Mining Newsgroups Using Networks Arising From Social Behavior

Rakesh Agrawal Sridhar Rajagopalan Ramakrishnan Srikant Yirong Xu

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

ABSTRACT

Recent advances in information retrieval over hyperlinked corpora have convincingly demonstrated that links carry less noisy information than text. We investigate the feasibility of applying link-based methods in new applications domains. The specific application we consider is to partition authors into opposite camps within a given topic in the context of newsgroups. A typical newsgroup posting consists of one or more quoted lines from another posting followed by the opinion of the author. This social behavior gives rise to a network in which the vertices are individuals and the links represent "responded-to" relationships. An interesting characteristic of many newsgroups is that people more frequently respond to a message when they disagree than when they agree. This behavior is in sharp contrast to the WWW link graph, where linkage is an indicator of agreement or common interest. By analyzing the graph structure of the responses, we are able to effectively classify people into opposite camps. In contrast, methods based on statistical analysis of text yield low accuracy on such datasets because the vocabulary used by the two sides tends to be largely identical, and many newsgroup postings consist of relatively few words of text.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; J.4 [**Computer Applications**]: Social And Behavioral Sciences; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph Algorithms*

General Terms

Algorithms

Keywords

Data Mining, Web Mining, Text Mining, Link Analysis, Newsgroup, Social Network

1. INTRODUCTION

Information retrieval has recently witnessed remarkable advances, fueled almost entirely by the growth of the web. The fundamental feature distinguishing recent forms of information retrieval from the classical forms [25] is the pervasive use of link information [3] [5]. Drawing upon the success of search engines that use linkage information extensively, we study the feasibility of extending link-based methods to new application domains.

Interactions between individuals have two components:

1. The content of the interaction — the "text."
2. The choice of person who an individual chooses to interact with — the "link."

Making determinations about values, opinions, biases and judgments purely from a statistical analysis of text is hard: these operations require a more detailed linguistic analysis of content [24] [28]. However, relatively simple methods applied to the link graph might discern a great deal.

1.1 Application Domain

We selected the mining of newsgroup discussions as our sandbox for this study. We chose this domain for the following reasons.

- Newsgroups are rich sources of publicly available discourse on any conceivable topic. For instance, Google has integrated the past 20 years of Usenet archives into Google Groups (groups.google.com), and offers access to more than 700 million messages dating back to 1981.
- Newsgroups discussions are mostly open, frank, and unadulterated. Newsgroup postings can provide a quick pulse on any topic.
- Finally, extracting useful information from a newsgroup using conventional text mining techniques has been hard because the vocabulary used in the two sides of an issue is generally identical and because individual postings tend to be sparse.

The structure of newsgroup postings Newsgroup postings tend to be largely "discussion" oriented. A newsgroup discussion on a topic typically consists of some seed postings, and a large number of additional postings that are responses to a seed posting or responses to responses. Responses typically quote explicit passages from earlier postings. We can use these quoted passages to infer an implicit "social network" between individuals participating in the newsgroup.

DEFINITION 1 (QUOTATION LINK). There is a *quotation link* between person i and person j if i has quoted from an earlier posting written by j.

Quotation links have several interesting social characteristics. First, they are created without mutual concurrence: the person quoting the text does not need the permission of the author to quote. Second, in many newsgroups, quotation links are usually "antagonistic": it is more likely that the quotation is made by a person

WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM 1-58113-680-3/03/0005.

challenging or rebutting it rather than by someone supporting it. In this sense, quotation links are not like the web where linkage tends to imply a tacit endorsement.

1.2 Contributions

Our target application is the categorization of the authors on a given topic into two classes: those who are "for" the topic and those who are "against". We develop a graph-theoretic algorithm for this task that completely discounts the text of the postings and only uses the link structure of the network of interactions. Our empirical results show that the link-only algorithm achieves high accuracy on this hard problem.

