

# AuGEAS (AUthoritativeness Grading, Estimation, and Sorting)

Ayman Farahat  
Palo Alto Research Center  
3333 Coyote Hill Road  
Palo Alto, CA 94304  
farahat@parc.com

Geoff Nunberg  
Palo Alto Research Center  
3333 Coyote Hill Road  
Palo Alto, CA 94304  
nunberg@parc.com

Francine Chen  
Palo Alto Research Center  
3333 Coyote Hill Road  
Palo Alto, CA 94304  
fchen@parc.com

## ABSTRACT

When searching for content in a large heterogeneous document collections like the World Wide Web it is not easy to know which documents provide reliable authoritative information about a subject. The problem is particularly pointed as it concerns content search for “high-value” informational needs such as retrieving medical information, where the cost of error may be high. In this paper, a method is described for estimating the authoritativeness of a document based on *textual*, non-topical cues. This method is complementary to estimates of authoritativeness based on link structure, such as the PageRank and HITS algorithms. This method is particularly suited to “high-value” content search where the user is interested in searching for information about a specific topic. A method for combining textual estimates of authoritativeness with link analysis is also presented. The types of textual cues to authoritativeness that are easily computed and utilized by our method are described, as well as the method used to select a subset of cues to increase the computation speed. Methods for applying authoritativeness estimates to re-ranking documents returned from search engines, combining textual authoritativeness with social authority, and use in query expansion are also presented. By combining textual authority with link analysis, a more complete and robust estimate can be made of a document’s authoritativeness.

## 1. INTRODUCTION

A notoriously difficult problem in using large heterogeneous document collections like the World Wide Web is that it is not easy to know which documents provide reliable authoritative information about a subject. The problem is particularly pointed as it concerns content search for “high-value” informational needs such as retrieving medical information, where the cost of error may be high.

Authoritativeness is commonly measured based on social networks represented by the link structure of the web (*e.g.*,

Google’s PageRank algorithm [5], Kleinberg’s HITS algorithm [14]). But these measures of authoritativeness do not consider the content of the documents, even though that is often a highly useful indicator of document authoritativeness, and is not derivable from link structure alone. As an example of social authority, when a newspaper says, “An authoritative source announced that the President would veto the bill,” we interpret “authoritative” to mean that the source was relatively close to the people who have social authority over the matter in question (in this case, presumably, to the President or his advisors).

In contrast, in the content based, or textual notion of authoritativeness, when we say, “Professor Jones has written an authoritative book on Roosevelt’s foreign policy,” we do not necessarily imply that Jones had any close relation to the people who had first-hand knowledge of the subject at hand, or for that matter that scholars or journalists are generally disposed to cite Jones’ book (though that may very well be the case). Rather, we mean that the book is authoritative on *internal grounds*: it reads as if it is well-researched, uses language in a skillful and appropriate way, contains numerous references of the right sort, and so forth. In this vein, our prediction of textual authoritativeness is a generalization of the readability index [8] that encompasses hyper-linked documents.

In society at large, as evidenced on the Web, there is much more heterogeneity in knowledge and viewpoint, and the mere fact that a text is widely referenced may not by itself assure that it is authoritative in the broader sense of the term. This point becomes particularly important when we come to issues where there is a large amount of misinformation abroad – such as in obtaining medical information. For example, when we gave the query “heterosexual transmission AIDS virus” to Google, the first 50 hits contained a number of pages that most people would judge as authoritative, but they also include some pages that the majority of health professionals would be unlikely to recommend: *e.g.*, a page about how federal AIDS policy is shaped by the “homosexual agenda,” and another that accuses the government of rewarding promiscuity by providing AIDS patients with housing assistance and other benefits. These pages came up well before other general-information pages from the HIV Insite project at UCSF or the Harvard AIDS Institute. In view of the fact that 85% of the users only view the first page of the search results [19], it becomes increasingly important to reliably identify and present search results.

By the same token, it often happens that a text that is

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.

CIKM’02, November 4–9, 2002, McLean, Virginia, USA.

Copyright 2002 ACM 1-58113-492-4/02/0011 ...\$5.00.

authoritative on internal grounds occurs in a site that is not widely linked to – for example, a government health institute report that someone has included on a Geocities site.

In the case of Web applications, we identify two types of searches “content searches” and “site searches”. In content search, the user is interested to learn more about a particular topic and could potentially visit a number of sites returned by a search engine to learn more about that topic. In a site search, the user is interested to find a specific site, that he or she is interested in and will scan through the results of the search engine until they find that site. For content searches, our ability to identify authoritative Web pages will be improved if we can perform an additional analysis of their textual content. In this paper, we will describe methods for performing such analyses and suggest some applications for them.

