# Ranking Target Objects of Navigational Queries

Louiqa Raschid, Yao Wu,
Woei-Jyh Lee
University of Maryland
College Park, USA
louiqa@umiacs.umd.edu
yaowu@cs.umd.edu
adamlee@umiacs.umd.edu

María Esther Vidal
Universidad Simón Bolívar
Caracas, Venezuela
mvidal@umiacs.umd.edu

Panayiotis Tsaparas
University of Helsinki
Helsinki, Finland
tsaparas@cs.helsinki.fi

Padmini Srinivasan,
Aditya Kumar Sehgal
The University of Iowa
Iowa City, USA
padmini-srinivasan@uiowa.edu
sehgal@cs.uiowa.edu

## ABSTRACT

Web navigation plays an important role in exploring public interconnected data sources such as life science data. A navigational query in the life science graph produces a result graph which is a layered directed acyclic graph (DAG). Traversing the result paths in this graph reaches a target object set (TOS). The challenge for ranking the target objects is to provide recommendations that reflect the relative importance of the retrieved object, as well as its relevance to the specific query posed by the scientist. We present a metric layered graph PageRank (lgPR) to rank target objects based on the link structure of the result graph. LgPR is a modification of PageRank; it avoids random jumps to respect the path structure of the result graph. We also outline a metric layered graph ObjectRank (lgOR) which extends the metric ObjectRank to layered graphs. We then present an initial evaluation of lgPR. We perform experiments on a real-world graph of life sciences objects from NCBI and report on the ranking distribution produced by lgPR. We compare lgPR with PageRank. In order to understand the characteristics of lgPR, an expert compared the Top K target objects (publications in the PubMed source) produced by lgPR and a word-based ranking method that uses text features extracted from an external source (such as Entrez Gene) to rank publications.

## Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services – Web-based Services; H.2.8 [**Database Management**]: Database Applications – Data Mining

## General Terms

Algorithms, Experimentation

*WIDM'06,* November 10, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-525-8/06/0011 ...$5.00.

## Keywords

Ranking, Navigational Query, Link Analysis, PageRank

## 1. INTRODUCTION

The last few years have seen an explosion in the number of public Web accessible data sources, Web services and semantic Web applications. While this has occurred in many domains, biologists have taken the lead in making life science data public, and biologists spend a considerable amount of time navigating through the contents of these sources, to obtain information that is critical to their research.

Providing meaningful answers to queries on life science data sources poses some unique challenges. First, information about a scientific entity, e.g., genes, proteins, sequences and publications, may be available in a large number of autonomous sources and several sources may provide different descriptions of some entity such as a protein. Second, the links between scientific objects (links between data entries in the different sources) are important in this domain since they capture significant knowledge about the relationship and interactions between these objects. Third, interconnected data entries can be modeled as a large complex graph. Queries could be expressed as regular expression navigational queries and can more richly express a user's needs, compared to simpler keyword based queries.

Consider the following navigational query: *Retrieve publications related to the gene 'tnf' that are reached by traversing one intermediate (protein or sequence) entry.* This query expresses the scientist's need to expand a search for gene related publications beyond those publications whose text directly addresses the 'tnf' gene, while still limiting the search to publications that are closely linked to gene entries.

Consider gene sources OMIM Gene and Entrez Gene, protein sources NCBI Protein and SwissProt, sequences in NCBI Nucleotide and biomedical publications in PubMed. Figure 1 represents the results of evaluating this navigational query against these sources. The result is a layered DAG; we refer to it as a result graph (RG). All paths in this directed result graph (RG) start with data entries in the sources OMIM Gene or Entrez Gene; this is the first layer. They visit one intermediate data entry in sources NCBI Protein, Swiss Prot or NCBI Nucleotide (second layer) and they terminate in a publication data entry in PubMed (final layer).

The query returns all objects in PubMed that are reached

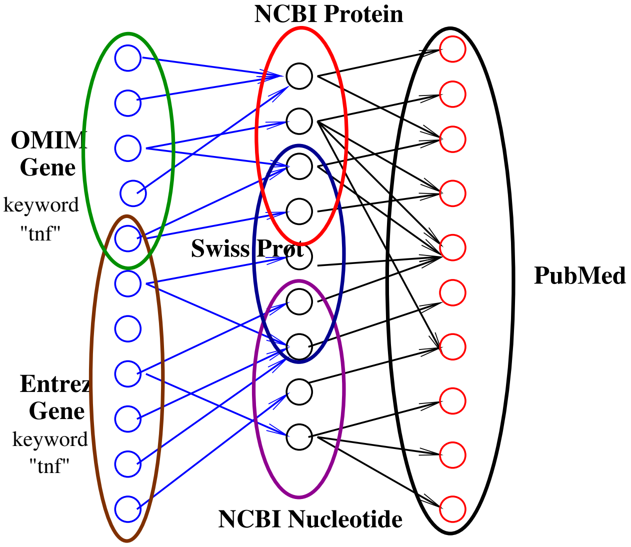

**Figure 1: An example of a result graph (RG)**

by traversing results paths; these PubMed entries are referred to as the *target object set* (TOS) reached by traversing the result paths of the RG. In contrast, a keyword based query would not have been able to specify the set of target publications. Navigational queries, the RG and the target object set (TOS) that answers the query are defined in the paper.