Figure 1 shows an example of the results produced by our algorithm when applied to social network extracted from newsgroup discussions on the claim handling of a specific insurance company.

1.3 Related Work

The work of pioneering social psychologist Milgram [19] set the stage for investigations into social networks and algorithmic aspects of social networks. There have been more recent efforts directed at leveraging social networks algorithmically for diverse purposes such as expertise location [17], detecting fraud in cellular communications [9], and mining the network value of customers [8]. In particular, Schwartz and Wood [26] construct a graph using email as links, and analyze the graph to discover shared interests. While their domain (like ours) consists of interactions between people, their links are indicators of common interest, not antagonism.

Related research includes work on incorporating the relationship between objects into the classification process. Chakrabarti et al. [6] showed that incorporating hyperlinks into the classifier can substantially improve the accuracy. The work by Neville and Jensen [22] classifies relational data using an iterative method where properties of related objects are dynamically incorporated to improve accuracy. These properties include both known attributes and attributes inferred by the classifier in previous iterations. Other work along these lines include co-learning [2] [20] and probabilistic relational models [10]. Also related is the work on incorporating the clustering of the test set (unlabeled data) when building the classification model [13] [23].

Pang et al. [24] classify the overall sentiment (either positive or negative) of movie reviews using text-based classification techniques. Their domain appears to have sufficient distinguishing words between the classes for text-based classification to do reasonably well, though interestingly they also note that common vocabulary between the two sides limits classification accuracy.

1.4 Paper Organization

The rest of the paper is organized as follows. In Section 2, we give a graph theoretic formulation of the problem. We present results on real-life datasets in Section 3, including a comparison with alternate approaches. Section 4 describes sensitivity experiments using synthetic data. We conclude with a summary and directions for future work in Section 5.

2. GRAPH-THEORETIC APPROACH

Consider a graph $G(V, E)$ where the vertex set V has a vertex per participant within the newsgroup discussion. Therefore the total number of vertices in the graph is equal to the number of distinct participants. An edge $e \in E$, $e = (v_1, v_2), v_i \in V$, indicates that person v_1 has responded to a posting by person v_2.

2.1 Unconstrained Graph Partitioning

Optimum Partitioning Consider any bipartition of the vertices into two sets F and A, representing those *for* and those *against* an issue. We assume F and A to be disjoint and complementary, i.e., $F \cup A = V$ and $F \cap A = \phi$. Such a pair of sets can be associated with the cut function, $f(F, A) = |E \cap (F \times A)|$, the number of edges crossing from F to A. If most edges in a newsgroup graph G represent disagreements, then the following holds:

PROPOSITION 1. The optimum choice of F and A maximizes $f(F, A)$.

For such a choice of F and A, the edges $E \cap (F \times A)$ are those that represent antagonistic responses, and the remainder of the edges represent reinforcing interactions. Therefore, if one were to have an algorithm for computing F and A optimizing f as above, then we would have a graph theoretic approach to classifying people in the newsgroup discussions based solely on link information.

This problem, known as the *max cut* problem, is known to be NP-complete, and indeed was one of those shown to be so by Karp in his landmark paper [15]. The situation on the problem remained unchanged until 1995, when Goemans and Williamson [11] introduced the idea of using methods from Semidefinite Programming to approximate the solution with guaranteed bounds on the error better than the naive value of $\frac{3}{4}$. Semidefinite programming methods involve a lot of machinery, and in practice, their efficacy is sometimes questioned [14].

Efficient Solution Rather than using semidefinite approaches, we will resort to spectral partitioning [27] for computational efficiency reasons. In doing so, we exploit particularly two additional facts that hold in our situation:

1. Rather than being a general graph, we have an instance which is largely a bipartite graph with some noise edges added.
2. Neither side of the bipartite graph is much smaller than the other, i.e., it is not the case that $|F| << |A|$ or vice versa.