The work presented in this paper can be viewed as a generalization of the query expansion approach presented by Glover *et al.* [11]. In their work, Glover *et al.* presented an approach for learning query modification when locating pages within a discrete set of preselected categories. Our approach generalizes that work by extending the set of preselected categories to a continuum and by computing a single authority rank that can be used to further refine the results of the search.

Our work has applications in improving the ranking of retrieved documents and in improving query expansion to return more authoritative documents. A *query dependent* framework for improving information retrieval through the integration of the textual content and the link structures of documents is presented in [3] [6] and [18]. In contrast, our approach for integrating textual and linkage information relies on the *textual* authority of a document which is *query independent* and can be computed and stored once for each document.

In the following sections, we describe our method for estimating textual authority, applications of textual authority estimation, and present results comparing textual and linkage based authority, and results in applications of textual authority. In section 2, we give a general overview describing our classes for characterizing a document as to textual authority. In section 3 we define the features that characterize textual authority and build on our definition of ordered textual authority to develop predictive models for estimating textual authority. In section 4 we describe methods for applying textual authoritativeness prediction to information retrieval. In section 5, we compare the observed performance of textual and linkage based authority. We also apply textual authority based information retrieval applications to a number of queries and compare the results to general search engines. Finally in section 6, we give our conclusions, and outline directions for future work.

## 2. OVERVIEW OF AUTHORITY RANKING

A large heterogeneous collection of documents such as the web can be classified in a number of ways. In our research, we developed a set of classes that cover a large portion of the documents available on the Internet and that are particularly relevant to high-value informational domains like medical and scientific information. These classes range from the most authoritative documents, that is, documents written by someone with a scientific background for an audience with a scientific background, to documents written by a ran-

**Table 1: Classification Attributes**

Number	Attribute	Values
1	Review	Reviewed (R), Not reviewed (N)
2	Author’s background	Professional(P), General(G)
3	Audience	Professional (P), General (G)
4	Author’s affiliation	Professional (P), Media (M), Commercial (C),None (N)

**Table 2: Assigning Document Authority**

Attribute				Class
1	2	3	4	
R	P	P	P	1
R	P	G	P	2
R	P	G	N	3
R	P,G	G	M	4
R	P,G	G	C	5
R,N	P,G	G	P,N ,M	6
N	G	G	N	7

dom person for anyone willing to read their web page.

In order to provide a more consistent unambiguous and non-arbitrary framework for classifying the documents, we examine the following attributes of each document; the author’s background, the targeted audience, the author’s institutional affiliation, and whether the document has been reviewed or examined by others. Table 1 lists these attributes as well as their possible values. Table 2 describes our framework for assigning an authority class to a specific document.

The set of classes that we developed are listed below and a short description or example is given; many of the examples are from the medical domain and are used for the purpose of illustration:

- 1. Scientific documents:** documents by a professional to other professionals, *e.g.*, scientific research papers, articles from Center for Disease Control, New England Journal of Medicine
- 2. General information–scientific:** documents provided by scientific organizations for the general public, *e.g.*, press releases from CDC, UCSF
- 3. Information from reputable sites:** *e.g.*, health sites such as drkoop.com
- 4. General information–news:** documents provided by news organizations for the general public *e.g.*, Time, CNN
- 5. Commercial pages:** *e.g.*, drugstore.com
- 6. Mail groups and discussion lists:** this also includes opinion and editorials from a newspaper
- 7. Home pages:** *e.g.*, personal home pages and organization home pages such as the Green Party
- 8. Links pages:** may contain a short paragraph describing each link. Although these pages may point to authoritative documents, they do not in themselves contain any authoritative information.

There is an implicit, if sometimes inexact, ordering of the authoritativeness of these classes. All things being equal, people regard scientific documents as more authoritative than press reports, and press reports as more authoritative than information found on newsgroups. The ordering relation allows us to rank and compare the authority of different documents. To that end, we mapped the authority of each class to an ordered set of the positive integers. In general, any monotonic map from the set of classes to the set of real numbers can be used to assign an authority rank. In particular, we defined the map from the class of documents to positive integers, corresponding to that shown in the above list.

### 3. ESTIMATING TEXTUAL AUTHORITY

In this section we describe our approach to estimating the authority rank of a general web page given our definition of textual authority as given in Section 2 and a set of web pages that have been classified accordingly. In particular, we first identify the most salient features of the web page, encode these features in feature vector  $\mathbf{x}$ , and then develop a predictive model that maps the feature vector  $\mathbf{x}$  to an authority rank  $a_{text}$ .