It is difficult for a user to explore all target objects in a reasonable amount of time and it is important to provide a ranking of the TOS. As is well known, word based ranking methods are very good at identifying the most *relevant* results, typically using features extracted from the contents of the target objects. For example [13] produces a ranking of documents in PubMed that are most relevant to a gene. In contrast, PageRank [11] focuses on the importance of the target object and importance is transferred from other important objects via the link structure. A recent technique ObjectRank [1] addresses both relevance and importance; it exploits schema knowledge to determine the correct authority transfer between important pages. We note that there is also research on ranking paths [2]. For term-based query dependent ranking, we refer to [3, 12].

The focus of this paper is to produce a ranking method to select the best target objects in the RG that answer the navigational query. Our ideal ranking must identify target objects that are both relevant and important. The ranking must also be *query dependent* since we must guarantee that the target objects that are ranked indeed occur in the RG and answer the navigational query. Further, both relevance and importance must be determined with respect to the objects in TOS, rather than with respect to all the data entries (as is the case with PageRank).

We propose two ranking metrics for the layered graph RG; they are layered graph PageRank (lgPR) and layered graph ObjectRank (lgOR). lgPR extends PageRank by distinguishing different roles (intermediate node, answer node) which can be played by the same node in the result graph. It does not perform random jumps so as to respect the RG. Our second metric lgOR is an extension to ObjectRank; due to space limitations we only discuss it briefly.

We report on our preliminary evaluation of lgPR on a real dataset from NCBI/NIH. For some navigational queries, we apply lgPR to the corresponding RG and use the ranking distribution for lgPR to illustrate that lgPR indeed discriminates among the TOS objects. We also apply the original PR metric to the object graph of life science data (against which we evaluate the query). We compare with applying lgPR to the actual RG to illustrate that lgPR and PR produce dissimilar rankings.

Finally, we report on an initial user experiment. We consider a set of complex queries typical of a scientist searching for gene related PubMed publications, and the Top K results of a word based ranking technique (Iowa) that has been shown to be accurate in answering gene queries [13]. We compare the Iowa Top K publications with the lgPR Top K publications, for some sample gene related queries, using criteria that reflect both relevance and importance. We use these criteria to understand the characteristics of lgPR.

The paper is organized as follows: Section 2 describes the data model, navigational query language and layered DAG result graph. Section 3 presents PageRank, lgPR, ObjectRank, and lgOR. 4 reports on preliminary results of an experimental study with NCBI data and concludes.

## 2. DATA MODEL

We briefly describe a data model and navigational query language for the life science graph. Details in [6, 9, 14].

### 2.1 Data Model for the Life Science Graph

The data model comprises three levels: ontology, source and data (Figure 2). At the ontology level, a domain ontology describes the universe of discourse, e.g., a gene, a pro-

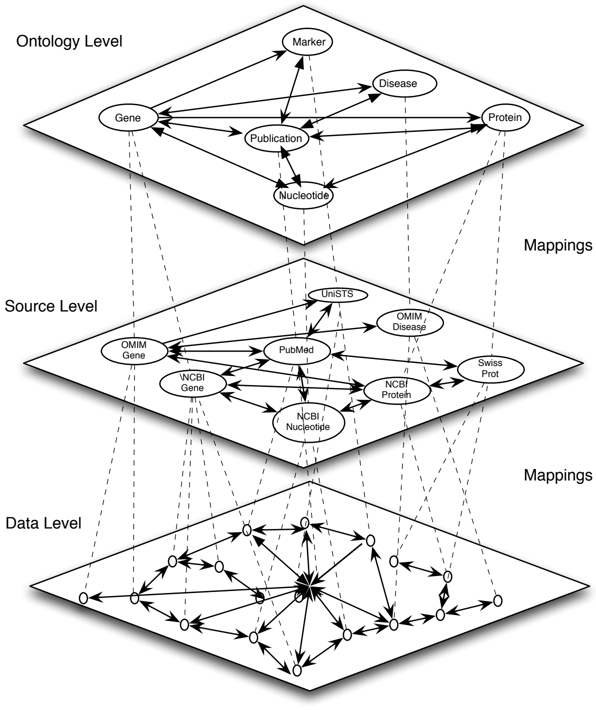

**Figure 2: A Data Model for the Life Science Graph**

tein, etc., and the relationships among them. An ontology graph $OG = (C, L_C)$ models the domain ontology, where nodes in $C$ represent classes, and edges in $L_C$ correspond to relationships among classes. For example, *genes* and *publications* are classes in $OG$ and the association *discuss* relates publications with genes. In this paper, we only consider one type of link, *isRelatedTo*, to capture the semantics of a relationship; therefore, we omit all link labels.

At the source level, a source graph $SG = (S, L_S)$ describes data sources and links that implement logical classes ($C$) and associations ($L_C$) in $OG$, respectively. For example, PubMed and Entrez Gene are sources that implement the logical classes *publications* and *genes*, respectively. A mapping defines logical classes in $C$ in terms of the sources in $S$ that implement them. A link between sources represents a hyper-link, a service or an application that connects these two sources. At the data level, a *Data Graph* is a graph $(D, L_D)$, where $D$ is a set of data entries and $L_D$ is a set of references between entries. A mapping $m_S$ establishes which data entries in $D$ are published by source $S$.

