Knowl Inf Syst (2006) 9(4): 492–511
DOI 10.1007/s10115-005-0214-9

Knowledge and Information Systems

REGULAR PAPER

Miles Efron

Using cocitation information to estimate political orientation in web documents

Received: 8 November 2004 / Revised: 20 December 2004 / Accepted: 5 February 2005 /
Published online: 9 September 2005

Abstract This paper introduces a simple method for estimating cultural orientation, the affiliation of online entities in a polarized field of discourse. In particular, cocitation information is used to estimate the political orientation of hypertext documents. A type of cultural orientation, the political orientation of a document is the degree to which it participates in traditionally left- or right-wing beliefs. Estimating documents' political orientation is of interest for personalized information retrieval and recommender systems. In its application to politics, the method uses a simple probabilistic model to estimate the strength of association between a document and left- and right-wing communities. The model estimates the likelihood of cocitation between a document of interest and a small number of documents of known orientation. The model is tested on three sets of data, 695 partisan web documents, 162 political weblogs, and 198 nonpartisan documents. Accuracy above 90% is obtained from the cocitation model, outperforming lexically based classifiers at statistically significant levels.

Keywords Opinion mining · Style analysis · Document classification · Political orientation

1 Introduction

The current popularity of online diaries known as weblogs (blogs) raises an old problem in information filtering. Often impassioned and partisan, author-maintained weblogs remind us of the importance of opinion and sentiment in the

M. Efron (✉)
School of Information, Sanchez Building 564, 1 University Station D7000, Austin, TX 78712-0390, USA
E-mail: miles@ischool.utexas.edu

filtering task. According to the website *blogstreet.com*,[1] *Cut on the Bias*[2] and *Liberal Oasis*[3] 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. Although both sites are about politics, their orientation with respect to their topic diverges.

This paper describes a hyperlink-based method of gauging the subjective affiliations of documents in a polarized field of discourse. To motivate the technique, this paper explores the political orientation (left- or right-leaning) of web documents. The focus is on politics because political orientation is a particularly salient dimension of author and reader opinion in the weblog domain, an increasingly important area of retrieval. The internet portal Yahoo!, for example, classifies blogs into 18 categories, of which politics is the largest, a count which does not include separately listed blogs devoted to the current war in Iraq (so-called war blogs). The marriage of politics and weblogs has surfaced most visibly in the 2004 US presidential election. Led by the Democratic candidate Howard Dean, the major campaigns this year all used blogs to disseminate political opinion [21].

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 [24], which follows Firth's notion that we may learn the meaning of a word "by the company it keeps" [11]. Likewise, we may use hyperlink structure to learn about a document on the Internet [1, 6, 12]. 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[4] `link:` command to create an odds ratio that approximates the target entity's cultural orientation.

The remainder of the paper is organized as follows. Section 2 gives background on the matter of opinion as it relates to text retrieval and personalization systems. Additionally, this section defines the notions of political orientation and the broader idea of cultural orientation. In Sect. 2.1 Turney and Littman's pointwise mutual information from information retrieval (PMI-IR) algorithm is reviewed, along with the details of its adaptation for measuring political opinion in documents. To evaluate the success of the proposed estimator of political orientation, Sect. 4 reports the results of three experiments, one using traditional web documents, another explicitly concerned with weblogs, and a third analyzing nonpartisan documents. Finally, Sects. 5 and 6 reflect on the performance and limitations of the proposed method.

2 Background

Although it has long been known that relevance is a multidimensional construct [20, 22], recent developments in the literature suggest that techniques for addressing aspects of relevance aside from topicality are gaining new favor in IR research.

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

In particular, a large body of recent work is concerned with so-called opinion mining. Opinion mining classifies documents about a given topic with respect to their opinion on that subject. Thus Dave et al. [9] and Turney [23] use product reviews to construct classifiers (using a variety of methods) to detect statements of positive or negative sentiment about a given item. These efforts have achieved good results, with classification accuracy above 80% commonly reported. Extending the domain of opinion mining, Pang et al. [18] used support vector machine classification for movie reviews, achieving 82.9% accuracy. The difficulty in opinion mining and other types of sentiment classification lies in the linguistics of affect. While topicality is often detectable by lexical evidence, nuances of opinion involve subtle aspects of discourse and pragmatics. A sentence such as *I usually love Sony TVs, but this is not their best effort* contains emphatically "positive" words. But they are qualified in ways that complicate the classification problem.

2.1 PMI-IR for opinion mining

The current work is based on the PMI-IR algorithm proposed by Turney and Littman [24]. Their work defines an unsupervised approach to learning the semantic orientation of words [13, 26]. 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 described by Church and Hanks [8], the pointwise mutual information between two words w_i and w_j is given by Eq. (1):

$$PMI(w_i, w_j) = \log_2 \left(\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right), \tag{1}$$

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 Eq. (1), Turney and Littman define $SO(w_i)$ (the semantic orientation of w_i based on PMI) by Eq. (2):

$$SO(w_i) = PMI(w_i, \textbf{positive}) - PMI(w_i, \textbf{negative}), \tag{2}$$

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:

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

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

Because the counts in Eq. (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 [23] 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 approach appears to operate at state of the art levels for opinion mining [9, 18].