#### 3.1 Feature Selection and Features

In order to fully capture and accurately represent documents that are typically encountered in web searches, we selected a large number of features that capture both the linguistic and presentation (*e.g.*, colors, tables) content. The features used to predict authoritativeness are of various types:

- use of particular characters in the plain text (*e.g.*, question marks, semicolons)
- numerals or particular styles of numerals (*e.g.*, “34.56,” “1957:31”)
- particular words (*e.g.*, “I,” “Mr.,” “Dr.,” “today”)
- word-classes (*e.g.*, words with learned prefixes like “pseudo-” or “hetero-” or learned suffixes like -acious, -metric, -icity)
- words in certain grammatical locations (*e.g.*, sentences beginning with adverbs)
- HTML features (hyperlinks, tables, images, page color)
- abbreviations and classes of abbreviations (*e.g.*, “pp.,” “c.c.”)
- text characteristics (*e.g.*, document length, average and standard deviation of word, sentence, and paragraph lengths)
- part of speech tagging features (*e.g.*, number of noun phrases, verb phrases)
- different readability indices (*e.g.*, Gunning Fog or Flech [8]).

These features provide a much richer and web-embodying view than those used by traditional readability measures.

Textual authoritativeness appears to correlate with a relatively large number of features. If all the features are used in

evaluating the authoritativeness of a document, the features that are less informative as to authoritativeness add noise to the decision. Furthermore, these features decrease the speed with which decisions can be made. Therefore, from the large number of features, a subset was selected for use in the ranking of documents with respect to authoritativeness levels. This was done using a training set of documents that was manually labeled with respect to authority using the map defined in Section 2. The variables were selected using stepwise regression with the “Efroymson” method of the S software package [2]. Efroymson’s method is an iterative method that adds a new variable to the selected set at each iteration, and then considers whether any of the variables currently in the subset should be dropped based on partial correlations between the new and selected set of features. In particular, in the first step, the best feature for predicting the textual authority of the classes is identified and selected. In each following step, the combination of the best feature and each of the other features is evaluated one at a time, and the best feature for predicting authoritativeness from the set of test features is selected. Partial correlations between the feature selected in the current step and the features that have already been selected is used to determine if any of the variables currently in the subset of selected features should be dropped.

#### 3.2 Prediction Models

The task of predicting the textual authority of a document represented by a feature vector  $\mathbf{x}$  can be viewed as a cost sensitive multi-class classification problem [20]. Because of the relative ranked relationship between classes, the cost of misclassification is not the same between each pair of classes, *e.g.*, the cost of misclassifying a home page as a scientific document is much higher than the cost of cost of misclassifying a general information document by a scientific organization as a scientific document

While a number of approaches such as ordinal regression [13] [9] and multi-class classification [21] have been suggested for handling these problems. We experimented with a number of learning algorithms and found that the two best performing algorithms in terms of training complexity and in terms of generalization and prediction were metric-regression algorithms and boosted decision trees.

In view of the above discussion, we developed a set of linear regression models which we shall refer to as “Augeas” and boosted decision trees which we shall refer to as “Boost”.

##### 3.2.1 Augeas

The linear regression model was developed using the reduced feature set as identified in Section 3.1 and the manually labeled training set. The authority of each document in a test set separate from the training set was then estimated using the regression model.

An important factor that impacts the performance of the regression models is the *transformation* of the dependent variable which in our case is the textual authority. While the textual authority scheme described in Table 2, gives an implicit ordering of authoritativeness, the exact numerical values might not produce the best separation between the classes. In order to allow for a richer class of dependent

variables, we consider transformations of the form

$$\begin{aligned} y(\lambda) &= \frac{y^\lambda - 1}{\lambda} \\ y &= \log(y); \lambda = 0 \\ 0 < y &\leq 1 \end{aligned} \quad (1)$$

This transformation is the well known Box-Cox transformation [4]. We experimented with different values of the parameter  $\lambda$ , and used the average precision and recall performance as described in section 5 to select the optimal values.

### 3.2.2 Boost

The boosted decision tree directly classifies the document into one of the  $N$  possible classes described in Table 2. For the boosted decision tree, we used the AdaBoost algorithm [10] with 10 stages of “C4.5” decision tree [17] tree as a base learner. AdaBoost manipulates the training examples to generate multiple hypothesis. AdaBoost maintains a probability distribution  $p_i(x)$  over the training examples. At each iteration  $l$ , Adaboost draws a training set by sampling with replacement according to  $p_l(x)$ . A decision tree learning algorithm is constructed and then its error rate  $\epsilon_l$  weighted according to  $p_l(x)$  is computed. The error rate of the decision tree is used to compute  $p_{l+1}(x)$ , which essentially gives a higher weight to the examples that have been misclassified by the base learner. The procedure is repeated until a number of base learners have been selected or the error rate  $\epsilon_l$  exceeds 0.5.