In such situations, we can transform the problem into a min-weight approximately balanced cut problem, which in turn can be well approximated by computationally simple spectral methods. We detail this process next.

Consider the co-citation matrix of the graph G. This graph, $D = GG^T$ is a graph on the same set of vertices as G. There is a weighted edge $e = (u_1, u_2)$ in D of weight w if and only if there are exactly w vertices $v_1 \cdots v_w$ such that each edge (u_1, v_i) and (u_2, v_i) is in G. In other words, w measures the number of people that u_1 and u_2 have both responded to. Clearly, w can be used as a measure of "similarity", and we can use spectral (or any other) clustering methods to cluster the vertex set into classes. This leads to the following observations.

OBSERVATION 1 (EV ALGORITHM). The second eigenvector of $D = GG^T$ is a good approximation of the desired bipartition of G.

OBSERVATION 2 (EV+KL ALGORITHM). Kernighan-Lin heuristic [18] on top of spectral partitioning can improve the quality of partitioning [16].

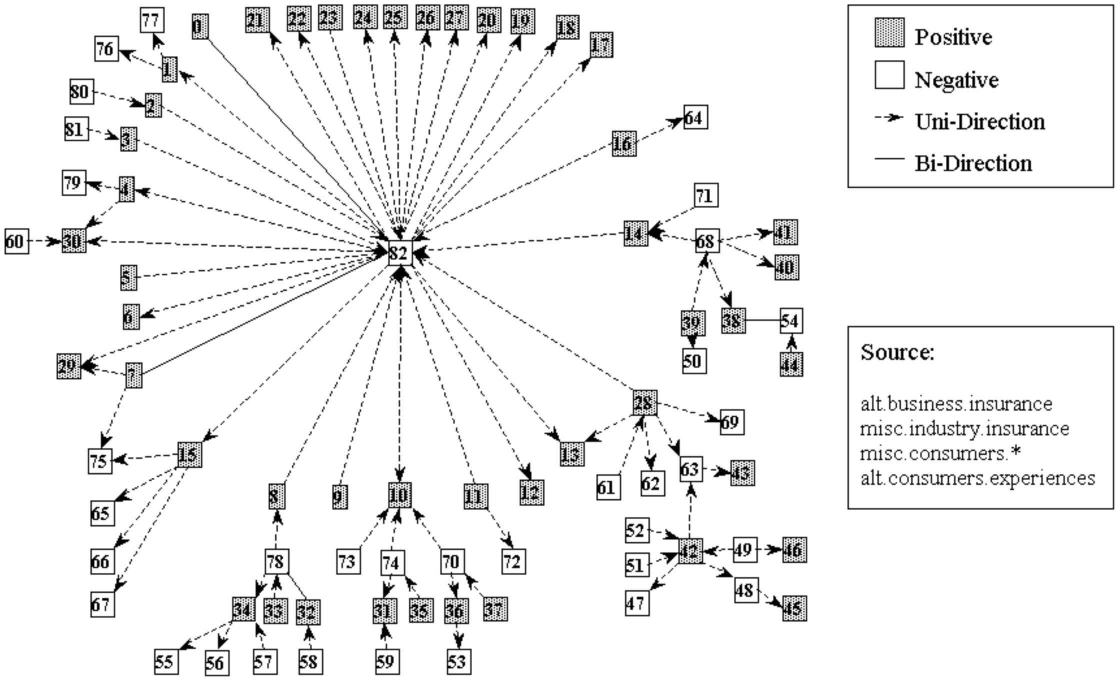

Figure 1: Claim handling of an insurance company.

2.2 Constrained Graph Partitioning

A limitation of the approach discussed so far is that while high accuracy implies high cut weight, the reverse may not be necessarily true. We may need to embellish our partitioning algorithms with some focusing to nudge them towards partitions that have high cut weight as well as high accuracy.

