

# Predicting Performance for Gene Queries

Aditya Kumar Sehgal  
Department of Computer Science  
The University of Iowa  
Iowa City, IA 52242  
sehgal@cs.uiowa.edu

Padmini Srinivasan  
School of Library and Information Science  
The University of Iowa  
Iowa City, IA 52242  
padmini-srinivasan@uiowa.edu

## ABSTRACT

We propose a method to predict the average precision score when ranking document sets for gene queries. Compared to a baseline predictor our method reduces error by 35%. We obtain significant correlations between rankings of queries by predicted and actual average precision scores.

## 1. INTRODUCTION

Our goal is to predict the performance of a post-retrieval document ranking strategy for gene queries. Over the last several decades IR research has contributed various automatic methods for improving retrieval results. Although many methods offer good improvements on average, they sometimes fail for particular queries. For such users the original unmodified queries or result set are the best. Being able to predict such cases up front make for more effective retrieval systems.

In recent research (in preparation for publication) with 4647 human gene queries we observed similar query specific performance patterns. Figure 1 illustrates this. In it two post-retrieval ranking strategies are assessed against a baseline. (The particulars of the strategies are not important for the example). Bars above the X axis indicate improvements compared to the baseline and the height of a bar indicates the extent of improvement. The X axis shows gene queries binned by baseline average precision. Observe that positive returns occur only if baseline performance is 0.6 or less. Thus being able to predict the baseline score will allow us to avoid negatively impacting some queries.

The ability to effectively rank documents about genes is especially crucial for the expanding body of research using literature-based profiles for genes from whole genomes. These are being used to analyze gene clusters in DNA microarray experiments [5, 3, 7]. We hope our research on performance prediction in this bioinformatics context will generate interesting discussions on methods at the SIGIR workshop.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '05, August 15 - August 19, 2005, Salvador, Brazil.  
Copyright 2005 ACM 1-58113-331-6/01/0009 ...\$5.00.

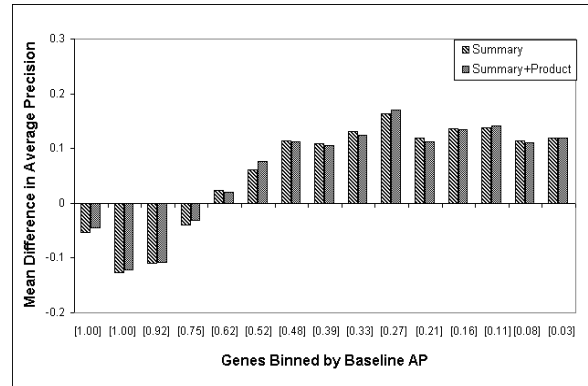


Figure 1: Difference in Average Precision (AP).

## 2. METHODS

Our overall strategy for predicting the average precision (AP) score for gene queries is to use a regression model. We adopt a 5 fold training-testing where we calibrate the model using training data and evaluate on test data.

### 2.1 Dataset

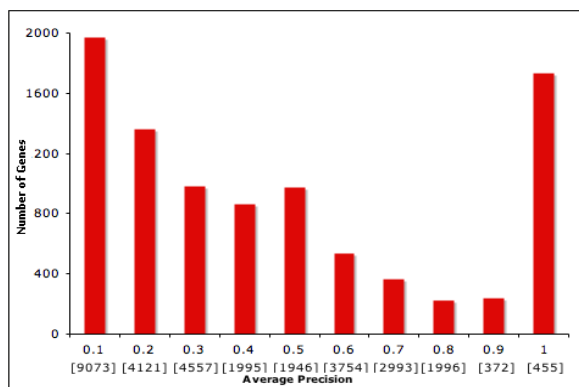
We use a collection of 9240 human genes identified from LocusLink<sup>1</sup>. For each gene LocusLink provides names and symbols and a set of relevant documents that are manually curated. Recent TREC genomics tasks indicate that LocusLink relevant documents are incomplete. However, this does not bias our research in favor or against particular queries.

A typical PubMed boolean search for gene queries is a disjunction of gene names, symbols and aliases. PubMed ranks only by date. Thus we use the same query (without the disjunction operator) for ranking. Ranking is done using the tf\*idf weighting scheme and cosine similarity. Figure 2 provides the distribution of average precision scores (AP) obtained from our ranking. Genes are binned along the X axis by AP score, with the average retrieved set size shown. Our goal is to predict these AP scores.

We randomly divide the 9240 genes into five equal parts. Each split is defined by a training set formed from four parts with the fifth taken as the test set.