## 4. APPLICATIONS OF TEXTUAL AUTHORITY

In section 3.2, we described a general framework for computing the textual authority of a document. The computed textual authority can be used to improve the quality of the search either by directly using it to re-rank the search results based on the estimated textual authority, or by combining it with social authority. It should be noted that since the textual authority of the document is query independent, the authority can be precomputed once for the document. Applications that leverage the textual authority can then use the precomputed authority. This is different from the social authority measure as computed by HITS, which depends on the particular query and has to be re-computed for each query.

### 4.1 Re-ranking Search Results

In a large, heterogeneous and constantly evolving collection such as the world wide web, the results returned by a search engine in response to a specific query often includes a wide range of documents that encompass all ranges of authoritativeness. While this might be a desirable feature in some situations, users are more likely to be interested in a specific class of documents *e.g.*, scientific documents. An immediate application of the textual authority is to reorder and filter the search results according to the textual authority, and then return all the documents that fall within a certain authority range, *e.g.*, scientific documents.

### 4.2 Combining Social and Textual Authority

The social authority of a page in a networked structure reflects how other members in that structure view that page.

In that sense, the more members in the community that point to a specific page, the higher the authority of that page. However, not all pages that refer to other pages are equally selective in terms of the pages that they point at. The original HITS algorithm [14], defines the notion of “hub.” A hub is a specific page that points to a high-authority pages. Conversely, a high-authority page is pointed at by high-quality hubs. Following [14], we associate a set of hyper-linked pages  $V$  with a directed graph  $G = (V, E)$  with the nodes corresponding to the pages and a directed edge  $(p, q) \in E$  indicates the presence of an edge from  $p$  to  $q$ . The graph structure can be conveniently represented by the adjacency matrix  $A$  with entry  $a[i][j] = 1$  if there is a link from node  $i$  to node  $j$  and is set to 0 otherwise. In view of the above discussion, we define the authority rank  $auth(p)$  and the hub rank  $hub(p)$  of page  $p$  as follows.

$$auth(p) = \sum_{q:(q,p) \in E} hub(q) \quad (3)$$

$$hub(p) = \sum_{q:(q,p) \in E} auth(q) \quad (4)$$

Kleinberg proved that the authority weights correspond to the entries of the principal eigenvector of the matrix  $A^T A$  and that the hub weights correspond to the entries of the principal eigenvector of the matrix  $AA^T$ . The page rank algorithm used by the Google search engine replaces the adjacency matrix  $A$  with the matrix  $M = \alpha U + (1 - \alpha)B$  where  $U$  is the transition matrix of uniform transition probability and represents a random transition to any page, and  $B$  is a normalized version of  $A$  such that all rows sum to 1. The parameter  $\alpha$  is in the range 0.1-0.2 and represents the probability that a user will “jump” to a random page.

The adjacency matrix in its current form assigns equal weights to all the links. At the conceptual level, one might weight certain links more heavily than other links based on factors other than link structure. Bharat and Henzinger [3] use query expansion to expand the initial query into a broader query term  $q$  and then measure the similarity  $d(i, j)$  between every document  $j$  in the collection of documents and the expanded query. The entries of the adjacency matrix that correspond to links from document  $j$  are weighted by the similarity  $d(j, q)$  between the document and the expanded query.

The textual authority of a page gives an estimate of the *intrinsic* quality of the page and is therefore an indicator of the quality of the pages linked to by that document. In view of the above discussion, we replace the entries of the adjacency matrix corresponding to page  $j$  by the textual authority of page  $j$ . In particular we set  $a[i][j]$  to the textual authority of page  $i$  if page  $i$  points to page  $j$  and zero otherwise.

$$a[i][j] = \begin{cases} \frac{auth_{text}(i)}{|i|} & \text{if } i \text{ points to } j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where  $|i|$  is the out degree of page  $i$ .

The weighted authority ranks are the entries of the principal eigenvectors of the matrix  $A^T A$ , or the matrix  $M^T M$  in the case of page rank. Note that adding a few keywords or linking to good hubs wouldn’t significantly change the textual authority and that only an authoritative rewrite of the page would change the textual authority. In that spirit, the textual authority produces a more robust weighting that

**Table 3: Contingency Table for LLR Test**

	term $t$	other terms
relevant	$T(t, R)$	$T(\neg t, R)$
non-relevant	$T(t, N)$	$T(\neg t, N)$

can not be easily “spoofed”.

