Cultural Orientation: Classifying Subjective Documents by Cociation Analysis

Miles Efron
School of Information
Sanchez Building (SZB) 564
Austin, TX 78712-0390
(512) 471-3821
miles@ischool.utexas.edu

Abstract

This paper introduces a simple method for estimating *cultural orientation*, the affiliation of hypertext documents in a polarized field of discourse. Using a probabilistic model based on cocitation information, two experiments are reported. The first experiment tests the model's ability to discriminate between left- and right-wing documents about politics. In this context the model is tested on three sets of data, 695 partisan web documents, 162 political weblogs, and 72 non-partisan documents. Accuracy above 90% is obtained from the cocitation model, outperforming lexically based classifiers at statistically significant levels. In the second experiment, the proposed method is used to classify the home pages of musical artists with respect to their mainstream or "alternative" appeal. For musical artists the model is tested on a set of 227 artist home pages, achieving 88 % accuracy.

Introduction

This paper presents a simple method for estimating the *cultural orientation* of documents in a hypertextual system. Cultural orientation is the degree of association between an object and a community in a polarized field of discourse. As an example, this paper presents a method for estimating the orientation of Web documents about politics–the extent to which these documents participate in left- or right-wing ideologies. Additionally, the method is used to discriminate between musical artists' home pages with respect to their mainstream and alternative appeal.

Underpinning both examples is a simple probabilistic model. Throughout this paper it is assumed that in a hypertextual system such as the Web, cultural orientation is evident in the hyperlink structure of a given community. In particular, this paper presents a model based on cocitation analysis. Starting with a small number of documents on a given topic, each of whose orientation is known beforehand, cultural orientation is estimated by analyzing the likelihood that a document of interest is co-cited with exemplars from each distinct community.

Estimating cultural orientation is of interest for developers of personalized information systems. In particular this work was initiated during the development of a recommender system for weblogs (blogs). Often opinionated and partisan, blogs lend new challenge to old problems in information retrieval and filtering. In particular, retrieval from web logs depends on a suitable model of author sentiment. Blogs about politics, for instance, my be of interest to widely divergent audiences. According to the website blogstreet.com, *Cut on the Bias*[1] and *Liberal Oasis*[2] are both popular blogs about politics. Yet their perspectives on the same political subjects are at odds. The home page of *Liberal Oasis* reads, "Where the Left is Right and the Right is Wrong." In *Cut on the Bias*, on the other hand, a recent entry decried the "lameness and pandering" of the 2004 US Democratic presidential candidates. Readers of *Liberal Oasis* are unlikely to be interested in keeping abreast of *Cut on the Bias*. Although both sites are about politics, their orientation with respect to their topic is at odds.

In addition to politics, many blogs are about popular music. Again, however, the type of music treated within such blogs forms a number of distinct communities. Thus people interested in the industrial improvisations of Merzbow and The Swans are unlikely to pay much heed to blogs about boy bands or Brittney Spears.

The thesis of this paper is that detecting such cultural distinctions is feasible by adopting a social metaphor. The model explicated here is based on previous work by Turney and Littman (Turney & Littman 2002), which follows Firth's notion that we may learn the meaning of a word "by the company it keeps" (Firth 1957). Likewise, we may use hyperlink structure to learn about a document on the Internet (Gibson, Kleinberg, & Raghavan 1998; Brin & Page 1998; Agrawal *et al.* 2003). This paper takes Firth's metaphor literally. The proposed estimator gauges an object's cultural orientation by noting which documents it tends to be cocited with. Beginning with a core of exemplar documents, the estimator uses AltaVista's[3] `link:` command to create an odds ratio that approximates the target entity's cultural orientation.

The remainder of the paper is organized as follows. The following section contextualizes this study within the literature on personalized information filtering and opinion mining. Within these contexts, the following section offers a

[1] http://bias.blogfodder.net
[2] http://www.liberaloasis.com
[3] http://www.altavista.com

definition of cultural orientation, along with a method for estimating the orientation of Web documents. The bulk of the paper reports results from two experiments that apply the proposed classifier to documents about politics and about popular music. The final section reflects on the outcome of these experiments, proposing directions for future research.