3 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, semantic orientation is often of less interest than partisanship as such. Readers of progressive weblogs, for instance, are unlikely to identify with conservative news media. Cultural orientation applies to domains besides politics, as well. For example, fans of mainstream music such as Britney Spears are unlikely to appreciate the hardcore punk rock of Black Flag. While this paper focuses on political orientation of documents, modeling cultural orientation more generally would be useful for personalized information systems and is feasible under the proposed framework.

3.1 Political orientation defined

Political orientation is a type of cultural orientation. *The New York Times* recently appointed Daniel Okrent as its public editor, charged with advocating on behalf of the paper's readership. To describe his biases Okrent writes, "When you turn to the paper's designated opinion pages tomorrow, draw a line from the Times's editorials on the left side to William Safire's column on the right: you could place me just about at the halfway point" [17]. Okrent alludes to the notion that American political opinion spans a spectrum, with liberal thinking commonly associated with the left, and conservatism with the right. The goal of the current work is to locate documents along this continuum, much as Okrent imagines metaphorically.

Let the political orientation of a document be defined as the degree to which it participates in the ideologies of either the political left or right. Ideally an algorithm for estimating political orientation should display several properties:

- Pages with similar outlooks should receive similar scores.
- Pages with opposing outlooks should receive inverse scores.
- Pages that are nonpolitical in their content, or that treat politics in a nonpartisan way should receive neutral scores.

With these desires in mind the following section offers a probabilistic framework for estimating cultural orientation based on hyperlink information.

3.2 Cultural orientation from cocitation analysis

While some research confronts the linguistics of affect head-on [13, 26], other work has looked to nonlinguistic features for use in opinion mining. This turn motivates the current work. 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 [1]. 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 Eq. (1) we define the pointwise mutual information between citations[5] to d_i and to another document d_j by Eq. (4):

$$PMI(d_i, d_j) = \log_2 \left(\frac{p(d_i, d_j)}{p(d_i)\, p(d_j)} \right), \tag{4}$$

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; Eq. (4) yields 0 bits of information. But if documents d_i and d_j tend to be cited together in the discursive community Eq. (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 Eq. (5):

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

where link:**R** is the AltaVista command *link:*r_1 *OR link:*r_2 ... *OR link:*$r_{|R|}$.[6] To avoid division by zero, we follow Turney in adding 0.01 to the first element of the numerator and denominator of Eq. (5) as a form of Laplace smoothing.

[5] For rhetorical simplicity, this discussion uses the words *citation* and *link* interchangeably. Although links and scholarly citations certainly imply unique semantics, this simplification allows us to describe the proposed model succinctly.

[6] 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.

3.3 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 $\mathbf{K}$. 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 Eq. (5) as a classifier corresponds with choosing the most likely distribution given the data. Eq. (5) may be rewritten as Eq. (6):

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

By definition, the numerator of the antilog of Eq. (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 $\mathbf{R}$. That is, out of $|link{:}\mathbf{R}|$ draws from B_R we observe $|link{:}d_i\ AND\ link{:}\mathbf{R}|$ "successes." Likewise, the denominator gives the MLE of p_L, the estimated probability of seeing d_i co-cited with an exemplar in set $\mathbf{L}$.

Because the number of documents citing the exemplars in $\mathbf{R}$ and $\mathbf{L}$ is large, Eq. (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 $\mathbf{R}$ is greater than the probability of a cocitation with a document in $\mathbf{L}$ then Eq. (6) is positive. Conversely, a higher likelihood of cocitation with $\mathbf{L}$ gives a negative value for Eq. (6).

Classifying a document with respect to the sign of Eq. (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 Eq. (7):

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

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 [2]. Finally, a 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 Eq. (5).

3.4 Using cocitations to estimate political orientation

To estimate the political orientation of a document d_i, 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 (also known as DMOZ).[7] Left-leaning documents were chosen by selecting those documents in the Open Directory's category *Politics:Liberalism:Social Liberalism* that had

[7] http://www.dmoz.org

Table 1 Exemplar partisan documents

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

(according to AltaVista) at least 1000 incoming hyperlinks, on December 7, 2003. Likewise, the set of right-wing exemplars was chosen from *Politics:Conservatism*. The matter of which documents should be included in these exemplar sets is discussed in Sect. 4.5.

Equation (5) describes the strength of association of d_i to the left- and right-leaning paradigms. Pages that are cocited with a preponderance of right-wing documents will score positively, while documents that appear mostly with left-wing material get a score below zero. The stronger these tendencies, the larger the absolute value of Eq. (5) will be. Documents cocited with none of the exemplars will get a constant score reflecting the ratio of in-links for each set of exemplars. If the prior probabilities (i.e. the link counts) of each class are equal, then documents with no partisan cocitations score 0, indicating neutrality.