### 4.3 Query Expansion

In a large number of situations, the intended topic of a query is broader than the specific query terms. Thus matching the query against the documents is usually not sufficient. Instead of directly using the query term, the query is first expanded into a broader query topic. The query expansion operation has two phases. First, a search engine is used to get an initial set of relevant documents. Second, the most frequent terms in the initial set are used to define a candidate set of query terms. The query expansion terms are extracted from the candidate set using statistical tests such as log likelihood ratio [1], [7]. The selection of the initial set can have a significant effect on the precision as reported by Glover *et al.* [11], who restrict the initial set to a specific category *e.g.* home pages. The concatenation of the new and old query terms forms a new query that is given to a search engine. The search results for the new query provide a richer and more relevant set of documents than the original query.

In order to insure that highly authoritative documents are returned, it is important to insure that the expanded query contains terms that correlate with textually authoritative documents. To that end, the candidate query terms are extracted from textually authoritative documents whose textual authority exceeds a certain threshold and not from all the documents in the initial set. The log likelihood ratio test (LLR) [7] is then used to test whether the distribution of each of the candidate terms in the relevant documents is significantly different from its distribution in a general collection of documents. If the distribution of specific term is significantly different, then the term is included in the set of query expansion terms. The LLR test can be conveniently formulated as contingency Table 3, where  $T(t, R)$  is the number of times term “ $t$ ” occurs in relevant documents,  $T(t, N)$  is the number of times term “ $t$ ” occurs in non-relevant documents,  $T(\neg t, R)$  is the number of times term other than “ $t$ ” occurs in relevant documents, and  $T(\neg t, N)$  is the number of times term other than “ $t$ ” occurs in non-relevant documents. The counts for the  $T(t, N)$ ,  $T(\neg t, N)$  were computed from a general corpora [15].

## 5. RESULTS

To test our approach, we choose an authority scheme and a number of search topics that are representative of the application at hand, *i.e.*, topics that have generated a much interest from a wide spectrum, and where there is a high cost of misinformation. The selected topics were from the medical and freedom of expression domains.

For each of these topics, we issued a query to the search engine “Google”, retrieved the top 100 results of the search and then reordered the results based on the textual authority. The results for the query “alcohol addiction” sorted by textual authority are given in Table 4 while the results sorted by Google are given in Table 5. The top documents in Table

4 tend to be longer length pages that have more scientific and medical content. For example, the top document is from the Brain science group at Brown University that describes in some detail the program that offers training for alcohol addiction professionals. The third-ranked document in Table 4, which was ranked number 59 by Google, is a highly authoritative document on addiction by Stanton Peele, PhD, JD Fellow, The Lindesmith Center - Drug Policy Foundation. The sixth entry in Table 5 center.butler.brown.edu, is the home page of the Butler center at Brown university and while having scientific authority by virtue of its affiliation with the center, the document does not provide information on alcohol addiction.

The results for the query “Internet filtering” sorted by textual authority are shown in Table 6. These results tend to contain more content pages that have useful information on the issue as compared to the Google ranked pages shown in Table 7 that have a high social authority.

Another important evaluation measure is precision and recall [1]. In order to quantify these measures, we first submitted queries to the search engine Google, and then used our textual authority method to automatically rank the top 100 documents returned by Google in response to each query. Independently, we manually ranked the documents based on the quality of their content using the authority ranking scheme outlined in section 2. The manual ranking provided us with a set of relevant documents to be used in evaluating the precision and recall. We selected the query terms based on two criterion. First we wanted “high value” terms where there is a large number of documents and the cost of miss-classification is high. We also wanted to perform content search where the user is interested to learn more about the topic. Content searches are encountered frequently in practice as evidenced by search engine logs [12].

We compared the manual ranking to the textual authority rank as computed by linear regression model (Augeas) and the boosted decision tree (Boost) on one hand, and by Google on the other hand. Results of the precision at 11 standard recall levels are shown in Figures 1, 2, and 3, where the relevant set was the top 10, 20 and 30 documents, respectively. In each case, we averaged the results for the queries “genetic modification of food”, “child abuse”, and “Ebonics”. The results indicate that while the first few highest-ranked socially authoritative sites returned by Google are of high quality, the remaining sites are of mixed authoritativeness. The Augeas and Boost systems on the other hand, return informative sites that would have not been highly ranked by Google and as such can be viewed as complementary to the Google search engine. For these queries, the Augeas and Boost systems consistently outperforms the Google system. At higher recall levels, the difference in precision is much more pronounced *e.g.*, for the top 30 documents at 10% recall, a user would need to examine “3” documents by Google compared to “4” documents by Augeas. On the other hand, at 100% recall, a user would need to examine “75” documents by Google compared to “60” documents by Augeas or Boost. These results are further supported by the average precision as shown in Table 8.