Previous Work

The proposed method for determining the cultural orientation of documents has ties to at least two fields of research: personalization through collaborative filtering and sentiment classification (or opinion mining). This section contextualizes the experiments reported here with respect to these literatures.

Social Models for Information Filtering

Using a social metaphor for information filtering is not a new idea. Certainly recommender systems based on collaborative filtering have done this for years. Collaborative filtering involves predicting which objects will be of interest to a user by comparing that user's expressed interests with the interests of other users of the system (Konstan 2004). In the case of document management, the well-known GroupLens project has used collaborative filtering to predict Usenet articles that will be interesting to particular users (Konstan *et al.* 1997).

However, long before the advent of collaborative filtering, researchers in the field of bibliometrics developed models for organizing documents by analyzing patterns of citation among scholarly articles (Borgman 1990; White & McCain 1989). Particularly relevant to the current study is the notion of cocitation among documents (Small 1973). A forerunner of contemporary link analysis, cocitation study aims to quantify the relationship between two documents d_i and d_j by noting which documents they both cite. The underlying principle in cocitation analysis is that if two documents both cite many of the same documents we can infer that they enjoy some similarity even if they do not directly cite each other.

PMI-IR for Opinion Mining

The current work is based on the PMI-IR algorithm proposed by Turney and Littman (Turney & Littman 2002). Their work defines an unsupervised approach to learning the semantic orientation of words (Wiebe 2000; Hatzivassiloglou & McKeown 1997). For a given word w_i, the semantic orientation $SO(w)_i)$ is modeled as a real-valued number. Positive values indicate that the word has positive connotations, while negative values imply bad connotations. To gauge the semantic orientation of a word w_i, Turney and Littman estimate the pointwise mutual information between w_i and two sets of words **negative** and **positive**. These sets are comprised of seven words that exemplify negative and positive sentiment, respectively. As defined in (Church & Hanks 1989), the pointwise mutual information between two words w_i and w_j is given by Equation 1:

$$PMI(w_i, w_j) = \log_2(\frac{p(w_i, w_j)}{p(w_i)p(w_j)}) \quad (1)$$

where $p(wi, wj)$ is the probability that the words co-occur in a document. The pointwise mutual information between a word and a set of words **S** is obtained by treating the members of **S** as an equivalence class when calculating the probabilities. Based on Equation 1, Turney and Littman calculate $SO(w_i)$ (the semantic orientation of w_i based on PMI) by Equation 2:

$$SO(w_i) = PMI(w_i, \mathbf{positive}) - PMI(w_i, \mathbf{negative}) \quad (2)$$

where **positive** and **negative** are the sets of exemplar words. To estimate the necessary probabilities, Turney and Littman use the AltaVista search engine to extract an estimated semantic orientation:

$$\hat{S}(w_i) = \log_2(\frac{|w_i \text{ NR } p\_query| \times |n\_query|}{|w_i \text{ NR } n\_query| \times |p\_query|}) \quad (3)$$

where $|w_i$ NR p_query $|$ is the number of pages returned for the query w_iNEAR(pos$_1$ ORpos$_2$ ... OR ... pos$_{|p|}$). Words with a positive $\hat{S}$ are classified as having positive semantic orientation, and vice versa.

Because the counts in Equation 3 are derived from a search engine, Turney and Littman call the method semantic orientation from pointwise mutual information from information retrieval (SO-PMI-IR). Using this simple algorithm, Turney and Littman report results on a par with the more involved approach described by Hatzivassiloglou and McKeown, with classification accuracy on a set of 3596 words above 80%.

More immediately relevant to this work, in (Turney 2002) Turney applies the SO-PMI-IR algorithm to the problem of opinion mining. Using a document's average $\hat{S}$ over syntactically tagged two-word phrases as evidence of overall sentiment, Turney reports 84% accuracy for author sentiment in automobile reviews, with slightly lower numbers for other types of products. Thus the PMI-IR apporach appears to operate at state of the art levels for opinion mining (cf. (Dave, Lawrence, & Pennock 2003; Pang, Lee, & Vaithyanathan 2002)).