## 2.2 Navigational Query Language

We define a query as a path expression over the alphabet $C$ in $OG$, where each class occurrence can optionally be annotated with a Boolean expression. The simplest Boolean expression is the comparison of a `Field` to a particular value. In this paper, a field can be either `source` or `Object_content`, and the relational operators can be "=" for source and "contains" for Object_content. A condition over `source` and the relational operator "=", (source = "name-of-source"), restricts the query to some specific sources that implement the class. A condition on `Object_content` and the relational operator "contains", specifies the set of keywords that must occur within objects in the Data Graph. The symbol $\epsilon$ is a wild-card matching any class and the "." represents any relationship.

The query: *Retrieve publications that are related to the gene "tnf or aliases in human" in OMIM or Enrtez Gene, and are reached by traversing one intermediate resource*, is expressed in the navigational query language as follows: $Q =$ `Gene`[*Object_content* contains {"tnf" and aliases in human} and source = OMIM or Entrez Gene] $\cdot\ \epsilon\ \cdot$ `Publication`

The answer to a query $Q$ is defined at the three levels of the data model. It comprises three sets of paths: $\sigma_{OG}(Q)$, $\sigma_{SG}(Q)$ and $\sigma_{DG}(Q)$. The meaning of query $Q$ with respect to the ontology graph $OG$, $\sigma_{OG}(Q)$, is the set of simple paths in $OG$ that correspond to words in the language induced by the regular expression $Q$. The meaning of the query with respect to the source graph $SG$, $\sigma_{SG}(Q)$, is the set of all simple paths in $SG$ that correspond to mappings of the paths in $\sigma_{OG}(Q)$. Finally, the answer for query $Q$ with respect to the data graph $DG$, $\sigma_{DG}(Q)$, is the set of simple paths in $DG$ that are the result of mapping the paths in $\sigma_{SG}(Q)$ using mapping function $m_S$. A simple path does not repeat (revisit) the same class, data source or data entry (in the same path).

The queries that are presented in this section are typical queries posed by researchers. At present, there are no query evaluation engines to answer navigational queries and researchers must rely on manual navigation via browsers or they must write scripts; the latter involves labor to keep writing the scripts and the scripts may be inefficient in answering these queries.

## 2.3 Result Graph

The union of paths in $\sigma_{DG}(Q)$ is the result graph $RG$. We note that for our query language, all the paths that satisfy a query are of the same length, i.e., all the paths in the sets $\sigma_{OG}(Q)$, $\sigma_{SG}(Q)$ and $\sigma_{DG}(Q)$ are of the same length. We model a result graph $RG_Q = (D_{RG}, L_{RG})$ for a query $Q$, as a layered directed acyclic graph comprising $k$ layers, $L_1$, ..., $L_k$, where $k$ is determined by the query. The set of nodes $D_{RG}$ corresponds to the union of the data entries that appear in the paths in $\sigma_{DG}(Q)$. $L_{RG}$ represents the links among these data entries. A layer $L_i$ is composed of the union of the data entries in the paths $\sigma_{DG}(Q)$ that appear in the i-th position of the paths. The data entries in the k-th layer are called the target objects and they form the target object set (TOS) of the RG.

Note that since the result graph has multiple paths, and since a source may occur in different layers of these paths, the same data entry may appear multiple times in the different layers, depending on its connectivity to other data entries. In this case, each occurrence of the data entry is represented independently within each layer/path in which it occurs. The result graph framework distinguishes the different roles (intermediate node, answer node) which can be played by the same node in the result graph.

Figure 1 is a layered RG for the following query: *Retrieve publications related to the gene "tnf" traversing one intermediate source*; it has three layers. The first layer corresponds to the genes in the sources OMIM Gene and Entrez Gene that are related to the keyword "tnf". The second layer are the entries in the sources NCBI Protein, Swiss Prot or NCBI Nucleotide that are reached by objects in the first layer. Finally, the target objects in the third layer (TOS) are the publications in PubMed that are linked to the objects in the second layer.

# 3. RANKING METRICS

We briefly describe the PageRank metric [11] and then discuss our metric lgPR for layered DAGs. We briefly discuss the ObjectRank metric [1] and our extension lgOR.

## 3.1 PageRank

PageRank assumes that links between pages confer authority. A link from page $i$ to page $j$ is evidence that $i$ is suggesting that $j$ is important. The importance from page $i$ that is contributed to page $j$ is inversely proportional to the outdegree of $i$. Let $N_i$ be the outdegree of page $i$. The corresponding random walk on the directed web graph can be expressed by a transition matrix $A$ as follows:

$$A[i,j] = \begin{cases} \frac{1}{N_i} & \text{if there is an edge from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

Let $E$ be an arbitrary vector over the webpages, representing the initial probability of visiting a page. Let $d$ be the probability of following a link from a page and let $(1-d)$ be the probability of a random jump to a page. The PageRank ranking vector $R = dA \cdot R + (1-d)E$. $R$ converges for the web graph with any $E$, since generally the web graph is aperiodic and irreducible[5, 10].

PageRank cannot be directly applied to a layered graph. A Markov Chain is *irreducible* if and only if the graph contains only one strongly connected component. RG is not irreducible since the last layer in RG contains nodes with no outgoing links with respect to the query.

There are several potential ways to extend PageRank for RG. First, one can ignore links that point to pages without outgoing edges since these pages do not affect the ranking of other pages [11]. However we are specifically interested in obtaining a ranking for the TOS or the objects in the last layer of the layered result graph RG with no outgoing links, we cannot ignore these pages. Another possibility is modifying the transition matrix probability so that one takes a random jump from a node in the TOS [5]. This will ensure that the graph will be irreducible and aperiodic. However, this would arbitrarily modify RG whose structure is determined by the query; modifying RG will not assure that it answers the query. To summarize, the extensions to PageRank in the literature cannot be applied to the problem of ranking the target object set TOS of RG.