Another important measure of the quality of each of the three rankings is to measure how well it correlates with the human rank *i.e.*, the *rank correlation*. The two most commonly employed measures are the “Spearman Rank-Order

**Table 4: AuGEAS Results for Alcohol Addiction**

URL	Authority	AuGEAS Rank	Google Rank
www.brainscience.brown.edu/news/postdocsalcohol.html	2.36	0	63
english.pravda.ru/fun/2001/08/23/13143.html	2.52	1	32
www.peele.net/lib/gambling.html	2.66	2	59
etoh.niaaa.nih.gov/	2.89	3	16
www.addiction-ssa.org	2.91	4	39
www.macad.org	2.93	5	44
www.bma-wellness.com/Addiction/EtOHPsychobiology.htm	2.94	6	48
www.bfe.org/alco.htm	2.96	7	84
ssw.unc.edu/fcrp/Cspn/vol4_no4/gender_and_alcohol.htm	3.03	8	58
chemcases.com/alcohol/alc-11.htm	3.07	9	61
immuners.org/00/12step/recovery/treatment.htm	3.10	10	22

**Table 5: Google Results for Alcohol Addiction**

URL	Authority	Google Rank	AuGEAS Rank
www.health.org	4.64	0	6
www.health.org/recoverymonth	5.51	1	78
www.niaaa.nih.gov	3.34	2	22
www.well.com/user/woa	3.66	4	37
center.butler.brown.edu	4.52	5	58
www.healthrecovery.com	7.15	6	91
www.utexas.edu/research/asrec	5.53	7	79
www.utexas.edu/cons/wcaar/	4.74	8	65
www.thirteen.org/wnetschool/origlessons/alcohol/alcoholov.htm	5.16	9	75
www.thirteen.org/wnetschool/origlessons/alcohol/alcoholproc.html	4.16	10	49

**Table 6: AuGEAS Results for Internet Filtering**

URL	Authority	AuGEAS Rank	Google Rank
www.frc.org/papers/insight/index.cfm?	0.02	0	67
www.surfcontrol.com/news/press_releases/content/07_30_2001.html	0.953	1	22
www.surfcontrol.com/	1.99	2	21
abcnews.go.com/sections/scitech/DailyNews/library_filters001220.html	2.06	3	62
www.software4parents.com/index-xdetect.html	2.18	4	47
www.alia.org.au/advocacy/filtering.html	2.46	5	76
www.ftrf.org/internetfilteringmemo.html	2.48	6	9
www.ala.org/alaorg/oif/filt_stm.html	2.57	7	4
www.newsbytes.com/news/01/166171.html	2.60	8	34
www.newsbytes.com/news/01/166332.html	2.63	9	33

**Table 7: Google Results for Internet Filtering**

URL	Authority	AuGEAS Rank	Google Rank
www.n2h2.com	2.766	0	11
www.n2h2.com/products/index.php	2.849	1	13
www.we-blocker.com	3.18	2	22
www.bluehighways.com/tifap	3.517	3	27
www.ala.org/alaorg/oif/filt_stm.html	2.567	4	7
www.ala.org/alaorg/oif/filtersandfiltering.html	3.95	5	47
www.tispa.org/info/kinnaman/filtering.htm	2.914	6	15
www.cyberpatrol.com	6.28	7	93
www.surfwatch.com	3.54	8	28
www.ftrf.org/internetfilteringmemo.html	2.48	9	6
www.safesurf.com	4.4	10	62

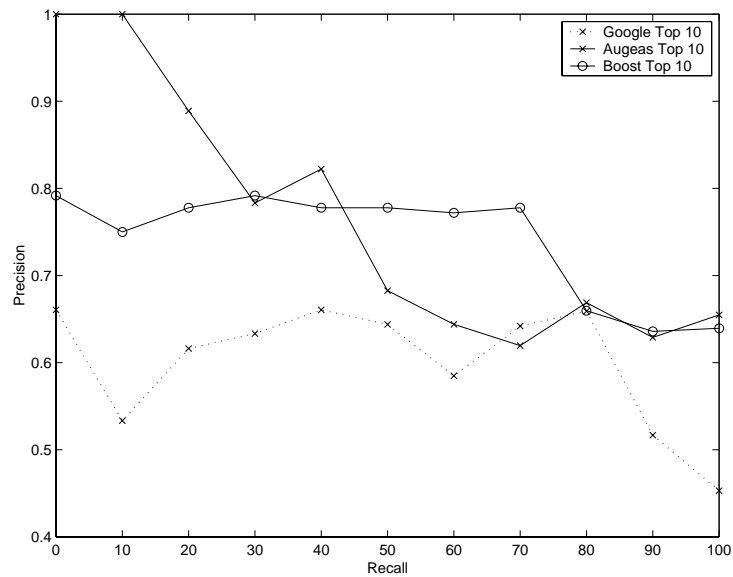


Figure 1: Top 10 documents precision at 11 standard recall levels.