The method proposed in this paper adapts the PMI-IR algorithm from a focus on terms to a focus on documents. This shift allows us to generalize from estimation of semantic orientation to the notion of cultural orientation described below. Additionally, shifting our attention from co-occurrence of terms to cocitation among documents allows us a desirable stability in estimation. As noted by Agrawal *et al.*, a hyperlink-based approach to opinion mining is both accurate and robust (Agrawal *et al.* 2003)

The Idea of Cultural Orientation

Estimating semantic orientation by PMI allows us to approach the linguistics of opinion socially, judging words "by the company they keep." We can extend the utility of this framework by applying its social metaphor at the document, rather than the term level. This section generalizes Turney and Littman's social metaphor to estimate *cultural orientation:* the degree to which a given document participates in an identifiable and polarized community of discourse.

In the context of political writing or popular music, semantic orientation is of less interest than cultural orientation as such. Readers of progressive weblogs, for instance are unlikely to identify with conservative news media. Likewise, fans of mainstream music are unlikely to appreciate avant garde jazz. But identifying the cultural orientation of websites about politics and music is problematic due to the unreliability of authors' self-identification. For instance, the Institute for Historical Review describes itself as "non-ideological, non-political, and non-sectarian."[4] However the chief subject matter of this site is a so-called revisionist view of the Holocaust. In a much less controversial vein, many musicians with a highly popular fan-base describe themselves as "alternative rock" (artists such as Pearl Jam and Linkin Park come to mind). It is ill -advised to take at face value the statements of authors in polarized fields of discourse.

Estimating Cultural Orientation

In the context of opinion mining, Agrawal *et al.* argue that "the vocabulary used in the two sides of an issue is generally identical," a fact that obviously frustrates lexical machine learning (Agrawal *et al.* 2003). Thus Agrawal *et al.* turn to the hyperlink structure between documents in their analysis. Their graph partitioning-based approach to classifying positive or negative sentiment yields results superior to most term-based methods, achieving accuracy near 84% for certain topics.

Let **L** and **R** be two sets of URLs. Also let d_i be a document. By adapting Equation 1 we define the pointwise mutual information between citations to d_i and to another document d_j by Equation 4:

$$PMI(d_i, d_j) = \log_2(\frac{p(d_i, d_j)}{p(d_i)p(d_j)}) \quad (4)$$

where $p(d_i)$ is the probability that a document in the community of interest contains a link to d_i and $p(d_i,\ d_j)$ is the probability that documents d_i and d_j are co-cited—that is to say, the probability that a document in the population contains links to both documents. If the probability of a document d_k linking to d_i is independent of whether or not it also links to d_j then we gain no information by knowing the links of d_k; Equation 4 yields 0 bits of information. But if documents d_i and d_j tend to be cited together in the discursive community Equation 4 quantifies the information we gain by knowing the links of d_k. Equation 4 thus estimates the degree of independence with respect to cocitation between documents.

If the URLs that comprise **L** and **R** are exemplars of resources committed to opposing points of view with respect to a given topic *T*, then we estimate the orientation of d_i with respect to *T* by Equation 5:

$$\hat{\omega}(d_i, T) = \log_2(\frac{|\text{link:}d_i \text{ AND link:}\mathbf{R}| \times |\text{link:}\mathbf{L}|}{|\text{link:}d_i \text{ AND link:}\mathbf{L}| \times |\text{link:}\mathbf{R}|}) \quad (5)$$

where *link:* **R** is the AltaVista command *link:* r_1 *OR link:* r_2 ... *OR link:* $r_{|R|}$[5]. To avoid division by zero, we follow Turney in adding 0.01 to the first element of the numerator and denominator of Equation 5 as a form of Laplace smoothing.

Probabilistic Motivation

Let d_i be a document whose cultural orientation we wish to estimate, as in the previous section. We begin by assuming that two Bernoulli processes generate all documents on the topic of interest. Each process B_k generates documents that contain a hyperlink to documents in the exemplar set K. That is, each process generates documents with a given orientation towards the topic. Finally each distribution B_k produces documents that contain a hyperlink to d_i with probability p_k. Using the sign of Equation 5 as a classifier corresponds with choosing the most likely distribution given the data. Equation 5 may be re-written as Equation 6:

$$\hat{\omega}(d_i, T) = \log_2(\frac{|\text{link:}d_i \text{ AND link:}\mathbf{R}| \ / \ |\text{link:}\mathbf{R}|}{|\text{link:}d_i \text{ AND link: }\mathbf{L}| \ / \ |\text{link:}\mathbf{L}|}). \quad (6)$$