The basic idea will be to manually categorize a small number of prolific posters and tag the corresponding vertices in the graph. We then use this information to bootstrap a better overall partitioning by enforcing the constraint that those classified on one side by human effort should remain on that side during the algorithmic partitioning of the graph. We can now define the constrained graph partitioning problem:

PROBLEM 1 (CONSTRAINED PARTITIONING). Given a graph G, and two sets of vertices C_F and C_A, constrained to be in the sets F and A respectively, find a bipartition of G that respects this constraint but otherwise optimizes $f(F, A)$.

We can achieve the above partitioning by using a simple artifice: all the positive vertices are condensed into a single vertex, and similarly for the negative vertices, before running the partitioning algorithm. Since EV and EV+KL algorithms cannot split the single vertex, all we need to check is that the final result has these two special vertices on the correct sides. The corresponding algorithms will be referred to as the *Constrained EV* and *Constrained EV+KL* algorithms.

3. EMPIRICAL EVALUATION

We next present the results of our empirical evaluation of the link-only approach.

3.1 Experimental Setup

3.1.1 Raw Data

We created three datasets from the archives of the Usenet postings:

- **Abortion:** The dataset consists of the 13,642 postings in *talk.abortion* that contain the words “Roe” and “Wade”.
- **Gun Control:** The dataset consists of the 12,029 postings in *talk.politics.guns* that include the words “gun”, “control”, and “opinion”.
- **Immigration:** The dataset consists of the 10,285 postings in *alt.politics.immigration* that include the word “jobs”.

3.1.2 Extracting Social Network

As stated in Section 1, links are implicit in the postings in the form of “in response to” tags. We retained from the raw data only those postings that contained both the author and the person whom the author was responding to. Each such posting yields a link. We were thus able to use 65% to 80% of the postings (see Figure 2). The remaining postings were either not responses, or we were not able to find or match the name of the original poster.

Because of the sampled nature of the data, if we were to form the graph corresponding to all the responses, parts of the graph would contain too little information to be analyzed effectively [26]. Each data set contained a core connected component (see Figure 3) that comprised almost all the postings within the data set. We, therefore, omitted all vertices (and corresponding documents) that do not connect to the core component. Figure 3 also shows the total number of authors and the average number of postings per author

	Abortion	Gun Control	Immigration
Postings	13,642	12,029	10,285
Extracted Responses	10,917	8,640	6,691
... as % of Postings	80%	72%	65%

Figure 2: Number of response postings extracted from the raw data.

	Abortion	Gun Control	Immigration
Responses	10,917	8,640	6,691
Retained Responses (R)	10,821	8,380	6,473
Authors (A)	2,525	2,632	1,648
Ratio (R/A)	4.3	3.2	3.9

Figure 3: Characteristics of the core component. Between 97% and 99% of the extracted responses were in a single connected subgraph.

in the core component.

3.1.3 Links

To understand the nature of links in the newsgroups, we manually examined 100 responses (split equally between the three datasets). We found two interesting behaviors:

- The relationship between the two individuals in the newsgroup network is much more likely to be antagonistic than reinforcing.
- Many authors go off-topic and cause the discussion to drift off to an unrelated subject.

We summarize the results in Figure 4. Overall, around 74% of the responses were antagonistic, 7% reinforcing, and 19% off-topic. The results were quite uniform across all 3 datasets.

3.1.4 Test Data

For each dataset, we manually tagged about 50 random people (not postings) into “positive” and “negative” categories to generate the test set for our experiments. Figure 5 gives the number of people in each category for the three datasets.

3.2 Results

Before presenting results for the graph-theoretic approach, we first give results obtained using two classification-based approaches: text classification and iterative classification. These results should be viewed primarily as reference points for whether link-based methods can yield good results.

3.2.1 Text-based Classification

	Abortion	Gun Control	Immigration	Overall
Total	34	33	33	100
Antagonistic	26	23	25	74
Reinforcing	1	4	2	7
Off-topic	7	6	6	19

Figure 4: Analysis of responses.