## 3.2 Layered Graph PageRank(lgPR)

We describe layered graph PageRank to rank the TOS.

### 3.2.1 The Metric

Table 1 lists the symbols used to compute lgPR.

| Symbol | Meaning |
|---|---|
| $RG(V_{RG}, E_{RG})$ | Result Graph, a layered DAG, with objects $V_{RG}$ and edges $E_{RG}$ |
| $e_{E_{RG}}$ | an edge in $E_{RG}$ |
| $R$ | ranking vector for objects in $RG$ |
| $R_{ini}$ | initial ranking vector |
| $A_{lg}$ | the transition matrix for objects in $RG$ |
| $k$ | the number of layers in the result graph |
| $OutDeg^{RG}(u_p)$ | outdegree from object $u$ at layer $p$ (across multiple link types to objects in layer $p+1$ |

**Table 1: Symbols used by lgPR**

The layered DAG result graph RG is represented by a transition matrix $A_{lg}$ to be defined next. Note that an object in the object graph may occur in multiple paths of the result graph, in different layers; it will be replicated in the transition matrix for each occurrence. Each object $u$ at layer $p$ will have an entry in the transition matrix to some object $v$ at layer $q$. We denote the occurrence of them as $u_p$ and $v_q$ respectively.

The ranking vector $R$ is defined by a transition matrix $A_{lg}$ and initial ranking vector $R_{ini}$, is as follows:

$$R = A_{lg}^{k-1} R_{ini} = (\prod_{l=1}^{k-1} A_{lg})\ R_{ini}$$

We pick $R_{ini}$ as follows: the entry for an object in $R_{ini}$ is 1 if this object is a link in start layer and 0 otherwise. The transition matrix $A_{lg}$ is computed as follows:

$$A_{lg}[i_u^p, j_v^q] = \begin{cases} \frac{1}{OutDeg^{RG}(u_p)} & \text{if } OutDeg^{RG}(u_p) > 0 \\ & \text{and } e(u_p, v_q) \in E_{RG}, \\ & \\ 0 & \text{otherwise.} \end{cases}$$

Note that we define the outdegree of each object in RG to only consider those edges that actually occur in RG and link to objects in the next layer. This reflect the probability that a user follows an object path in the RG. In contrast, PageRank considers all outgoing edges from a page.

Unlike PageRank, lgPR differentiates the occurrence of a data entry in different layers, as well as the links to entries in subsequent layers; lgPR is thus able to reflect the role of objects and links (from the entire graph of data entries) in answering a navigational query. Suppose an object $a$ occurs in an intermediate layer as well as in the TOS of the RG. It is possible that $a$ is able to convey authority to other objects in the TOS. However, $a$ may not rank very high in the TOS for this query. This characteristic is unique to lgPR. Thus, the score associated with the object is query dependent to reflect the role played by the object in the result graph.

### 3.2.2 Convergence Property

This transition matrix $A_{lg}$ is neither irreducible nor aperiodic as all rows for target objects contain only 0's. The matrix $A$ is a *nilpotent* matrix and the number of layers is the index. We provide two defintions (details in [8]).

Definition 3.1. *A square matrix $A$ is a nilpotent matrix, if there exists some positive integer $k$ such that $A^k = 0$ but $A^{k-1} \neq 0$. Integer $k$ is known as the index of $A$.*

Definition 3.2. *Let $k$ be the index of $A$. $\{A^{k-1}x, A^{k-2}x, ..., Ax, x\}$ form a* Jordan Chain*, where $x$ is any vector such that $A^{k-1}x \neq 0$.*

A characteristic of a nilpotent matrix is that its only eigenvalue is 0. The consequence is that any vector $x$ is an eigenvector of $A$ as long as $Ax = 0$. From the previous definition $\{A_{lg}^{k-1}R_{ini}, A_{lg}^{k-2}R_{ini}, ..., A_{lg}R_{ini}, A_{lg}\}$ forms a Jordan Chain, since $A_{lg}^{k-1}R_{ini} \neq 0$.

We show following two lemmas without providing proof in this paper.

Lemma 3.3. *Jordan chain $\{A_{lg}^{k-1}R_{ini}, A_{lg}^{k-2}R_{ini}, ..., A_{lg}R_{ini}, A_{lg}\}$ is a linearly independent set.*

Lemma 3.4. *$\{A_{lg}^{k-1}R_{ini}, A_{lg}^{k-2}R_{ini}, ..., A_{lg}R_{ini}, A_{lg}\}$ consists of a sequence of ranking vectors. In $R_{ini}$, only objects in layer 0 have non-zero scores; In ranking vector $A_{lg}^{m}R_{ini}$, only objects in layer $m$ receive non-zero scores.*

The final ranking vector by lgPR is the first eigenvector in the Jordan Chain, given the above initial ranking vector $R_{ini}$ and the transition matrix $A_{lg}$. While the traditional PageRank algorithm converges on a ranking in multiple iterations, lgPR can be computed in exactly $k-1$ iterations. Note that because $RG$ is a layered DAG, we can use link matrices, each of which represents links between neighboring layers, instead of the single transition matrix $A_{lg}$ for the entire graph. We also use keywords to filter query answers at each iteration.