By definition, the numerator of the antilog of Equation 6 is the maximum likelihood estimate of p_R, the probability of success in the Bernoulli process that generates documents with links to items in the exemplar set **R**. That is, out of $|link : \mathbf{R}|$ draws from B_R we observe $|link : d_i \textit{ AND } link : \mathbf{R}|$ "successes." Likewise, the denominator gives the MLE of p_L, the estimated probability of seeing d_i co-cited with an exemplar in set **L**.

Because the number of documents citing the exemplars in **R** and **L** is large, Equation 6 approximates a log-odds ratio; it is symmetric around 0. If the probability of seeing a cocitation of d_i and an exemplar in **R** is greater than the probability of a cocitation with a document in **L** then Equation 6 is positive. Conversely, a higher likelihood of cocitation with **L** gives a negative value for Equation 6.

Classifying a document with respect to the sign of Equation 6 thus corresponds to choosing the class whose underlying distribution is most likely to have generated the document. We thus define the classifier for document d_i with respect to topic *T* by Equation 7:

$$\hat{o}(d_i, T) = sign(\hat{\omega}(d_i, T)) \quad (7)$$

Of course our estimator $\hat{\omega}$ could be improved. As an odds ratio, one could derive a confidence interval for ω using the asymptotic standard error estimate (Agresti 2002). Finally, a standard hypothesis test for the difference between two binomial proportions could be employed. However, in this experiment, neither the confidence interval nor hypothesis testing approach improved classification significantly. Thus we retain the point estimate of Equation 6.

[4] http://www.ihr.org/main/about.shtml

[5] Due to instability in AltaVista's counts on conjunctions of *link*: commands, the numbers reported here were obtained by issuing a separate query for each member of the exemplar sets. Work is currently under way to construct a local database of link data to obviate this.

Left	Right
www.commondreams.org	www.nationalreview.com
www.thenation.com	www.townhall.com
www.zmag.org	www.worldnetdaily.com
www.prospect.com	www.family.org
www.buzzflash.com	www.insightmag.com
www.counterpunch.org	www.frontpagemag.com
www.cursor.org	www.rightwingnews.com
www.tompaine.com	www.sierratimes.com
www.motherjones.com	www.enterstageright.com
www.moveon.org	www.cc.org
www.progressive.org	www.lucianne.com
www.utne.com	www.infowars.com
www.inthesetimes.com	www.spectator.org
www.liberaloasis.com	www.etherzone.com
www.pacifica.org	www.aclj.org
www.movingideas.org	www.anxietycenter.com
www.wage-slave.org	www.rutherford.org
www.turnleft.com	www.conservative.org
www.progressivemajority.org	www.dailyrepublican.com

Table 1: Exemplar Partisan Documents

Experimental Evaluation

To examine the accuracy of the classifier proposed in equation 7 experiments were undertaken in two domains: political discourse and popular music. The goal of these experiments was to compare the performance of the cocitation-based classifier with respect to established methods of documents classification.

Orientation of Political Documents

Experiment 1 analyzed the cultural orientation of documents about politics, determining the degree to which these documents participated in traditionally left- or right-wing sentiment. While the left/right polarization is an oversimplification of the world of political discourse, it carries strong currency in contemporary debate. Especially in the arena of blog retrieval, it was hypothesized that developing a means to filter documents with a particular partisan bias would be useful for personalization.

To evaluate the suitability of the proposed estimator in the political arena, two sub-experiments were performed. The first tests the approach on a corpus of 695 partisan web documents. In experiment two, analysis was limited to a set of 162 political weblogs.

To estimate the cultural orientation of a document d_i with respect to the topic *politics*, we begin by defining two sets of exemplar documents **left** and **right**. For this study, each set of exemplars was populated with 19 highly partisan documents. A list of these documents appears in Table 1.