	Abortion	Gun Control	Immigration
Positive	29	14	28
Negative	22	36	24

Figure 5: Distribution of People in the Test Data.

		Abortion	Gun Control	Immig-ration
(a)	Majority Assignment	57%	72%	54%
(b)	SVMTorch II	55	42	55
	Naive Bayes	50	72	54
(c)	Iterative Classification	67	80	83
(d)	EV	73	78	50
	EV+KL	75	74	52
(e)	Constrained EV	73	84	88
	Constrained EV+KL	73	82	88

Figure 6: Accuracy (in percentage).

We experimented with two commonly used text classification methods: Support Vector Machines [4] [29] and Naive Bayes classifiers [12] [21]. For SVM, we used Ronan Collobert’s SVMTorch II [7]. For Naive Bayes, we used the implementation described in [1].

The training set for text-based classification was obtained by taking the postings of each of the manually tagged user and treating them as one document (after dropping quoted sections). The 10-fold cross-validation results for both Naive Bayes and SVM are shown in Figure 6(b). We also show in Figure 6(a) the accuracy for the three datasets if everyone was simply assigned to the majority class.

Both SVM and Naive Bayes were unable to distinguish the two classes, for any of the datasets. The only *apparent* exception is the 72% accuracy on Gun Control with Naive Bayes; however, this result was obtained by classifying all examples into the majority class. The reason for the low accuracy is that in a newsgroup discussion the vocabulary used by two sides is quite similar. Meaningful words are contained equally frequently in both the positive and negative classes, and the “distinguishing words” are meaningless.[1]

3.2.2 Iterative Classification

Next, we experimented with an algorithm based on the iterative classification proposed in [22]. The proposal in [22] used both the text as well as the link structure. However, given the inadequate accuracy of the text-based classifiers, we experimented with a purely link-based version, which we describe next. Let the total number of iterations be m. In each iteration i:

1. Use links from labeled data to predict class labels on unlabeled data.
2. Sort predicted labels by confidence.
3. Accept k class labels, where $k = N(i/m)$, and N is the number of instances in the test data.

Let w_{ij} be the weight of the link between vertices v_i and v_j. Let the vertices in the training set have scores of either +1 or -1

[1] More sophisticated methods may potentially improve the accuracy of text-based classification. Unlike link-based methods, these methods can also classify loners – people who neither respond to other postings nor have any responses to their postings.

(depending on their class). The score for labeled vertices in the test set is their score in the previous iteration; the score for unlabeled vertices is 0. The score s for a vertex v_i in the test set is computed as:

$$s(v_i) := \frac{\sum_j -s(v_j) \times w_{ij}}{\sum_j w_{ij}}$$

The sign of $s(v_i)$ gives the predicted class label, and $|s(v_i)|$ gives the confidence of the prediction.

Figure 6(c) gives the 10-fold cross-validation results for this algorithm. Clearly, iterative classification using the link structure does much better than the text-based classification. In some sense, this algorithm can be considered as a heuristic graph partitioning algorithm. Thus the question is: will the traditional graph partitioning algorithms do better than this heuristic approach?

3.2.3 Unconstrained Partitioning

Figure 6(d) shows accuracy results obtained using the unconstrained EV and EV+KL algorithms for the three datasets. Unlike the two earlier approaches, unconstrained partitioning does not need the manually-labelled postings as training data; thus the accuracy was computed simply by checking how many labeled individuals were assigned correct partitions.

Note that the iterative classification algorithm should really be compared to the constrained graph partitioning algorithm (that also uses training data), not the unconstrained graph partitioning. Nevertheless, it is quite interesting that even unconstrained graph partitioning does quite well on two of the datasets: Abortion and Gun Control. It does about as well as iterative classification, and much better than text-based classification. We will explain shortly why the accuracy obtained is not high for the third dataset (Immigration). KL marginally improves EV for Abortion and Immigration, but degrades EV for Gun Control.