## 3.3 Layered Graph ObjectRank(lgOR)

PR is computed a priori on the complete data graph and is independent of the RG. A recent technique ObjectRank [1] extends PageRank to consider relevance of query keywords. It exploits schema knowledge to determine the correct authority transfer in a schema graph. In ObjectRank, the authority flows between objects according to semantic connections. It does so by determining an *authority weight* for each edge in their schema graph. The ranking is (keyword) query dependent.

Due to space limitations, we do not provide the details of the ObjectRank metric. Instead, we briefly describe how the transition matrix for lgPR can be extended to consider the authority weights associated with edges that occur in RG.

Consider a metric layered graph ObjectRank(lgOR). The difference from lgPR is the transition matrix $A_{OG}$. It is as follows:

$$A[i_u^p, j_v^q] = \begin{cases} \alpha(e_{E_{RG}}) & \text{if } e(u_p, v_q) \in E_{RG}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\alpha(e_{E_{RG}}) = \begin{cases} \frac{\alpha(e_{E_{SG}})}{OutDeg(u_p, e_{E_{SG}})} & \text{if } OutDeg(u_p, e_{E_{SG}}) > 0 \\ 0 & \text{if } OutDeg(u_p, e_{E_{SG}}) = 0 \end{cases}$$

Let the edge between $u_p$ and $v_q$ map to an edge $E_{SG}$ in the $SG$. $\alpha(E_{SG})$ represents the authority transfer weight associated with $E_{SG}$. $OutDeg(u_p, e_{E_{SG}})$ is the outdegree in $RG$ of type $E_{SG}$.

As discussed in [1], the success of ObjectRank depends on correctly determining the authority weight to be associated with each link. Figure 3 (next section) illustrates the source graph that we use in our evaluation of navigational queries. For lgOR to be successful, an authority weight may have to be associated with each link in each result path (type) in the RG. Experiments with users to determine the correct authority weights for lgOR is planned for future work.

Currently the importance is computed after query evaluation. We compute result graph first, then ranking, for the reason that the transition matrix is defined in terms of outdegree in the RG. This motivates further research of combination of two problems, whose ideal solution is to ranking objects during query evaluation.

## 4. EXPERIMENTS ON LGPR

We report on experiments on real world data. We show that the lgPR ranking distribution has the ability to differentiate among the target objects of the RG and it is different from PageRank. A user compared the Top K results of lgPR and a word based ranking (Iowa) [13], using criteria that reflect both importance and relevance, to determine their characteristics.

### 4.1 Experiment Setting

NCBI/NIH is the gatekeeper for biological data produced using federal funds in the US[1]. We consider a source graph $SG$ of 10 data sources and 46 links. Figure 3 presents the source graph used in this task.We used several hundred keywords to sample data from these sources (the EFetch utility) and followed links to the other sources (the ELink utility). We created a data graph of approximately 28.5 million objects and 19.4 million links. We note that several objects are machine predicted objects so it is not uncommon that they have no links. The object identifiers for the data entries (nodes of the data graph) and the pair of object identifiers (links) were stored in a DB2 relational database.

Table 2 identifies the queries and keywords that were used in this experiment. The symbols $g$, $p$, $n$, $s$ refer to classes gene, publication, nucleotide and SNP, respectively. Note that $\epsilon$ is the wild card and can match all the classes and sources (in the source graph).

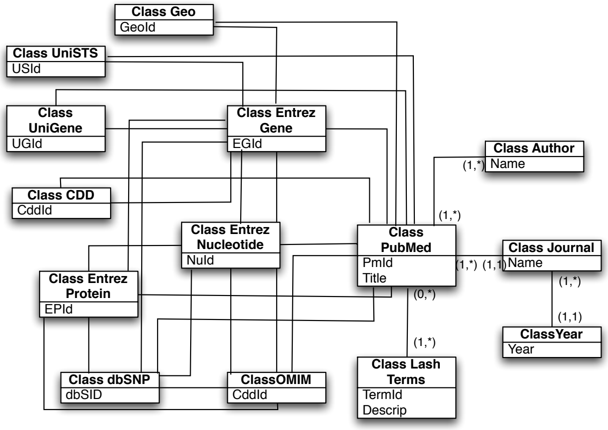

**Figure 3: Source Graph for User Evaluation**

| Queries | $g.n.p$, $g.s.p$, $g.n.s.p$, $g.s.n.p$, $g.s.g.n.p$, $g.s.n.g.p$, $g.\epsilon.p$, $g.\epsilon.\epsilon.p$, $g.\epsilon.\epsilon.\epsilon.p$ |
|---|---|
| Keywords | “parkinson disease”, “aging”,“cancer”<br>“diabetes”, “flutamide”, “stress”<br>“degenerative joint”,“tnf”,”insulin”<br>“fluorouracil”, “osteoarthritis”,“sarcoma” |

**Table 2: Experiment setting**

For each navigational query, the source paths that answer the query were determined using an algorithm described in [14]. Evaluating the paths in the data graph for each source path was implemented by SQL queries. Since a result graph RG could involve multiple source paths whose computation may overlap we applied several multiple query optimization techniques. The SQL queries were executed on DB2 Enterprise Server V8.2 installed on a 3.2 GHz Intel Xeon processor with 1GB RAM. The execution time for these queries varied considerably, depending on the size and shape of RG. If we consider the query g.n.p with keyword “degenerative joint” used to filter ’g’, one source path was ranked in approximately 1 second. However, the query (g.$\epsilon$.$\epsilon$.$\epsilon$.p) with the keyword `aging` used to filter ’g’ created a very large result graph and the execution time for this was approximately 2000 seconds. Typically the We note that computing the high scoring TOS objects of the RG efficiently is a related but distinct optimization problem.