Schematically, we may think of these scores as locations along the political spectrum, with neutrality centered at 0, as seen in Table 2. Table 2 shows a sample of home pages from widely known political websites, locating each URL as a point on the real number line, which approximates the notion of political spectrum. As we would expect *The New York Times* website falls near the center, but slightly to the left of the more conservative *Wall Street Journal*. On the other hand, the website of the progressive Green party lies far to the left, at the opposite end of the spectrum from The Heritage Foundation, a think tank known for its right-wing policy recommendations.

Equation (5) models political orientation as a social phenomenon. Applying it assumes that hyperlink structures approximate intellectual communities. While this ignores the fact that many links imply disfavor rather than endorsement of

Table 2 Example websites

Organization	URL	$\hat{\omega}$
The Green Party	www.gp.org	−4.00
The New York Times	www.nytimes.com	−0.52
The Wall Street Journal	www.wsj.com	0.51
The Heritage Foundation	www.heritage.org	1.94

Table 3 Categories for test set construction

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

their target documents, the success of link-based approaches to data mining suggests that this oversimplification introduces relatively little error [1].

4 Experimental evaluation

To evaluate the suitability of the proposed estimator of political orientation, three experiments were performed. The first experiment tests the approach on a corpus of 695 partisan web documents. In experiment two, analysis was limited to a set of 162 political weblogs. Experiment three analyzes the behavior of the proposed estimator on nonpartisan political documents.

4.1 Web data evaluation

A test corpus was constructed based on the web directories provided by Yahoo![8] and the Open Directory Project (DMOZ). For this experiment all URLs listed in the directories shown in Table 3 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:

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

Table 4 Summary of experiment one

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

Table 5 Cocitation classifier confusion matrix

	Predicted	
Actual	Left	Right
Left	276	28
Right	13	378

- Cocitations as described in Eq. (7)
- Naive Bayes: A naive Bayes classifier fitted after removal of stopwords [16]
- SVM: a support vector machine classifier trained on the same data as the naive Bayes model [7]. The SVM used a linear kernel function.

The word-based classifiers were trained on a corpus of 2412 documents.[9] 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 an attempt to mitigate the tendency of the naive Bayes and SVM methods to overfit the training data. A summary of these results appears in Table 4.

More detailed results appear in Table 5, the confusion matrix for the cocitation-derived classifications.

For left-leaning documents, the cocitation-based method achieved 90% accuracy on the test data. On the other hand it reached 97.6% accuracy for the right-leaning documents. Part of this asymmetry can be ascribed to the fact that in this experiment, documents that had zero incoming links were assigned to the most likely category (*right*). Alternately, we could assume that such documents must be classified as nonpartisan. Changing the results to reflect this heuristic, yields Table 7.

If we assume that classifying a document as nonpartisan is an error, this formulation yields an F-measure of 0.9564, with classification accuracy of 91.65%. Defining zero-linked documents as failures gives classification accuracy of 90.46% for *left* documents and 92.82% for *right* documents. Unless otherwise noted the remainder of this study assumes that it is known a priori that all documents are about politics and should be assigned the most probable score in the event of poor link evidence.

Tables 4 through 7 show a clear advantage for the link-based approach over the tested lexical methods. To formalize this, the proportions of correct to incorrect classifications for left- and right-wing documents found under the cocitation-based

[9] 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.

Table 6 SVM confusion matrix

	Predicted	
Actual	Left	Right
Left	202	90
Right	89	281

Table 7 Cocitation classifier confusion matrix

	Predicted		
Actual	Left	Right	None
Left	275	23	6
Right	13	362	16

and SVM methods were compared. Submitting the resultant contingency tables to a standard 2 × 2 chi-square test yielded $p < 0.001$ for both left- and right-leaning documents. Additionally, Table 7 was used to estimate the error rate of the cocitation-based classifier, counting nonclassified documents as errors. A chi-squared test on the error rates for the SVM and this more stringent heuristic also yielded $p < 0.001$.

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. Due to the evident inferiority of the naive Bayes lexical classifier, the remainder of this paper reports results only from the cocitation-based approach and the SVM model described above.

4.2 Blog data evaluation

As a second test, the estimator of Eq. (5) 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 of 119 left-wing blogs was downloaded from *Progressive Gold*,[10] a self-described progressive blog. The maintainer of *Progressive Gold* keeps a directory of decidedly left-leaning blogs; these comprised our sample on the left.

On the right, test data came from the conservative portal *Townhall.com*,[11] which maintains a directory of right-wing blogs. Unfortunately, *Townhall's* directory contained only 43 blogs at the time of this writing, and thus our sample is skewed to the left.

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

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

Table 8 In-links for web and blog data

In-links	Web data	Blog data
% with 0 in-links	3.16	0.6
1st quartile	14	40.5
Median	67	1548
3rd quartile	758.5	8026
Max.	249,800	118,900

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 nine 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 webpage data described above. Table 8 summarizes the connectedness of each data set.

The weblogs yielded richer link information on average than the standard web pages. While 3% of the sampled web pages had zero in-links, only 1 of the 162 (0.6%) weblogs had no 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.