Although the results for Gun Control are not that much better than assigning everything to the majority class, these results were not obtained by assigning everything to the majority class. In fact, all the errors were due to the misassignments of majority class to minority class. This phenomenon can be explained as follows. When the distribution of classes is skewed, it is quite likely that there will be some majority class people who will interact more with other majority class people than with the minority class. So they will be misclassified. On the other hand, most of the minority class people will be correctly classified since they will interact much more with the majority class.

To understand why EV did not do well on Immigration, we ran KL 1000 times starting with different random partitions, and measured cut weight versus accuracy. The results are shown in Figure 7. This figure confirms the observation we made in Section 2 that high accuracy implies high cut weight, but not vice versa. For Immigration, accuracy for a cut weight of 84% ranges between 50% and 87%. We therefore need to use constrained partitioning to nudge EV towards partitions that have high cut weight as well as high accuracy.

3.2.4 Constrained Graph Partitioning

Figure 6(e) shows the result of 10-fold cross-validation obtained by choosing different small subsets of the hand classified vertices as the manually tagged ones. We see that the accuracy results (compared to unconstrained graph partitioning) are now about the same for Abortion, better for Gun Control, and dramatically better for Immigration. Constrained graph partitioning does consistently (4% to 6%) better than iterative classification on all three datasets.

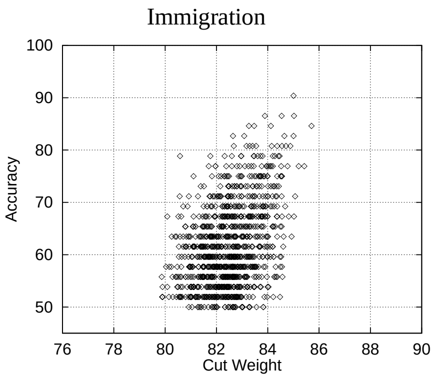

Figure 7: While high accuracy implies high cut weight, the reverse is not true.

$\vert V\vert$	Number of users	1000
I	Average postings per user	3
θ	for Zipf distribution of postings	0.8
S	Fraction of users who are "Positive"	0.6
P	Purity	0.8

Figure 8: Baseline parameters for the synthetic data experiments. In each case, when we varied one or other of our parameters, the others were kept fixed at the baseline levels.

4. SENSITIVITY EXPERIMENTS

In this section, we explore the impact of various data set parameters on the performance of the algorithms. Since it is infeasible to obtain real world data for a variety of parameter settings, we resort to synthetic data. We do this study largely to verify that the proposed algorithms are not dependent on some peculiarities of the datasets used in the experiments, and do work under a variety of settings.

4.1 Synthetic Data Generation

Figure 8 gives the baseline parameters used to create the synthetic datasets. We chose to make our classes unbalanced, i.e. 60% of the users are chosen in the majority (in our case, positive) class. Purity P is the fraction of links that cross from one class to the other; i.e. purity determines the expected number of antagonistic links in the network.

The following is a description of the data generation algorithm.

1. For each author v, the number of responses p_v that v posts is a random variable drawn from a Zipf distribution [30] with mean I and theta θ. All 3 real datasets follow a Zipf distribution for the number of postings versus rank of author, as shown in Figure 9 for the Gun Control dataset.
2. Randomly set S fraction of authors as "for" and the remaining as "against".
3. For each author, select the other users this author responds to. Let author v have p_v postings assigned in Step 1. For each of the p_v postings:

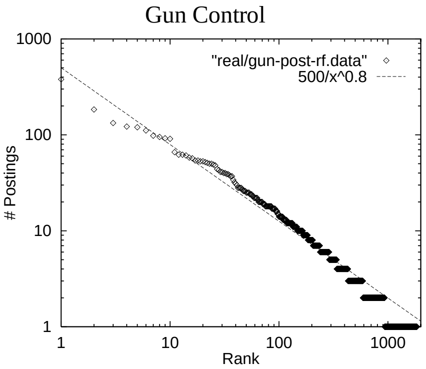

Figure 9: Prolificity follows a Zipf distribution.