### 4.2 lgPR Distribution

We report on the query ($g.\epsilon.p$), i.e., paths from genes to publications via one intermediate source.

Figures 4 and 5 report on the distribution of scores produced by the lgPR metric for the target objects in TOS for some representative queries. The first 10 bars represent scores in the range (0.00-0.01) to (0.09-0.1) and the last bar represents the range (0.1-1.0). Fig 4 shows that a small number of objects have very high score and the majority have a low score. As expected, many queries and keywords produced distributions that were similar to Figure 4. Most of the objects in TOS, in this case approx. 12,000 objects, had a very low score, and less than 200 object had a score in the range (0.1-1.0).

However, we made an interesting observation that some queries produced distributions that were similar to Figure 5. In this case, while many of the results (approx. 120) had low scores in the range (0.00-0.02), 46 objects had scores in the range (0.1-1.0) and 120 objects had scores in between.

[1] `www.ncbi.nlm.nih.gov`

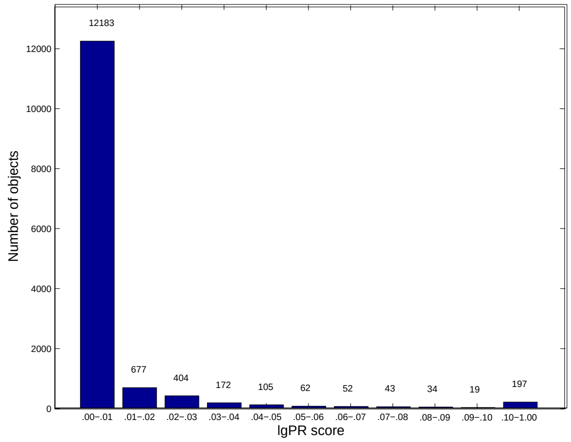

**Figure 4: Histogram for query: g[Object_content contains "aging"]$\cdot\epsilon\cdot$p**

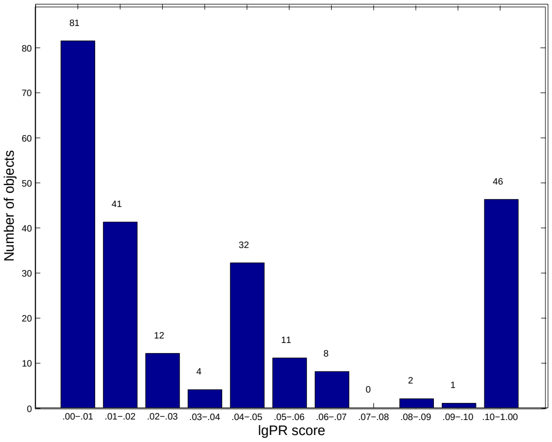

**Figure 5: Histogram for query: g [Object_content contains "degenerative joint"]$\cdot\epsilon\cdot$p**

Finally, we compared the ranking produced by lgPR and PageRank. We apply PageRank to the entire data graph of 28.5 million objects and 19.4 million links described in section 4.1. For the three sample queries (described in the next section), there are no PubMed ID's in common to the Top 25, 50, 100 for each of the queries, except that the top 50 of query with Lash term "allele" have 1 PubMed publications in common, and top 100 of same query have 3 in common. We speculate that the link structure of the RG is distinct compared to the link structure of the data graph; hence applying lgPR to the RG results in dissimilar ranking compared to a priori applying PageRank to the entire data graph.

We summarize that the lgPR score can both identify those objects with a very low ranking that may not be of interest to the user. However, it can also be used to discriminate amongst objects in the TOS whose ranking has a much lower variation of scores. Finally, lgPR ranking is not the same as that produced by PageRank applied to the entire data graph.

## 4.3 User Evaluation

In our user evaluation of lgPR, we consider a set of complex queries typical of a scientist searching for gene related PubMed publications, and the Top K results of a word based ranking technique (Iowa) that has been shown to be accurate in answering gene queries [13]. We compare the Iowa Top K publications with the lgPR Top K publications, for some sample gene related queries. We use criteria that reflect both relevance and importance to identify characteristics of lgPR.

Researchers are particularly interested in genetic and phenotypic variations associated with genes; these phenomena are often studied in the context of diseases, in a chromosomal region identified by a genomic marker (a unique known sequence) associated with the disease. Genetic and phenotypic knowledge are described using terms of the Lash controlled vocabulary [7]. We focus on a branch of the Lash vocabulary that relates to phenotypes and population genetics. Terms of interest include `linkage disequilibrium`, `quantitative trait locus` and `allele`. Figure 6 presents a portion of the Lash controlled vocabulary (term hierarchy). `LD` is not listed as the synonym to the term `linkage disequilibrium`, because `LD` may often refer to another concept. In the following experiment, we did not consider the plural form of some terms, such as *alleles* to *allele*, but this can be extended in the future studies.