4.3 Performance on nonpartisan documents

The high accuracy of the link-based approach discussed in the previous two sections suggests that the political orientation of the test documents might have been extreme. Since the web and blog test documents had been explicitly categorized as partisan, this would come as no surprise. An important question then is: how does our estimator $\hat{\omega}$ perform when presented with documents whose political orientation is less extreme? If Eq. (5) is in fact estimating each site's political orientation, it should deliver scores closer to 0 for nonpartisan documents than for documents with a clear political bias.

To test the hypothesis that politically moderate documents tend to have an orientation estimate closer to 0 than highly opinionated documents, a corpus of putatively nonpartisan documents was constructed. This corpus was built by selecting the contents of the three categories from Yahoo! the Open Directory, shown in Table 9. These categories were chosen because they were retrieved by a Google search for the phrase *nonpartisan politics*, and on inspection they appeared to

Table 9 DMOZ and Yahoo! Categories for nonpartisan documents

Source	Category
DMOZ	Government: US Government: Legislative Branch: Agencies: General Accounting Office
DMOZ	Government: US Government: Politics:Elections: Organizations: Voter Ed. & Awareness: League of Women Voters
Yahoo!	Regional: North America: US: Govt: Society/Culture: Politics
Yahoo!	Regional: North America: US: Govt: Elections: VoterEducation
Yahoo!	Society: Politics: Directories

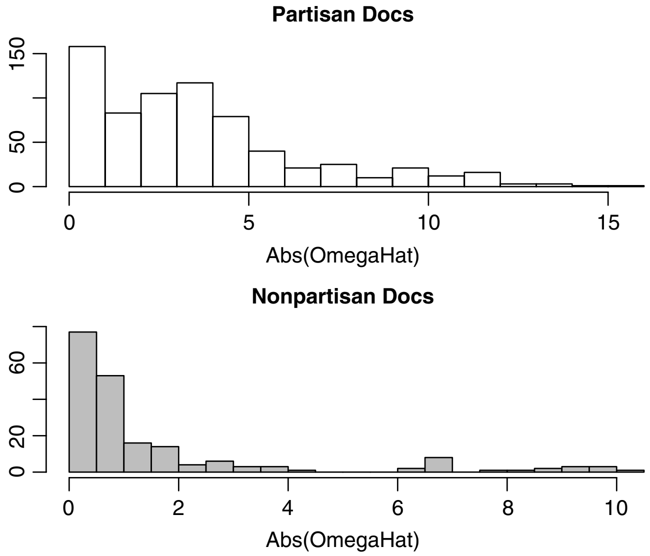

Fig. 1 Abs. Val. ($\hat{\omega}$) for partisan and nonpartisan docs

contain few highly biased documents. Selecting the documents from each of these categories led to a set of $N = 198$ putatively nonpartisan test documents.

However, it must be stressed that constructing a test set of nonpartisan political documents is difficult. Many websites that claim to offer unbiased information do so in efforts to advance a highly politicized agenda. Moreover, the categories listed in Table 9 do not explicitly preclude inclusion of biased documents. Thus the 198 test documents may in fact have a political bias. However, in an admittedly subjective sense, most of the documents in these categories appeared to offer unbiased information.

Figure 1 shows histograms of the absolute value of the orientation estimate for the 695 web documents (top panel) and the 198 nonpartisan documents (bottom panel). The general web documents appear to be bimodal, with a pronounced peak near 4 (the block of partisan documents scoring near 0 is an artifact of weak cocitation information, as described in the following section). The median absolute value of $\hat{\omega}$ for partisan documents was 3.009. On the other hand, the nonpartisan documents show a strong tendency to score near 0, with sample median

Table 10 Cocitation classifier confusion matrix

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

of 0.971. Additionally, the partisan documents show scores extending as high as $abs(\hat{\omega}) = 16$, while the highest estimates for nonpartisan documents was near 11. It must be admitted that a score of 11 suggests a highly partisan reading for a supposedly nonpartisan document. However, upon inspection, it was noted that this high scoring document was for the pro-choice voter education website *votingProChoice*,[12] a URL that entered the nonpartisan sample due to the method of gathering data, but which is unlikely to be truly above the partisan fray.

Performing a Wilcoxon rank sum test on the scores for partisan and nonpartisan documents yielded $p < 0.001$. This test suggests that the estimator modulated its output when presented with documents whose political orientation was not extreme. The significance of this test is especially encouraging since the collection of "nonpartisan" documents may have contained outlier documents that were in fact biased.

4.4 The problem of sparse link data

The obvious shortcoming of the proposed use of cocitation information in estimating political orientation lies in the problem of sparse link data. If a given document d_i has zero incoming links, the proposed approach has no evidence with which to make an estimate. Even if a document has some incoming links, if none of those entails a cocitation with an exemplar, the algorithm remains at a loss. This problem is exacerbated by the instability of the log-odds ratio when class counts sink below approximately 4.