### 2.2 Features

<sup>1</sup>www.ncbi.nlm.nih.gov/LocusLink



**Figure 2: Distribution of AP Scores.** (The peak on the 1.0 bin involves many queries with very few known relevant documents that are all ranked at the top positions).

Features fall into two groups. The first includes counts from the retrieved document set, normalized in two different ways. The second explicitly targets gene term ambiguity. Let  $D$  be the set of documents retrieved for a query,  $Me$  the set of MeSH (Medical Subject Heading) terms in  $D$ ,  $RN$  be the set of RN (Registry Number) terms, which are chemical terms assigned by the Chemical Abstracts Service (CAS), in  $D$ , and  $Ti$  be the number of words in titles of documents in  $D$ . Features in group 1 are:

- (F1)  $|D|$  - Number of documents.
- (F2)  $|Me_{uniq}|/|D|$  - Average number of unique MeSH terms per document.
- (F3)  $|RN_{uniq}|/|D|$  - Average number of unique RN terms per document.
- (F4)  $|Ti_{uniq}|/|D|$  - Average number of unique title words per document.
- (F5)  $|(Me + RN)_{uniq}|/|D|$  - Average number of unique MeSH and RN terms per document.
- (F6)  $|Me_{uniq}|/|Me|$  - Fraction of the total number of MeSH terms that are unique.
- (F7)  $|RN_{uniq}|/|RN|$  - Fraction of the total number of RN terms that are unique.
- (F8)  $|Ti_{uniq}|/|Ti|$  - Fraction of the total number of title words that are unique.
- (F9)  $|(Me + RN)_{uniq}|/|Me + RN|$  - Fraction of the total number of MeSH and RN terms that are unique.

As  $|D|$  increases we expect lower AP. Other features estimate whether the retrieved set is homogenous (about the same gene topic) or not. Queries retrieving homogenous document sets are likely to yield high AP scores. Such documents will have similar MeSH and RN terms and title words. Therefore features 2-9 count the number of unique MeSH terms, RN terms and title words. Features 2-4 normalize by document set size while features 6-8 normalize by the total

number of MeSH terms, RN terms and title words respectively. Features 5 and 9 treat the MeSH and RN terms as a single metadata unit. We expect features 2-9 to correlate negatively with score. Features in the second group include:

- (F10) Other biological meanings ( $NumBio$ )
- (F11) General English meanings ( $NumEng$ ).
- (F12) References to other genes ( $NumGNs$ ).

**NumBio:** Gene symbols may have other biological meanings, eg. ACR also means *albumin/creatinine ratios* and *acute to chronic ratio*. We use the Schwartz and Hearst [8] algorithm to recognize short form (A) - long form (B) pairs appearing as  $A(B)$  in text. For a search term  $t$ , we identify its long forms in the retrieved documents (set  $E_t$ ). If  $|E_t| \leq 1$ ,  $t$ 's count is 0 otherwise,  $|E_t| - 1$ . A single long form for  $t$  could be the correct one, hence the subtraction. We add counts for search terms in the gene query to get  $NumBio$ . Higher  $NumBio$  values will likely yield lower scores.

**NumEng:** Search terms may have general English language meanings (eg. GAB and RAGE). For a gene query, we identify the number of search terms with English meanings ( $NumEng$ ) by a lookup of WordNet<sup>2</sup>. Larger  $NumEng$  values will likely yield lower AP scores.

**NumGNs:** Gene terms sometimes refer to more than one gene. For each term we identify the number of LocusLink genes for which it occurs as a name or an alias. Adding these for a gene, we get  $NumGNs$ . Larger values will likely yield lower AP scores.

## 3. RESULTS AND ANALYSIS

### 3.1 Correlations

Our objective is to model performance, namely average precision, in terms of some or all of the twelve independent variables described before. But first we examine the correlations which are shown in Table 2. Since the values of features 1 through 9 are skewed we applied a log transformation ( $\ln(1+x)$ ) before doing a Pearson's correlation run.

Group 1 features are strongly correlated with score (Table 1). Group 2 features are weakly correlated and we do not consider them further. Within group 1 we observe a single strong anti-correlation with feature 1:  $|D|$ . Interestingly, for features 2 - 9, although our initial intuition regarding the strength of the correlation is supported, unexpectedly these are all negative. Our hindsight explanation is as follows.