1. EPIGENETIC ALTERATION
   ...
2. GENOMIC SEGMENT LOSS
   ...
3. GENOMIC SEGMENT GAIN
   ...
4. GENOMIC SEQUENCE ALTERATION
   ...
5. PHENOTYPIC ASSOCIATION
   (synonym: phenotype, trait)
   (a) locus association (synonym: locus, loci)
      i. linkage
      ii. quantitative trait locus (synonym: QTL)
   (b) allelic association (synonym: allele)
      i. linkage disequilibrium

**Figure 6: Branch 5 in Hierarchical controlled vocabulary of genetics terms (Lash Controlled Vocabulary)**

The navigational query used in our evaluation experiment can be described in English as follows: "*Return all publications in PubMed that are linked to an Entrez Gene entry that is related to the human gene TNF (or its aliases). The entry in PubMed must contain an STS marker and a term from the Lash controlled vocabulary.*"

We used the query term "*TNF AND 9606[TAXID]*" [2] to sample data from Entrez Gene. We then followed 8 paths to PubMed. Table 3 reports on the number of entries in Entrez Gene as well as the cardinality of the TOS for some sample queries [3].

We briefly describe the word-based ranking method (Iowa) that focuses on ranking documents retrieved by PubMed

[2] Note the Taxonomy ID for human is 9606 [4], and term *9606[TAXID]* was used to select human genes.

[3] We use g["tnf" and aliases in human] to denote g[Object_content contains {"tnf" and aliases in human}]; the entries in the first column of Table 3 are similar.

| Query | Cardinality of TOS |
|---|---|
| g[“tnf” and aliases in human] | 649 |
| g[“tnf” and aliases in human] $\cdot$ $\epsilon$·p[STS marker and “allele”] | 2777 |
| g[“tnf” and aliases in human] $\cdot$ $\epsilon$·p[STS marker and “linkage disequilibrium”] | 257 |
| g[“tnf” and aliases in human] $\cdot$ $\epsilon$·p[STS marker and “quantitative trait locus”] | 22 |

**Table 3: Cardinality of TOS**

for human gene queries [13], so that relevant documents are ranked higher than non-relevant documents. This method relies on using post-retrieval queries (ranking queries), automatically generated from an external source, viz., Entrez Gene (Locus Link), to rank retrieved documents. The research shows that ranking queries generated from a combination of the Official Gene Symbol, Official Gene Name, Alias Symbols, Entrez Summary, and Protein Products (optional) were very effective in ranking relevant documents higher in the retrieved list. Documents and ranking queries are represented using the traditional vector-space representation, commonly used in information retrieval. Given a gene, the cosine similarity score between the ranking query vector for the gene and each document vector is computed. Cosine scores are in the $[0, 1]$ range and documents assigned a higher score are ranked higher than documents with a lower score. In the absence of summary and protein product information, ranking queries generated from the gene symbol, name and aliases are used to rank retrieved documents. In this experiment study we are working on the Bio Web documents alone.

We use the following criteria to compare the Top K results from Iowa and lgPR, to understand basic characteristics of the two methods. Criteria labeled **R** appear to judge the *relevance* of the paper and those labeled **I** appear to judge importance. Some criteria appear to judge both and are labeled **R,I**.

1. **R**: Does the title or abstract of the article contain the term TNF or its aliases in human? Does the article discuss immune response?

2. **R,I**: Does the article contain any disease related terms? Does the article contain any genomic components (genes, markers, snps, sequences, etc.)?

3. **R,I**: Does the article discuss biological processes related to the Lash terms?

4. **R,I**: What is the connectivity of the article to gene entries in Entrez Gene that are related to TNF? Note that as shown in Table 3, there are 649 Entrez Gene entries that are related to human gene *TNF*. Each PubMed publication was reached by following a result path through the result graph RG that started with one of these Entrez Gene entries. However, some PubMed publications may have been reached along multiple paths in the RG reflecting much greater connectivity.

5. **I**: What is the category of the article (review, survey, etc.). Does the article address some specific topics or is it a broad brush article?

6. **I**: Where did the article appear? What is the journal impact factor? Has the article been highly cited?

| Top 10 | Rel. | Imp. | Criteria | | | | | |
|---|---|---|---|---|---|---|---|---|
| PMID | (0-5) | (0-5) | 1. | 2. | 3. | 4. | 5. | 6. |
| 16271851 | 4 | 2 | H | M | H | L | L | L |
| 1946393 | 4 | 4 | H | L | H | M | H | H |
| 12217957 | 4 | 4 | H | H | H | L | H | H |
| 12545017 | 4 | 4 | H | M | H | L | H | H |
| 9757913 | 3 | 3 | H | L | H | L | H | H |
| 8882412 | 4 | 4 | H | M | H | L | H | H |
| 2674559 | 4 | 3 | H | M | H | L | H | L |
| 7495783 | 4 | 3 | H | H | H | L | H | L |
| 15976383 | 5 | 4 | H | H | H | H | H | L |
| 10698305 | 3 | 3 | H | L | H | L | H | H |

**Table 4: Relevance and Importance of Top 10 Pulications Reported by Iowa Ranking Method**

| Top 10 | Rel. | Imp. | Criteria | | | | | |
|---|---|---|---|---|---|---|---|---|
| PMID | (0-5) | (0-5) | 1. | 2. | 3. | 4. | 5. | 6. |
| 7560085 | 5 | 5 | H | H | H | H | H | H |
| 12938093 | 5 | 5 | H | H | H | H | H | H |
| 10998471 | 3 | 3 | M | H | H | L | H | L |
| 11290834 | 5 | 4 | H | H | H | H | H | L |
| 11501950 | 4 | 3 | H | H | H | L | H | L |
| 11587067 | 5 | 4 | H | H | H | H | H | L |
| 11845411 | 2 | 4 | L | H | H | L | H | H |
| 12133494 | 5 | 4 | H | H | H | H | H | L |
| 12594308 | 4 | 4 | H | H | H | L | H | H |
| 12619925 | 5 | 5 | H | H | H | H | H | H |