Our test collection of 695 web documents contained 22 documents for which AltaVista offered no incoming links. More vexing, there were 137 documents that contained fewer than four cocitations with any of exemplar documents. Counting these documents as errors leads to a significant degradation in classifier performance, as noted in Table 10.

Counting documents with fewer than four cocitations as errors leads to 88% classification accuracy for left-wing documents and 69% accuracy for right-wing documents. From Table 6 we calculate the SVM's accuracy as 69% for left-wing documents and 76% for right-wing documents. The differences between the two methods' classification accuracy are not statistically significant under this definition.

Though sparse cocitation information constitutes a serious liability for the proposed approach, it bears mentioning that only a small amount of link information is needed to achieve good classification. For instance, of the 695 test documents, 80% (558 documents) showed at least 4 cocitations with exemplar documents. For these documents, the proposed method correctly classified 539, for an accuracy of

[12] http://www.votingprochoice.org

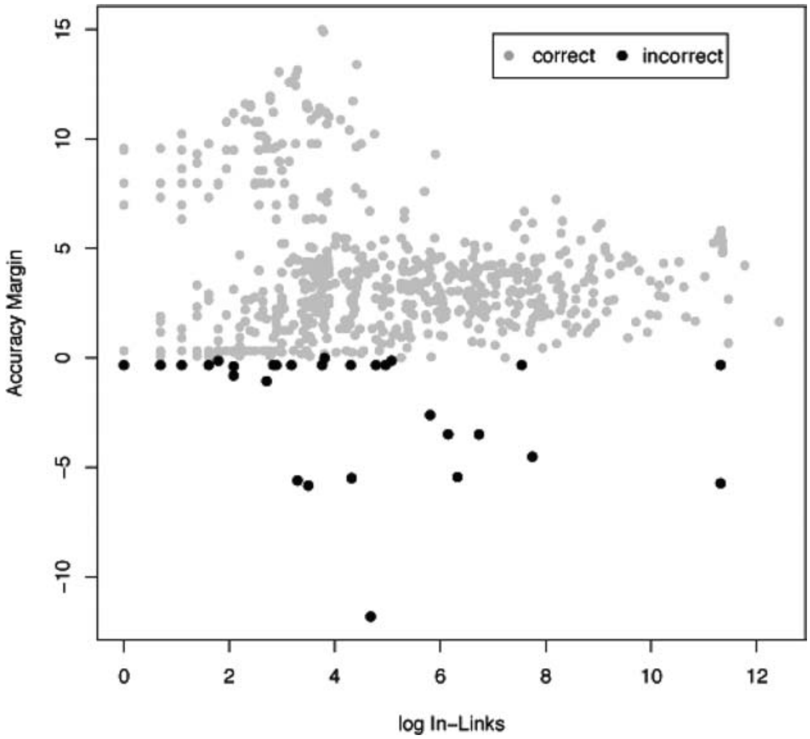

Fig. 2 log(in-links) vs. accuracy(URL)

96.6%. Thus even a small amount of link data appears to be sufficient for the technique to deliver good estimates.

This theory is borne out by Fig. 2, which plots the log of the in-link count against the margin of accuracy achieved by the proposed technique. That is, the y-axis is the magnitude of a URL's score, transformed such that positive values indicate a correct classification, while negative values show the margin of error.

While it must be borne in mind that many points in Fig. 2 overlap (e.g. 22 errors at in-link = 0), websites with low in-link counts do not appear to lead to egregious errors; instead, the errors toward the left of the plot are all small. The largest errors lie near the middle of the plot, where the majority of URLs are found, suggesting that such outliers are due to factors other than a paucity of data. The correlation between the absolute value of the political orientation estimate and the number of incoming links was only 0.026.

As discussed in Sect. 4.2, low link counts presented less of a problem on the blog data. Not only did weblogs contain more in-links on average than the general web documents, they also offered richer cocitation patterns with exemplar documents. While 137 (20%) of the 695 general web documents were cocited with fewer than four exemplar documents, only two (1.2%) of the 162 tested blogs provided untenable cocitation information. With $p < 0.01$ we conclude that blogs tend to have higher cocitation with the exemplars than non-blog political web pages. Thus sparse cocitation information appears to be less troublesome for blog data than for political web pages in general.

Table 11 Number of exemplars and model performance

Model	% accuracy	F-measure
M_{full}	94.1	0.97
M_7	93.21	0.97
M_2	65.99	0.85

4.5 Items in exemplar sets

An open question in the application of PMI-IR for concept learning in general (whether for semantic or cultural orientation of words or documents) is which items should comprise the paradigms. To motivate this choice for the experiments reported here, documents in intuitively relevant DMOZ directories that showed at least 1000 incoming links were selected. This led to the paradigms shown in Table 1. However, the selection of 1000 in-links was arbitrary. Turney and Littman use only seven exemplars in their learning paradigms, but for this study it was conjectured that the link-based approach would suffer from poor performance without sufficient exemplar data.