The documents were chosen by analysis of the Open Directory (http://www.dmoz.org). Left-leaning documents were chosen by selecting those documents in the Open Directory's category *Politics:Liberalism:Social Liberalism* that had (according to AltaVista) at least 1000 incoming hyperlinks, on December 7, 2003. Likewise, the set of right-wing exemplars was chosen from *Politics:conservatism*.

Left	Right
Yahoo!: US Politics : Political Opinion : Liberal	Yahoo!: US Political Opinion : Conservative
Yahoo!: US Politics : Political Opinion : Progressive	Yahoo!: US Political Opinion : Conservative : Magazines
Yahoo!: US Political Opinion : Progressive: Magazines	Yahoo!: US Political Opinion : Conservative : Organizations
Yahoo!: US Political Opinion: Progressive: Personal Opinion	Yahoo!: US Political Opinion : Conservative : Personal Opinion
DMOZ: Politics : Liberalism :Social Liberalism	DMOZ: Politics: Conservatism
DMOZ: Politics : Liberalism : Social Liberalism : Progressive	DMOZ: Politics: Conservatism : Organizations
DMOZ: Politics : Liberalism : Social Liberalism : News & Media	DMOZ: Politics: Conservatism : Personal Pages

Table 2: Categories for Test Set Construction

Web Data Evaluation A test corpus was constructed based on the web directories provided by Yahoo! And the Open Directory Project (DMOZ). For this experiment all URLs listed in the directories shown in Table 2 were used as test cases. Removing duplicates and any documents that appeared in the **left** and **right** training sets left a corpus of 695 documents, each of whose political orientation was ostensibly given by its Yahoo! or DMOZ category.

Each of the test documents was classified using three methods:

- Cocitations as described in Equation 7
- Naive Bayes: A naive Bayes classifier fitted after removal of stopwords (Mitchell 1997)
- SVM: A support vector machine classifier trained on the same data as the naive Bayes model (Burges 1998). The SVM used a linear kernel function.

The word-based classifiers were trained on a corpus of 2412 documents[6]. The training set was built by downloading the first two levels of hypertext from the websites that comprised the exemplars for the cocitation-based estimator. After initially dismal word-based performance 10 additional websites of each class (chosen from the relevant DMOZ directories) were downloaded in attempts to mitigate naive Bayes' and SVM's tendency to overfit the training data.

A summary of these results appears in Table 3. More detailed results appear in Table 4, the confusion matrix for the cocitation-derived classifications. An obvious problem with the cocitation based approach to classification is that some documents may have insufficient in-links to estimate an accurate odds ratio. Thus Table 4 contains a column labeled "none," which gives the number of documents that contain

[6]For the lexically based classifiers *N=662* test documents because some of the tested URLs were unavailable for download despite being listed in an online directory.

Method	Accuracy	F-Measure
Cocitation	94.1	0.97
Naive Bayes	64.71	0.79
SVM	72.96	0.84

Table 3: Summary of Experiment One

	Predicted		
Actual	**Left**	**Right**	**None**
Left	268	6	30
Right	13	271	107

Table 4: Cocitation Classifier Confusion Matrix

fewer than four in-links for both paradigms (left and right). Table 4 shows that 30 left-wing and 107 right-wing web documents had sparse cocitation data. The figures in Table 3 assume that all 137 of these are automatically classified into the most probable class (based on our training set for lexical classifiers, which was "right").

From Tables 3 through 5, an advantage for the cocitation-based classifier is evident. If we adopt the rationale of Table 3 (counting sparsely linked pages as "right"), the p-value for the resultant contingency tables is $p < 0.001$ for both left- and right-wing documents. If we omit sparsely linked documents from our analysis, then the cocitation-based classifier achieves 97% accuracy for left-wing documents and 95% accuracy for right-wing documents; the *p*-value remains effectively unchanged.

This result suggests that the link structure among politically oriented documents provides a less noisy source of evidence for opinion detection than do simple lexical features. Interestingly, for these data, using bigrams for term-based classification degraded performance over the unigram model, suggesting that the term-based models are suffering an overfitting effect, which the link-based approach avoids. However, the matter of sparse link information remains problematic for the cocitation-based classifier. This is analogous to the cold-start problem in collaborative filtering systems, where new items within a system are unlikely to be recommended due to poorly parameterized preference information (Huang, Chen, & Zeng 2004). However, two facts can mitigate concern over sparse link data. First is the possibility that in future work the cocitation-based classifier could easily be supplemented with lexical evidence. Recent work by Beineke *et al.* (Beineke, Hastie, & Vaithyanathan 2004) suggests that this is both practical and theoretically sound. As the next section shows, the second consideration is that in domains such as weblogs, sparse cocitation information is far less problematic than it appeared to be in the test set described here.