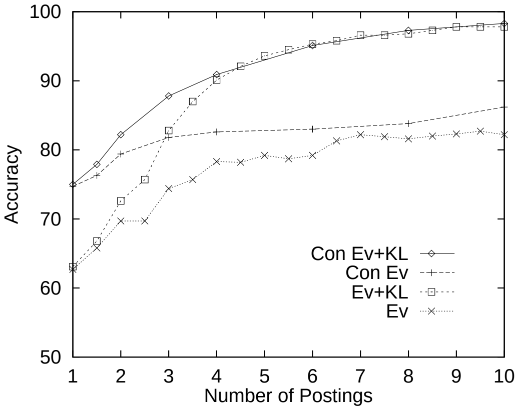

Figure 10: Accuracy rapidly increases with the number of postings.

- First, with probability P, the user is picked from the opposite side, and with probability $1-P$ from the same side.
- Within the set of users on either side, a random user is chosen to complete the link.

4.2 Experimental Results

In the first experiment, we vary the average number of postings per user, while keeping purity fixed at 0.8. Thus the ratio of antagonistic links to reinforcement links is fixed. However, what affect accuracy is not the overall fraction of reinforcement links, but rather how many users have fewer antagonistic links than reinforcement links. Not only will such users be classified incorrectly, they also make it much harder to correctly classify the people they are connected to. As the average number of postings per user increases, the probability that a specific user has more antagonistic links than reinforcement links increases significantly. Therefore accuracy increases correspondingly, as shown in Figure 10.

In the second experiment, we hold the average number of postings per author fixed at 3 and vary purity parameter P. Since the number of postings per user is small, the number of users with more antagonistic links than reinforcement links largely follows the purity fraction, and correspondingly accuracy also follows purity, as shown in Figure 11.

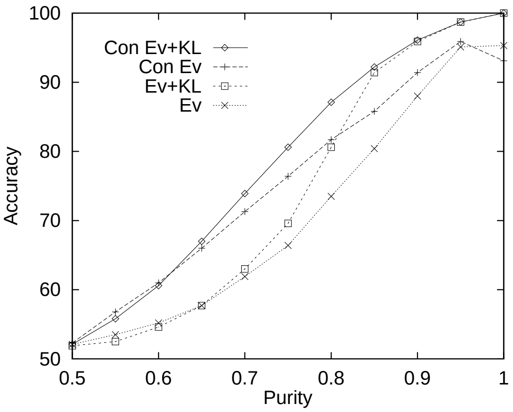

Figure 11: Accuracy follows purity when the number of postings is small.

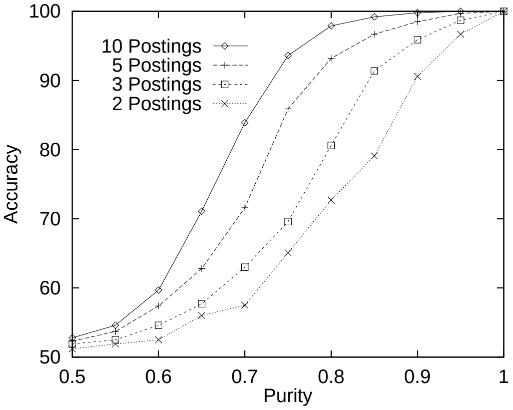

Figure 12: Accuracy jumps with the number of postings, for a given purity factor (EV+KL Algorithm).