To test this hypothesis, the experiment described in Sect. 4.1 was repeated using fewer exemplar documents. Let M_{full} be the 19-exemplar per-class model described above. Based on this, simpler models were constructed: M_7 and M_2, which contain the most highly linked seven and two exemplars for each class, respectively. Table 11 shows classification performance for the webpage data under each of the models.

From Table 11 it is clear that M2 leads to inferior performance, compared to the larger models. However, the distinction between the 19-exemplar and 7-exemplar models is negligible. A chi-square test on each model's error rate yielded $p = 0.806$, suggesting that they are statistically indistinguishable. For the data presented here, then, no more information is required for the cocitation-based classifier than was used in Turney and Littman's SO-PMI-IR study.

4.6 Estimator stability over time

Although websites (especially blogs) are dynamic, it may be expected that an estimator of a site's political orientation should remain relatively stable over time. Except for those sites that are nonpartisan, or whose partisanship changes, most left-leaning resources retain progressive sympathies, while right-leaning sites tend to remain conservative over time. Because the estimator proposed in this paper relies on information that is external to a given resource, this section undertakes an analysis of the estimator's behavior over time.

For this analysis, 250 blogs were downloaded from the Internet in June 2004. These blogs were acquired by issuing the queries *politics and Iraq* and *politics and economy* to the blog search engine *Daypop*.[13] The sample consists of 250 blogs (125 of each estimated political orientation) drawn from the first 1000 documents that Daypop returned for these queries. Following the initial identification of this

[13] http://www.daypop.com

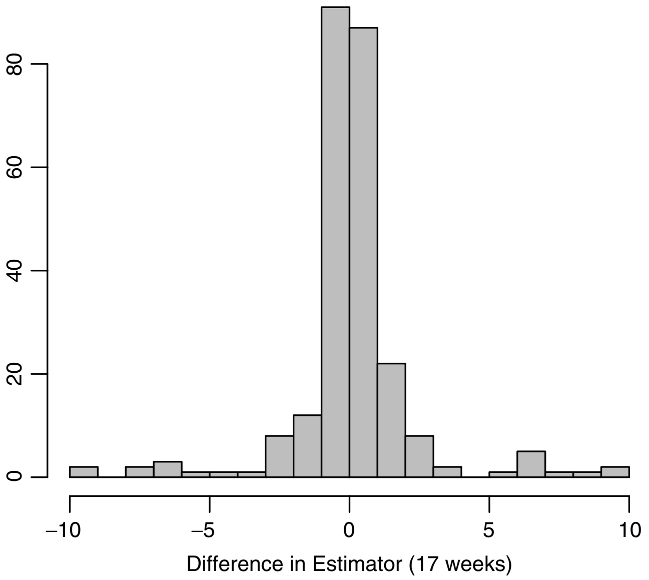

Fig. 3 Difference in $\hat{\omega}$ from week 1 to week 17

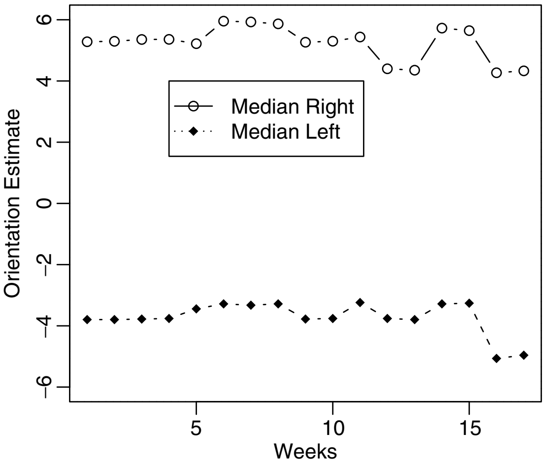

Fig. 4 Variation in $\hat{\omega}$ for two blogs

sample, the political orientation of each blog was measured weekly, for 17 weeks. It should be noted that this period included the US Presidential election.

Figure 3 shows the distribution of differences in $\hat{\omega}$ at week 1 and week 17. Although several blogs' estimated political orientation changed significantly over the four months, the majority changed very little. The mean of the difference between each site's $\hat{\omega}$ at week 1 and at week 17 was 0.147. A 95% confidence interval on $\hat{\omega}_1 - \hat{\omega}_{17}$ was $[-0.133, 0.428]$. The data suggest that over this four-month period the estimates of the blogs' political orientation did not change significantly.

Table 12 Blogs that changed orientation over 17 weeks

URL	In-links	Left cocites	Right cocites	$\hat{\omega}_{17}$
Left become Right				
http://lmeimei.blogspot.com/	226	0	1	-6.68
http://www.j-dreaming.com/weblog2.php	171	1	1	-0.02
http://hurryupharry.bloghouse.net/	90800	10429	23668	-1.21
Right become Left				
http://www.civicdialogues.org/weblog.php	1	1	0	6.63
http://www.blogsearchengine.com/blog/	5570	1	0	6.63
http://www.mjhinton.com/blogs/	180	2	1	0.96
http://www.whizzyrds.com/Windblog.html	403	9	8	0.14
http://theactivist.co.uk/Current.html	13	1	0	6.63

The general stability of the estimator, however, does evince some systematic variation, as seen in Fig. 4. This figure shows the estimated cultural orientation for two documents at each week's interval. The two documents are the blogs with median estimated cultural orientation on both the left and right. Both of these blogs show a very steady reading until approximately week 10, at which point they become less stable. Though it is not possible to ascribe a cause to this, it is worth noting that this period of time coincided with the 2004 US Presidential election, a period of high activity for political bloggers.