Blog Data Evaluation As a second test, the cocitation-based classifier was evaluated on a set of political blogs. Finding blogs about politics is easy. Unfortunately for our purposes, most blog directories do not subdivide the *Politics* category by partisanship. Because of this the test data for this evaluation was sparse, and a bit skewed. A sample

	Predicted	
Actual	**Left**	**Right**
Left	202	90
Right	89	281

Table 5: SVM Confusion Matrix

of 119 left-wing blogs was downloaded from *Progressive Gold*[7], a self-described progressive blog. The maintainer of *Progressive Gold* keeps a directory of decidedly left-leaning blogs; these comprised the sample on the left.

On the right, test data came from the conservative portal *Townhall.com*[8], which maintains a directory of right-wing blogs. Unfortunately, however, *Townhall's* directory contained only 43 blogs at the time of this writing, and thus our sample is skewed to the left.

Nonetheless, the performance of the cocitation-based classifier was impressive for these blogs. The method correctly classified all 43 right-wing blogs, and incorrectly classified only two of the 119 left-slanted blogs. In contrast, the SVM model mis-classified 9 right-wing documents and 55 left-wing documents. A chi-squared test on the resultant contingency tables of each method yielded $p < 0.001$ for the left-wing documents and $p = 0.003$ for the right-wing documents.

We may understand this result by comparing the hyperlink profile of the blog data versus the web-page data described above. The weblogs yielded richer link information on average than the standard web pages. While 3 percent of the sampled web pages had fewer than four in-links, only 1 of the 162 (0.6%) weblogs had sparse link information. And although the maximum number of in-links was higher for the plain web pages, the middle quartiles for the blogs were much higher.

The hyperlink is the coin of the realm in weblog communities. From a discursive standpoint, blogs operate much like annotated bibliographies, with authors using other authors' work as foils for their own reflections, citing their sources as they write. Although the small size of this sample leaves the matter open for future investigation, this study suggests that the community of political webloggers is more highly connected than the general domain of political websites is. This highly social behavior makes the weblog domain especially ripe for application of the cocitation-based approach to estimating political orientation.

Orientation of Musical Artist Web Pages

To examine the problem of classifying documents by cultural orientation more fully, a round of experiments was performed using websites about popular music. Popular music was chosen with an eye towards filtering in the domain of weblogs, where many authors write about bands, artists, and musical styles. Like politics, the discursive domain of popular music online is highly factionalized; the number of subgenres subsumed under the rubric of "rock music" is huge.

[7] http://progressivegoldbeta.blogspot.com

[8] http://www.townhall.com

Alternative	Mainstream
www.thewire.co.uk	www.rollingstone.com
www.insound.com	www.mtv.com
www.breakbeat.co.uk	www.vh1.com
www.forcedexposure.com	www.melodicrock.com
www.disinfo.com	www.imusic.com
www.popmatters.com	www.virgin.net
www.furious.com/perfect	www.billboard.com
www.trouserpress.com	www.andpop.com
www.pitchforkmedia.com	www.musictoday.com
www.fakejazz.com	top40-charts.com
www.dustedmagazine.com	www.muchmusic.com
www.stylusmagazine.com	
www.insound.com	
www.pseudo.com	
www.sandboxautomatic.com	
www.musicmoz.org	
www.xlr8r.com	
www.othermusic.com	

Table 6: Exemplar Music Sites

As an approximation of music's fractured discursive domain, this section describes an experiment into discrimination between the home pages of musical artists with "mainstream" versus "alternative" appeal. Both of these notions are admittedly subjective, but as a guiding principle this study took the operational notion that mainstream rock is likely to be played on commercial radio stations, while alternative music is likely to be played on college stations, if it sees airtime at all. Thus our concern is not with whether a band designates itself as alternative, but rather, whether listening audiences do.