Consider feature 7 ( $|RN_{uniq}|/|RN|$ ). It correlates negatively (around -0.37) with  $|RN_{uniq}|$  while  $|RN_{uniq}|$  correlates negatively (around -0.2) with score (not reported in the table). Thus if the number of unique RN terms in the retrieved set drops (rises) then the number of total RN terms drops (rises) even further. This is illustrated in figure 3 for a gene retrieving 53, 052 documents. We see what happens with the first 1000 documents, first 2000 documents, ..., the full set of documents (in the order returned by PubMed). These two negative correlations contribute to the positive correlation between ( $|RN_{uniq}|/|RN|$ ) and score. The same observations hold for features 2 through 9.

<sup>2</sup><http://www.cogsci.princeton.edu/~wn/>

| Correlations                     |      |      |      |      |      |      |         |      |      |          |            |           |             |
|----------------------------------|------|------|------|------|------|------|---------|------|------|----------|------------|-----------|-------------|
|                                  | (F1) | (F2) | (F3) | (F4) | (F5) | (F6) | (F7)    | (F8) | (F9) | (F10)    | (F11)      | (F12)     | (S)         |
| Group 1                          |      |      |      |      |      |      |         |      |      |          |            |           |             |
| (F1) $ D $                       | 1.0  | -0.9 | -0.7 | -0.7 | -0.9 | -0.9 | -0.8    | -0.9 | -0.9 | -0.2     | 0.3:0.4    | 0.4       | -0.51       |
| (F2) $ Me_{unq} / D $            |      | 1.0  | 0.8  | 0.9  | 1.0  | 0.9  | 0.9     | 0.8  | 0.9  | 0.1      | -0.3:-0.2  | -0.3      | 0.55        |
| (F3) $ RN_{unq} / D $            |      |      | 1.0  | 0.7  | 0.9  | 0.7  | 0.7     | 0.7  | 0.7  | 0.04:0.1 | -0.3:-0.2  | -0.3      | 0.45-0.46   |
| (F4) $ Ti_{unq} / D $            |      |      |      | 1.0  | 0.9  | 0.8  | 0.7:0.8 | 0.7  | 0.8  | 0.1:0.2  | -0.2       | -0.3      | 0.49-0.5    |
| (F5) $ Me_{unq} + RN_{unq} / D $ |      |      |      |      | 1.0  | 0.9  | 0.9     | 0.8  | 0.9  | 0.1      | -0.3:-0.2  | -0.3      | 0.54-0.55   |
| (F6) $ Me_{unq} / Me $           |      |      |      |      |      | 1.0  | 0.9     | 0.9  | 1.0  | 0.2      | -0.3:-0.2  | -0.3      | 0.58-0.59   |
| (F7) $ RN_{unq} / RN $           |      |      |      |      |      |      | 1.0     | 0.9  | 1.0  | 0.2      | -0.2       | -0.3      | 0.59-0.6    |
| (F8) $ Ti_{unq} / Ti $           |      |      |      |      |      |      |         | 1.0  | 0.9  | 0.2      | -0.3:-0.2  | -0.3      | 0.53-0.54   |
| (F9) $/( Me  +  RN )$            |      |      |      |      |      |      |         |      | 1.0  | 0.2      | -0.2       | -0.3      | 0.58:0.59   |
| Group 2                          |      |      |      |      |      |      |         |      |      |          |            |           |             |
| (F10) $NumBio$                   |      |      |      |      |      |      |         |      |      | 1.0      | -0.04:0.02 | 0.02:0.03 | 0.08:0.11   |
| (F11) $NumEng$                   |      |      |      |      |      |      |         |      |      |          | 1.0        | 0.2       | -0.15:-0.14 |
| (F12) $NumGNs$                   |      |      |      |      |      |      |         |      |      |          |            | 1.0       | -0.2:-0.18  |
| (S) AP Score                     |      |      |      |      |      |      |         |      |      |          |            |           | 1.0         |

Table 1: Pearson Correlation Matrix. Each cell indicates the range of values obtained for the 5 training - testing splits. Numbers are rounded to 1 significant digit, except for the last column - rounded to 2 significant digits. Cells with single numbers indicate that all 5 splits yield the same coefficient. Differences are usually from the third decimal place.

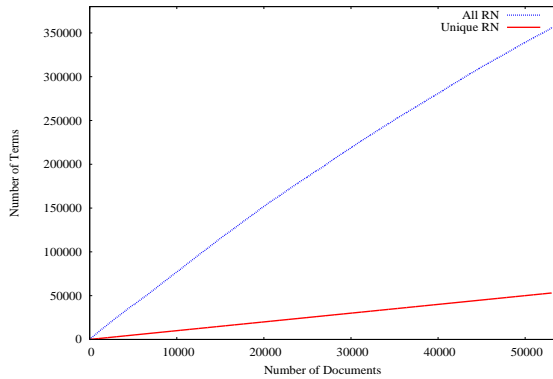


Figure 3: Example Gene.