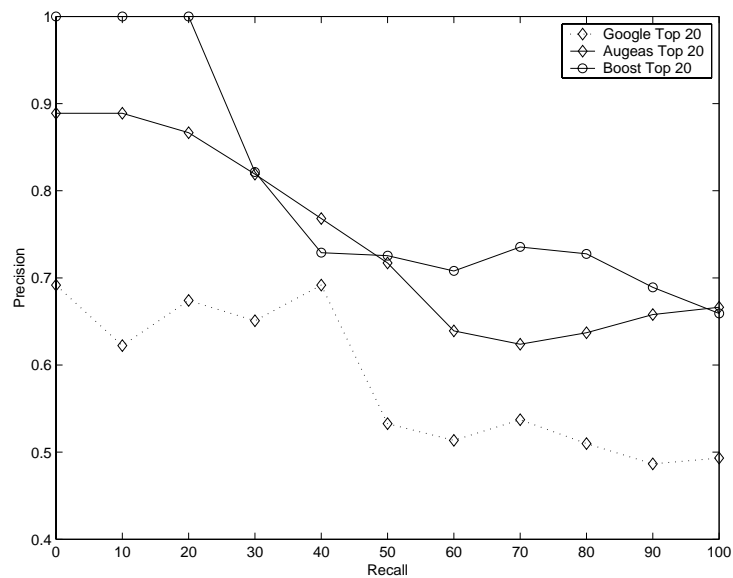


Figure 2: Top 20 documents precision at 11 standard recall levels.

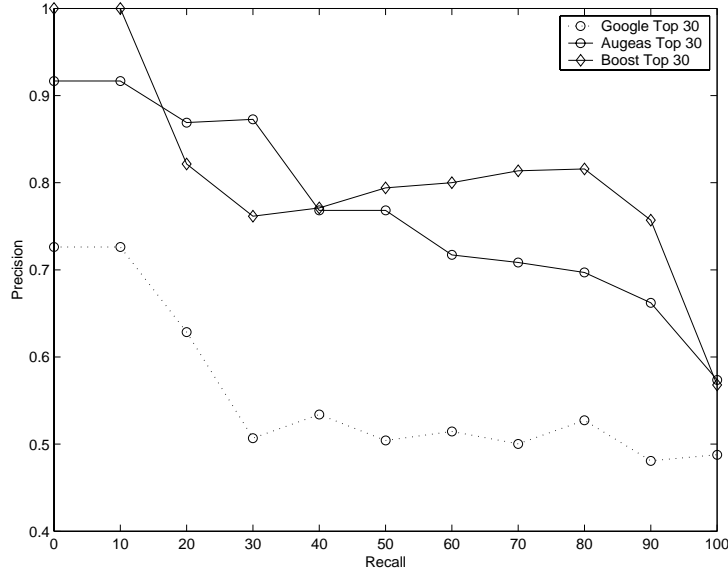


Figure 3: Top 30 documents precision at 11 standard recall levels.

Table 8: Average Precision

Relevant Set	Avg. Boost	Avg. Augeas	Avg. Google
Top 10	0.74	0.74	0.59
Top 20	0.78	0.72	0.57
Top 30	0.79	0.76	0.54

Table 9: Rank Correlation

Relevant Set	Google	Augeas	Boost
Spearman	0.2468	0.454	0.627
Kendall	0.173	0.322	0.476

coefficient” and the “Kendall Tau coefficient” [16]. These measures give an estimate of the distance between two ordered lists. In our case, we are interested in estimating the distance between the manually ranked list, and each of the automatically ranked lists. In particular let  $\gamma_{human}^{query_i}$ ,  $\gamma_{google}^{query_i}$ ,  $\gamma_{augeas}^{query_i}$ ,  $\gamma_{boost}^{query_i}$  denote the human, Google, Augeas and Boosted list of ordered search results for  $query_i$ . Furthermore we define the set of search results for a set of queries

$$\Gamma_{human} = [\gamma_{human}^{query_1} \dots \gamma_{human}^{query_n}] \quad (6)$$

$$\Gamma_{augeas} = [\gamma_{augeas}^{query_1} \dots \gamma_{augeas}^{query_n}] \quad (7)$$

$$\Gamma_{boost} = [\gamma_{boost}^{query_1} \dots \gamma_{boost}^{query_n}]. \quad (8)$$

Let  $\tau(\Gamma_{human}, \Gamma_{google})$ ,  $\rho(\Gamma_{human}, \Gamma_{google})$  denote the Kendall and Spearman rank correlation between the human and Google lists respectively.  $\tau(\Gamma_{human}, \Gamma_{augeas})$ ,  $\rho(\Gamma_{human}, \Gamma_{augeas})$ ,  $\tau(\Gamma_{human}, \Gamma_{boost})$ , and  $\rho(\Gamma_{human}, \Gamma_{boost})$  are similarly defined. Table 9 gives the rank correlation between the manual list and the three automatically generated lists. These results show that both the boosted decision tree model “Boost” and the linear regression model “Augeas” outperform the Google results.