To operationalize this definition, three DJs from a college radio station were asked to list what they considered to be authoritative websites on the matter of "alternative" music and on "mainstream" music. The resultant recommendations were used to populate the exemplar sets for the cocitation-based classifier, as shown in Table 6. Note that there are more alternative than mainstream exemplars. While this does reflect the tastes of the surveyed DJs, the total number total incoming links to these two sets are similar (131,761 for alternative and 194,528 for mainstream sites)[9].

As in the previous example, a word-based classifier was trained on the documents acquired by downloading the first two levels of hypertext from the exemplar sites. This led to a training set of 17,813 alternative pages and 12,020 mainstream pages. Due to the large size of the training set and limited computing resources, an SVM classifier was not feasible for the music data. Thus only a niave Bayes model was built.

Constructing a test collection was somewhat difficult, due to the subjectivity inherent in the alternative/mainstream dichotomy. Many bands who label themselves alternative would have no appeal for fans of the websites shown in Table 6. Conversely, many of the most avant garde musicians intentionally avoid any sort of stylistic self-identification. As such, typical web directories did not capture the semantics we hoped to test. Instead of turning to Yahoo! or DMOZ, then, we settled for a small, but hopefully more meaningful test collection. A list of 227 mainstream artists was compiled through two sources. First was the list of American top 40 singles published online at http://top40-charts.com This was supplemented by the list of the top selling 100 rock albums in Canada of 2000 and 2001 (published online a http://www.canoe.ca/JamMusicCharts/ALBUMS.html). Second, a list of 288 alternative artists was compiled from the website http://www.aquariusrecords.org. This website for a San Francisco record store was listed by two of the DJs as an alternative exemplar, but was withheld for test collection building. The 288 artists were culled from the last six months of "new releases" listings at the Aquarius website. For each of the artists in the 515 test cases (eight artists that appeared in both collections were removed from the sample), the official home pages of the corresponding artists were found manually. If no official home page was available, we took the first page returned for a Google search for the artist's name and record label[10].

	Predicted		
Actual	**Alt.**	**Main.**	**None**
Alt.	215	24	49
Main.	44	165	18

Table 7: Cocitation Classifier Confusion Matrix

	Predicted	
Actual	**Left**	**Right**
Left	61	227
Right	30	197

Table 8: Naive Bayes Confusion Matrix

The outcome of this experiment is shown in Tables 7 and 8 It is immediately obvious from Table 8 that the naive Bayes classifier had a strong bias towards mainstream documents. It achieved 86.78% accuracy on mainstream documents, versus only 21.18% for alternative documents. Its overall accuracy rate was thus 50.1%, worse than guessing at random. This appears to be at odds with the observed probabilities of the training set. However, the behavior of the naive Bayes model is more comprehensible in light of Table 9, which shows the ten terms with the highest information gain on the class variable. In fact almost all of the top 200 and most of the top 500 discriminators were proper names. Those that were not proper names often referred to individual portions of a website (e.g. "guided" and "tour" were among the top 200 terms). As in the political domain, then, this sample suggests that the vocabularies of each field of discourse

[9]It is worth stressing that this study was preliminary and of a small scale. No effort was made to find a canonical set of websites; quite the contrary. Instead, the goal here was to show how an *ad hoc* list of sites can be used to generate a useful filter

[10]searches were conducted July 7-11, 2004

largely overlap, leading lexical classifiers to overfit idiosyncrasies of the training data.

Not surprsingly, the cocitation-based classifier did have trouble with sparse link information, a fact that would have been worse if the test collection had included (often less popular) fan pages in addition to official band home pages. Nonetheless, the cocitation-based approach performed significantly better than naive Bayes. If we assume that documents with sparse in-link counts are classified as *alternative* (the more common class in the training set), then it achieved 91.67% accuracy on alternative documents and 72.69% accuracy on mainstream documents. Removing sparsely linked documents from the analysis give 89.96% and 86.34% accuracy for alternative and mainstream documents, respectively. The accuracy averaged across classes is thus 82.18% if sparse documents take the more probable class or 88.84% if those documents are ignored. Both formulations yield $p < 0.01$ with respect to the error rate of the naive Bayes classifier.

brian	rjd
unicorns	decemberists
burma	madvillain
rk	bj
eno	pedro

Table 9: Top 10 Discriminating Terms (Music)

Conclusion

Cultural orientation is a broad idea. Many documents about politics are biased to the left or right. On the other hand, many contemporary listeners put rock musicians along a spectrum of "mainstream appeal." Other types of cultural orientation abound. Some cancer patients, for instance, might be interested in alternative therapies, while others reject such practices as quackery. Cultural orientation also has a strongly individual meaning. Readers of personal blogs might like some bloggers while detesting the work of others; in such a setting each reader defines a personal notion of high versus low quality in the universe of blogs.