However, several documents' estimates did change over time. Of 125 blogs that had been classified as right-leaning at week one, three had changed to left-leaning classifications by week 17. Of the initial 125 left-leaning blogs, five switched direction at the end of 17 weeks. Details for these blogs appear in Table 12.

The majority of the blogs that changed political orientation during observation suffer from weak link information, as described above. This result demonstrates the skepticism that must accompany analysis of weakly linked resources by the methods proposed in this paper. Because the log-odds ratio is unstable on sparse data, the need for additional sources of evidence in classifying documents with respect to cultural orientation seems especially acute.

However, there are several blogs that changed orientation that have ample link evidence. In particular, *windBlog*[14] and *hurryUpHarry*[15] appear to have roughly equal proportions of readers on both the left and right, as shown in Fig. 5. With respect to link evidence, these blogs stay closely to the political center, a phenomenon that is notable in this sample due largely to its rarity.

5 Discussion

Using hyperlinks to infer communal relationships among documents is a relatively familiar and well-developed practice [3, 5, 6, 15, 25]. Gibson et al. [12] demonstrate the utility of such an approach to identifying discrete communities online. The imperative for IR to account for the communal behavior that is latent in hypertext is felt all the more keenly today, when, as Howard Rheingold argues, current technologies enable people to approach information seeking as a social

[14] http://www.whizzyrds,com/Windblog.html

[15] http://www.hurryupharry.bloghouse.net

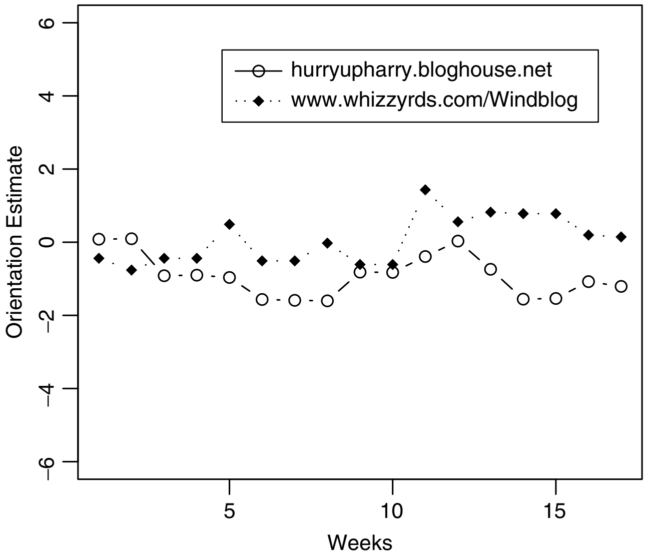

Fig. 5 Variation in $\hat{\omega}$ for two blogs

problem [19]. The popularity of highly interlinked weblogs exemplifies this trend, as does the advent of social network-based applications such as *Friendster*[16] and *Meetup.com*,[17] all of which have attained significance in the context of contemporary political discourse [10, 21].

The research presented here models political orientation as a social phenomenon, as opposed to a linguistic one. The failure of word-based classifiers to discriminate accurately between left- and right-leaning documents reinforces the intuitive notion that political orientation (like other facets of semantic affect) is communicated largely at linguistic levels above the lexicon. Unigram naive Bayes and SVM classifiers failed to discriminate adequately for these data because both liberal and conservative documents use largely the same vocabulary.

In future work it will be of interest to combine cocitation information with lexical and other sources of evidence for estimating cultural orientation. The framework provided by Beineke et al. [4] makes such an extension to the proposed model both feasible and well motivated.

Two surprising results emerge from this study. First is the fact that such a simple model of community appears to suffice for the purposes of document classification according to political orientation. As noted in Sect. 4.5, a simple 7-exemplar per-class paradigm led to 93% classification accuracy for the web data sampled here. This model obviously oversimplifies political opinion, ignoring legions of viable exemplars, not to mention other sources of evidence besides cocitation.

The second surprising result was the robustness of the cocitation-based approach to the complexity of link semantics. Central to the technique pursued here is the notion that cocitation implies similarity. However, it is the case that many authors create hyperlinks during the course of criticism, rather than endorsement.

[16] http://www.friendster.com

[17] http://www.meetup.com

Such patterns would introduce noise to the orientation estimates. However, the results reported here suggest that such errors are either not numerous, or otherwise offset by more conventional link semantics.