Our independent variables are strongly mutually correlated. All are positive except those involving  $|D|$ . Thus any single independent variable is sufficient for prediction. We select feature 7:  $|RN_{unq}|/|RN|$ , which is the number of unique RN terms divided by the total number of RN terms.

### 3.2 Regression Models

We calibrate two models using ordinary least square regression. Model 1 includes all 9 independent variables while model 2 includes only our chosen feature, i.e., feature 7. The adjusted R-square for model 1 ranges from 0.3484 to 0.3623 across the five training sets. All but 2 p-values (for one training set) are  $< 0.01$ . For model 2 the adjusted R-square ranges from 0.3483 to 0.3582. All p-values are  $< 0.01$ . Thus both models are equally powerful and are likely to be useful.

### 3.3 Test Set results

We use the two calibrated regression models to predict scores for each test set query. We evaluate performance by the mean squared error between the actual and predicted scores. As a baseline we take the mean score on the training set as the prediction for each test query. Overall, the error for the baseline is 0.1178 while it is 0.0751 and 0.0762 for model 1 and 2 respectively. Thus model 1 and model 2 reduce error by 36% and 35% respectively compared to baseline, and are therefore comparable in this respect.

Kendall-Tau correlation tests between queries ranked by

actual scores and by predicted scores produce correlation coefficients from 0.43 and 0.44 for both models across the five splits (significant at the 0.01 level). Again model 2 is comparable with model 1.

## 4. RELATED RESEARCH

Several researchers have addressed the problem of predicting query performance. Saracevic and Kantor [4] explore a variety of query characteristics such as clarity and specificity within a study of users and search methods. We show, for example that InfoSpiders, a type of web crawler, is sensitive to topic popularity [6]. More recently a query clarity score based on the relative entropy between a query language model and a collection language model was proposed [2]. Yom-Tov et al. explore prediction using a histogram based algorithm and using a modified decision tree [9]. Amati et al. explore information theoretic measures to predict hard queries[1]. Their goal was to avoid applying query expansion on the worst (most difficult) topics. They obtained results similar to results with no expansion for the worst topics and better than expansion on all topics.

## 5. CONCLUSIONS

Our method successfully predicts the AP score for ranked document sets obtained for gene queries. The method starts with correlation analysis to select features followed by calibration of regression models. We explored a variety of features. Most of them aim at estimating the homogeneity of the retrieved document set. Features that target specific kinds of ambiguities did not exhibit interesting correlations with the average precision score and hence were not useful. We found that a simple linear regression model built from a single feature: the number of unique RN terms divided by the total number of RN terms in the retrieved set, is effective at predicting the score. When compared to a baseline strategy this method reduces error by 35%. Finally the ranking of queries by predicted scores also correlates significantly (at the 0.01 level) with a ranking by actual scores. We believe that our approach and results will contribute to discussions at the SIGIR workshop.

## 6. REFERENCES

- [1] Amati G et al. Query difficulty, robustness and

- selective application of query expansion. Proc. 25th European Conference on IR,127-137,2004.
- [2] Cronen-Townsend S et al. Predicting query performance. SIGIR 299-306,2002.
  - [3] Raychaudhuri S. and Altman RB. A Literature-based method for assessing the functional coherence of a gene group, *Bioinformatics*, 19,396-401,2003.
  - [4] Saracevic T. and Kantor P. A study of information seeking and retrieving II. Users, questions, and effectiveness. *JASIS* 39(3),177-196, 1998.
  - [5] Shatkay H et al. Genes, Themes and Microarrays, *ISMB*, 317-328, 2000.
  - [6] Srinivasan, P et al. A General Evaluation Framework for Topical Crawlers. *Information Retrieval*, 8,417-447,2005.
  - [7] Srinivasan, P and Libbus, B. Mining MEDLINE for Implicit Links between Dietary Substances and Diseases. *Bioinformatics*, 20 Suppl 1:I290-I296,2004.
  - [8] Schwartz, AS, Hearst, MA. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *PSB*,451-462, 2003.
  - [9] Yom-Tov E et al. Improving document retrieval according to prediction of query difficulty. *TREC* 2004.