The technique described in this paper presents a general approach to classifying hypertext documents with respect to cultural orientation. Based on a small cohort of examples in each of several (we have presented only binary instances, but the odds ratio generalizes to more classes) "directions," the classifier presented in Equation 7 estimates the direction of unseen documents by analyzing which exemplars they tend to co-occur with online. The technique proposed here achieved accuracy over 90% for political documents and above 82% for a sample of home pages from popular musicians.

However, these experiments were not intended to demonstrate working systems. The value of an all-purpose political classifier is questionable, especially given the number of non-partisan documents and documents whose partisanship doesn't fall along the left-right spectrum used here. Instead, the value of the proposed method lies in its low barrier for implementation. Cocitation information would be put to best use with respect to cultural orientation by individual system users, who could specify the members of whatever *ad hoc* exemplar sets they deemed appropriate.

Thus future work along the lines presented here will embed the cocitation-based classifier in a system for filtering blogs with respect to topic and cultural orientation. This will require that the cocitation-based approach pursued here be integrated with lexical methods, an idea that the present experimental methodology precluded, but which is certainly desirable.

More modest research will involve further analysis of the error rate of the proposed technique. Towards this, studies will be undertaken that involve larger test collections, as well as classifiers in domains besides music and politics.

References

Agrawal, R.; Rajagopalan, S.; Ramakrishnan, S.; and Xu, Y. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the twelfth international conference on World Wide Web*. Budapest, Hungary: ACM Press. 529–535.

Agresti, A. 2002. *Categorical Data Analysis*. Wiley, 2 edition.

Beineke, P.; Hastie, T.; and Vaithyanathan, S. 2004. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Forthcoming.

Borgman, C. L., ed. 1990. *Scholarly Communication and Bibliometrics*. Newbury Park, CA: Sage Publications.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30:107–117.

Burges, Christopher, J. C. 1998. A tutorial on support vector machines. *Data Mining and Knowledge Discovery* 2(2):121–167.

Church, K. W., and Hanks, P. 1989. Word association norms, mutual information and lexicography. In *27th Annual Conference of the ACL*. New Brunswick, NJ: ACL. 76–83.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the twelfth international conference on World Wide Web*, 519–528. ACM Press.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Oxford: Philological Society. 1–32.

Gibson, D.; Kleinberg, J. M.; and Raghavan, P. 1998. Inferring web communities from link topology. In *UK Conference on Hypertext*. 225–234.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In Cohen, P. R., and Washlster, W., eds., *The Thirty-Fifth Annual Meeting of the Association for Computational Linguistics*. Somerset, New Jersey: ACL. 174–181.

Huang, Z.; Chen, H.; and Zeng, D. 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.* 22(1):116–142.

Konstan, J. A.; Miller, B. N.; Maltz, D.; Herlocker, J. L.; Gordon, L. R.; and Riedl, J. 1997. Grouplens: applying collaborative filtering to usenet news. *Commun. ACM* 40(3):77–87.

Konstan, J. A. 2004. Introduction to recommender systems: Algorithms and evaluation. *ACM Trans. Inf. Syst.* 22(1):1–4.

Mitchell, T. M. 1997. *Machine Learning*. McGraw Hill.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 79–86.

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24(4):265–269.

Turney, P., and Littman, M. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERB-1094, National Research Council Canada, Institute for Information Technology.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: ACL. 417–424.

White, H. D., and McCain, K. W. 1989. Bibliometrics. In *Annual Review of Information Science and Technology*, volume 24. Amsterdam: Elsevier. 119–186.

Wiebe, J. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, 735–740.