However, when the number of postings is large, we can get high accuracies even if a large fraction of interactions are between people on the same side. The explanation is the same as that for the first experiment. Figure 12 shows the accuracy of the EV+KL algorithm for average number of postings ranging from 2 to 10, and purity ranging from 0.5 to 1. Notice that even for a low purity of 0.7 (where 30% of the postings are between people on the same side), we are able to get an accuracy of 85% when there are 10 postings per user. At a purity of 0.8, we get 98% accuracy. Thus, given sufficient interactions, the algorithm is undisturbed by a substantial fraction of reinforcement links.

5. CONCLUSIONS

Drawing upon the huge success of exploiting the link information in computations over hypertext corpora, we conjectured that links arising out of who-interacts-with-whom might be more valuable than the text of the interaction. We tested this conjecture in a setting where we completely ignored the text and relied only on links. The application we considered in this experiment was the classification of people on the two sides of an issue under discussion in a Usenet newsgroup.

We developed a links-only algorithm for this task that exhibited significant accuracy advantage over the classical text-based algorithms. These accuracy results are quite interesting, particularly when one considers that we did not even ascertain whether a response was antagonistic or reinforcing, for that would have violated the no-looking-into-the-text dictum. We simply assumed that all links in the network were antagonistic. As long as the number of reinforcing responses are relatively few, or there are enough postings per author, the algorithm can withstand this error in building the network.

For future work, it will be interesting to explore if the accuracy of the proposed techniques can be further improved by incorporating advanced linguistic analysis of the text information. More generally, it will be interesting to study other social behaviors and investigate how we can take advantage of links implicit in them.

Acknowledgments We wish to thank Xianping Ge for his help with the text-based classification experiments.

6. REFERENCES

[1] R. Agrawal, R. Bayardo, and R. Srikant. Athena: Mining-based Interactive Management of Text Databases. In *Proc. of the Seventh Int'l Conference on Extending Database Technology (EDBT)*, Konstanz, Germany, March 2000.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proc. of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers, 1998.

[3] A. Broder and M. Henzinger. Information retrieval on the web: Tools and algorithmic issues. In *Foundations of Computer Science*, Invited Tutorial, 1998.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] S. Chakrabarti. Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, 1, 2000.

[6] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, Seattle, Washington, May 1998.

[7] R. Collobert and S. Bengio. SVMTorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[8] P. Domingos and M. Richardson. Mining the network value of customers. In *7th Int'l Conference on Knowledge Discovery in Databases and Data Mining*, pages 57–66, San Francisco, California, August 2001.

[9] T. Fawcett and F. J. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[10] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1309, 1999.

[11] M. X. Goemans and D. P. Williamson. .878-approximation algorithms for MAX CUT and MAX 2SAT. In *Proc. of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing*, pages 422–431, Montréal, Québec, Canada, 23–25 May 1994.

[12] I. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M.I.T. Press, 1965.

[13] T. Joachims. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski, editors, *Proc. of ICML-99, 16th Int'l Conference on Machine Learning*, pages 200–209, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.

[14] H. Karloff. How good is the Goemans-Williamson MAX CUT algorithm? In *Proc. of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing*, pages 427–434, Philadelphia, Pennsylvania, 22–24 May 1996.

[15] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, New York, 1975.

[16] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. Technical Report TR 95-035, University of Minnesota, Dept. of Computer Science, 1995.

[17] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[18] B. W. Kernighan and S. Lin. An efficient heuristic procedure for paritioning graphs. *The Bell System Technical Journal*, pages 291–307, 1970.

[19] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[20] T. Mitchell. The role of unlabeled data in supervised learning. In *Proc. of the Sixth International Colloquium on Cognitive Science*, San Sebastian, Spain, 1999.

[21] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[22] J. Neville and D. Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2000.

[23] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proc. of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 792–799, Madison, US, 1998. AAAI Press, Menlo Park, US.

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

[25] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, New York, 1989.

[26] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.

[27] D. Spielman and S. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *37th Annual Symposium on Foundations of Computer Science*, 1996.

[28] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, July 2002.

[29] V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, Heidelberg, DE, 1995.

[30] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.