how much time users spend on each ontology, how much time they spend on the pages following the search, and what actions they perform. Analyzing the data beyond the search page will give us a more complete picture of the user behavior and their implicit satisfaction with the search results.

7 Conclusions

Our framework provides an efficient way to compare various approaches to ontology ranking in a data-driven way by analyzing the user behavior in selecting search results. Our analysis of different ranking approaches for biomedical ontologies in BioPortal, shows that the majority of users always select the first search result, making good ontology ranking ever more important for user satisfaction.

Acknowledgements. This work was supported by the National Center for Biomedical Ontology, under grant U54 HG004028 from the National Institutes of Health. We are indebted to Barry Smith for encouraging us to look critically at the ranking of search results. We are grateful to our NCBO collaborators Helen Parkinson, Simon Jupp, and James Malone for their suggestions on features to use in ontology ranking.

References

1. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? A survey of ontology libraries. Journal of Web Semantics (JWS) 11, 96–111 (2011)
2. Musen, M.A., Noy, N.F., Shah, N.H., Whetzel, P.L., Chute, C.G., Storey, M.A., Smith, B.: The NCBO team: The National Center for Biomedical Ontology. Journal of American Medical Informatics Association 19, 190–195 (2012)
3. Pan, J.Z., Thomas, E., Sleeman, D.: Ontosearch2: Searching and querying web ontologies. In: IADIS International Conference WWW/Internet, pp. 211–219 (2006)
4. Buitelaar, P., Eigner, T., Declerck, T.: OntoSelect: A dynamic ontology library with support for ontology selection. In: Demo Session at the International Semantic Web Conference (ISWC 2004) (2004)
5. Alani, H., Noy, N.F., Shah, N.H., Shadbolt, N., Musen, M.A.: Searching ontologies based on content: experiments in the biomedical domain. In: 4th Int. Conf. on Knowledge capture (K-CAP 2007), pp. 55–62. KCAP, Whistler (2007)
6. Sabou, M., Lopez, V., Motta, E.: Ontology selection on the real semantic web: How to cover the queens birthday dinner? In: 15th Int. Conference on Knowledge Engineering and Knowledge Management (EKAW), Czech Republic (2006)
7. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
8. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: SIGIR 1994: 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 61–69. Springer-Verlag New York, Inc. (1994)
9. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)
10. Subhashini, R., Akilandeswari, J., Sinthuja, V.: Article: A review on ontology ranking algorithms. International Journal of Computer Applications 33(4), 6–11 (2011); Published by Foundation of Computer Science, New York, USA
11. Ding, L., et al.: Swoogle: A search and metadata engine for the semantic web. In: Conf. on Information and Knowledge Management (CIKM), Washington (2004)
12. Alani, H., Brewster, C., Shadbolt, N.R.: Ranking ontologies with AKTiveRank. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 1–15. Springer, Heidelberg (2006)
13. d'Aquin, M., Lewen, H.: Cupboard–a place to expose your ontologies to applications and the community. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 913–918. Springer, Heidelberg (2009)
14. Jonquet, C., Musen, M., Shah, N.H.: Building a biomedical ontology recommender web service. Journal of Biomedical Semantics 1(suppl. 1), S1 (2010)
15. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C.I., Tudorache, T., Musen, M.A.: BioPortal: Enhanced functionality via new web services. Nucleic Acids Research (NAR) 39(Web Server issue), W541–W545 (2011)
16. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: 28th Annual International ACM SIGIR Conference, pp. 154–161. ACM, Salvador (2005)
17. Kohavi, R., Henne, R.M., Sommerfield, D.: Practical guide to controlled experiments on the web: listen to your customers not to the HiPPO. In: 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Jose, CA (2007)
18. Hoehndorf, R., Dumontier, M., Gkoutos, G.V.: Evaluation of research in biomedical ontologies. Briefings in Bioinformatics (2012)