**Table 5: Relevance and Importance of Top 10 Pulications Reported by lgPR Ranking Method**

Tables 4 and 5 report Top 10 publications in PubMed that are linked to an Entrez Gene entry that is related to human gene *TNF* and contain the term *linkage disequilibrium*. The first column reports the PubMed identifiers (PMIDs) of the Top 10 publications returned by the Iowa and the lgPR ranking methods. The human evaluation results are reported in the fourth to the ninth columns using the the six criteria listed above. An **H** represents the publication is highly matched to the correspoinding criteria (**M** and **L** represents medium and low respectively). An **H** indicates:

1. The PubMed entry is linked to the human gene *TNF* with Entrez Gene identifier GeneID:7124.

2. The publication contains both diseases related terms and genomic components.

3. The publication contains multiple Lash terms.

4. The connectivity is high, if there are more than five related gene entries linked to the publication.

5. A research article considered more important than a review or a survey, and a more specific topic is better.

6. The article is published in a journal with the impact factor higher than 10.0, or the article is cited by ten or more publications.

We then score the relevance (rel.) and the importance (imp.) in the second and the third columns by combining the number of **H** and **M** reported in the six criteria. Criteria 1 weighs twice compared to the other five criteria. We use a number between 0 and 5, in which 5 indicates the corresponding PubMed entry is highly relevant or highly important to the given query. While both rankings appear to identify "good" documents, Iowa appears to favor relevant documents based on their word content. lgPR appears to exploit the link structure of the RG, and have higher interconnectivity to *TNF* related entries in Entrez gene. The publications retrieved by lgPR are more likely to contain diseases related terms or genomic components. The Iowa ranking has a primary focus on the relevance of documents (based on document contents; it is not able to differentiate the importance of these relevant documents. In contrast, lgPR has a primary focus on importance (based on the link structure of the result graph); it is not able to differentiate the relevance of important documents. We conclude that further study is needed to determine how we can exploit the characteristics of both methods.

There is no intersection between two sets of Top 10 publications returned by these two ranking methods. The first common PMID is 7935762, which is ranked 24 in the Iowa method and 21 by the lgPR method.

## 5. CONCLUSIONS

We have defined a model for life science sources. The answer to a navigational query are the target objects (TOS) of a layered graph Result Graph (RG). We define two ranking metrics layered graph PageRank (lgPR) and layered graph ObjectRank (lgOR). We also report on the results of experiments on real world data from NCBI/NIH. We show that the ranking distribution of lgPR indeed discriminates among the TOS objects of the RG. The lgPR distribution is not the same as applying PageRank a priori to the data graph. We perform a user experiment on complex queries typical of a scientist searching for gene related PubMed publications, and the Top K results of a word based ranking technique (Iowa) that has been shown to be accurate in answering gene queries the query. Using criteria that judge both relevance and importance, we explore the characteristics of these two rankings. Our preliminary evaluation indicates there may be a benefit or a meta-ranking.

We briefly presented layered graph ObjectRank (lgOR) which is an extension to ObjectRank. The challenge of ObjectRank is determining the correct authority weight for each edge. For lgOR, we need to find the weight for the edges that occur in RG. Experiments with users to determine the correct authority weights for lgOR is planned for future work. We expect that IR techniques can be used to determine authority weights.

## 6. REFERENCES

[1] Andrey Balmin, Vagelis Hristidis, and Yannis Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, pages 564–575, 2004.

[2] Magdalini Eirinaki, Michalis Vazirgiannis, and Dimitris Kapogiannis. Web path recommendations based on page ranking and markov models. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 2–9, New York, NY, USA, 2005. ACM Press.

[3] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM Press.

[4] *Homo sapiens* in NCBI Taxonomy Browser. `www.ncbi.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606`.

[5] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating pagerank computations. In *WWW*, pages 261–270, 2003.

[6] Z. Lacroix, L. Raschid, and M.-E. Vidal. Semantic model ot integrate biological resources. In *International Workshop on Semantic Web and Databases (SWDB 2006)*, Atlanta, Georgia, USA, 3-7 April 2006.

[7] Alex Lash, Woei-Jyh Lee, and Louiqa Raschid. A methodology to enhance the semantics of links between PubMed publications and markers in the human genome. In *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2005)*, pages 185–192, Minneapolis, Minnesota, USA, 19-21 October 2005.

[8] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathmatics, 2000.

[9] G. Mihaila, F. Naumann, L. Raschid, and M. Vidal. A data model and query language to explore enhanced links and paths in life sciences data sources. *Proceedings of the Workshop on Web and Databases, WebDB, Maryland, USA*, 2005.

[10] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, New York, NY, USA, 1995.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[12] Matthew Richardson and Pedro Domingos. Combining link and content information in web search. In *Web Dynamics '04: Web Dynamics - Adapting to Change in Content, Size, Topology and Use*, pages 179–194. Springer, 2004.

[13] Aditya Kumar Sehgal and Padmini Srinivasan. Retrieval with gene queries. *BMC Bioinformatics*, 7:220, 2006.

[14] Maria-Esther Vidal, Louiqa Raschid, Natalia Márquez, Marelis Cárdenas, and Yao Wu. Query rewriting in the semantic web. In *InterDB*, 2006.