The authority based query expansion procedure outlined in section 4.3 was used to extract query expansion terms for

a number of queries. Only documents with authority rank “3” or better were used to expand the query in the case of authority based query expansion while all documents were used in the baseline query expansion. The LLR test was then used to rank the terms and the top 30 terms were retained.

Table 10 shows the difference in the terms that were identified by the two approaches. The Augeas based query expansion for the query “Alcohol Addiction” was able to identify medical terms such as *dependence*, *disease*, *disorder*, *symptom*, and *patient*. The terms that were uniquely identified by the baseline approach tend to be of a very general nature.

The Augeas query expansion terms for the query “Internet filtering” contains terms such as *amendment*, *speech*, *court*, *decision*, and *liability*. These terms are indicative of the freedom of speech issues associated with Internet filtering and the legal implications of these issues. The Augeas query expansion terms for cancer cure contain the terms *case*, *diet*, *victim* and *vitamin*. Upon closer examination of the top Augeas sites for the query “cancer cure”, we found that a large number of these sites discussed the fact that while there are existing treatments for cancer, there is no medically proven cure. These sites go on to discuss and discredit claims of existing cures (in particular using vitamins).

We are currently running experiments to evaluate combining social and textual authority as outlined in section 4.2.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the notion of textual authority and its applications to information retrieval. Textual authority can complement and augment social authority and is particularly useful in high-value information retrieval. We presented three applications of textual authority to information retrieval. We tested our applications on a number of queries. The results indicate that the textual authority can significantly improve the quality of the results by emphasizing authoritative documents.



**Table 10: Query Expansion Terms for Alcohol Addiction**

Query	Terms in AuGEAS	Terms in Google
Alcohol Addiction	dependence, disease,disorder,withdrawal symptom,patient,mercy,lord	image,information,link,site ,age book,family,function,health
Internet Filtering	amendment,court,decision,employee environment,false,harmful,liability,speech	computer,content,include,list rating,school,system,user,win
Cancer Cure	case,diet,give,time,treat victim,vitamin	book,document,drug,make home,medicine,news

In future work, we plan to use textual authority to characterize the authoritativeness of a particular site. This will enable us to develop a *prior* estimate of the authoritativeness of a particular document. This estimate can then be formally combined with the measured estimate to give a more robust estimate of authority.

## 7. ADDITIONAL AUTHORS

Charles Mathis, Palo Alto Research Center,  
email: [mathis@parc.com](mailto:mathis@parc.com).

## 8. REFERENCES

- [1] Baeza-Yates and Riberio-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1999.
- [2] R. A. Becker, J. M. Chambers, and A. R. Wilks. *Splus Reference Manual*. Statistical Sciences, Inc., Seattle, Washington, 1991.
- [3] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Research and Development in Information Retrieval*, pages 104–111, 1998.
- [4] G. E. P. Box and D. R. Cox. An analysis of transformations. In *J. R. Statist. Soc.*, volume B 26, pages 211–252., 1964.
- [5] S. Brin and L. Page. The anatomy of a large-scale hypertextual (web) search engine. In *The seventh international world wide web conference*, 1998.
- [6] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.
- [7] T. E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- [8] R. Flesch. A new readability yardstick. *Journal of applied Psychology*, 32:221–233, 1948.
- [9] E. Frank and M. Hall. A simple approach to ordinal classification. In *Proceedings of the European Conference on Machine Learning*. Springer-Verlag, 2001.
- [10] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.
- [11] E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles. Web search – your way. *Communications of the ACM*, 1999. accepted for publication.
- [12] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths. Measuring search engine quality. *Information Retrieval*, 4(1):33–59, 2001.
- [13] R. Herbrich, T. Graepel, and K. Obermayer. Regression models for ordinal data: A machine learning approach, 1999.
- [14] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [15] G. Leech. 100 million words of english: the british national corpus, 1992.
- [16] E. Lehmann. Nonparametrics: Statistical methods based on ranks, 1975.
- [17] J. Quinlan. Programs for machine learning, 1993.
- [18] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Neural Information Processing Systems 14*, 2002.
- [19] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log, 1998.
- [20] P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.
- [21] J. Weston and C. Watkins. Multi-class support vector machines, 1998.