The ultimate utility of the work presented here lies in its application to personalization and recommender systems. In particular, this work anticipates the problem of supporting personalized information discovery in the weblog domain. In the highly opinionated discourse of blogs, filtering documents by political orientation would provide an important service. Likewise, the method could be applied to recommender systems, so that services such as online booksellers could more accurately gauge the tastes of their customers. At the risk of encouraging ideological insulation, the technique described here could derive political orientation scores for customers and authors, in efforts to satisfy user needs more effectively. In future work, the method will be extended to allow users to create their own sets of exemplar documents based on different areas of interest. This type of ad hoc personalization by hyperlink analysis was anticipated in Haveliwala's work on topic-sensitive PageRank [14]. As the name implies, however, Haveliwala's modification of the PageRank algorithm is inherently topical. As such it addresses a different type of personalization than the method pursued here.

Throughout this paper, it was assumed that all documents under analysis were about politics. Missing from this discussion is a consideration of how the proposed technique would fare when confronted with documents about, say, cars or travel. However, lexical methods have proven adept at such topical classification. Thus this omission is justified insofar as we may assume that routing political documents to the cultural orientation estimator could be handled by conventional methods. Perhaps more problematic is the matter of the method's performance on documents with moderate political views. The analysis of Sect. 4.3 yielded encouraging but preliminary results. The task remains of estimating the algorithm's error rate on moderate documents.

6 Conclusion

This study used hyperlink cocitations to estimate the political orientation of documents. Based on the work of Turney and Littman, the work extends the notion of pointwise mutual information between words and concepts, to derive a measure of association between documents and concepts. In this study, the concepts of interest were left- versus right-wing political orientation.

The proposed technique outperformed keyword-based approaches, achieving statistically significant improvement over unigram-based applications of both naive Bayes and SVM classifiers.

The technique does appear to suffer when test documents have very little incoming link data. However, although this effect was significant, the proposed technique was able to classify correctly many documents with low in-link counts. Likewise, the technique appears robust against reductions in model complexity, a quality that may lead to improved performance by recourse to alternate methods of exemplar selection.

References

1. Agrawal R, Rajagopalan S, Ramakrishnan S, Xu Y (2003) Mining newsgroups using networks arising from social behavior. In: Proceedings of the twelfth international conference on World Wide Web. ACM, Budapest, Hungary, pp 529–535
2. Agresti A (2002) categorical data analysis, 2nd edn. Wiley, Hoboken, NJ
3. Barabasi L (2002) linked: The new science of networks. Perseus, New York
4. Beineke P, Hastie T, 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, ACL, Barcelona, pp 263–270
5. Botafogo RA, Shneiderman B (1991) Identifying aggregates in hypertext. In: UK conference on hypertext, pp 63–74
6. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30:107–117
7. Burges CJC A (1998) Tutorial on support vector machines. Data Min Knowl Discov 2(2):121–167
8. Church KW, Hanks P (1989) Word association norms, mutual information and lexicography. In: 27th annual conference of the ACL, ACL, New Brunswick, NJ, pp 76–83
9. Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the twelfth international conference on World Wide Web, ACM, Budapest, Hungary, pp 519–528
10. Ehrlich E (2003) What will happen when a national political machine can fit on a laptop? The New York Times, December 14, p B01
11. Firth JR (1957) A synopsis of linguistic theory 1930–1955. Studies in linguistic analysis. Philological Society, Oxford, pp 1–32
12. Gibson D, Kleinberg J, Raghavan P (1998) Inferring web communities from link topology. In: Proceedings of the ninth ACM conference on hypertext and hypermedia: Links, objects, time and space—structure in hypermedia systems, ACM, Budapest, Hungary, pp 225–234
13. Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Cohen, P.R., Washlster W (eds) The thirty-fifth annual meeting of the association for computational linguistics, ACL, Somerset, NJ, pp 174–181
14. Haveliwala TH (2002) Topic-sensitive pagerank. In: Proceedings of the eleventh international conference on World Wide Web, ACM, Budapest, Hungary, pp 517–526
15. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
16. Mitchell TM (1997) Machine learning. McGraw-Hill, Boston, MA
17. Okrent D (2003) An advocate for times readers introduces himself. The New York Times, December 7, p 2
18. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: The 2002 conference on empirical methods in natural language processing (EMNLP), pp 79–86
19. Rheingold H (2002) Smart Mobs: The next social revolution. Perseus, New York
20. Schamber L (1994) Relevance and information behavior. In: Williams ME (ed) Annual review of information science and technology, vol 29. American Society for Information Science, Medford, NJ, pp 3–48
21. Shapiro SM (2003) The dean connection. New York Times Magazine, December 7, p 56
22. Tang R, Solomon P (1998) Towards an understanding of the dynamics of relevance judgement: an analysis of one person's search behavior. Inf Process Manage 34:237–256
23. 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, ACL, Philadelphia, PA, pp 417–424
24. Turney P, 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
25. Watts DJ (2003) Six degrees: The science of a connected Age. W.W. Norton, New York
26. Wiebe J (2000) Learning subjective adjectives from corpora. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence, AAAI Press/MIT Press, Austin, TX, pp 735–740