The MultiRank Bootstrap Algorithm: Semi-Supervised Political Blog Classification and Ranking Using Semi-Supervised Link Classification

Frank Lin and **William W. Cohen**
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
frank,wcohen@cs.cmu.edu

Abstract

We present a new semi-supervised learning algorithm for classifying political blogs in a blog network and ranking them within predicted classes. We test our algorithm on two datasets and achieve classification accuracy of 81.9% and 84.6% using only 2 seed blogs.

Introduction

We propose a novel algorithm that both classifies political blogs and ranks the blogs within the predicated class. We see a link to a blog of a certain political faction as a link that *endorses* that faction. In predicting the link label, we exploit a linking property found in the political blogosphere: blogs with similar political leaning tend to link to each other (Adamic & Glance 2005). We bootstrap the classification of the blogs and the links and the ranking of the blogs by propagating political leaning from an initial set of known seed nodes. We show that our algorithm achieves high classification accuracy when applied to networks of liberal and conservative political blogs using very few seeds.

Proposed Algorithm

PageRank (Page *et al.* 1998) is widely used to determine the importance or authority of a web site. However, different communities of users might attach different degrees of authority to the same site. This suggests assessing authority with an extended version of PageRank, in which every web site (and every inter-site link) is associated with a different community, and authority scores propagate only within a community. In the context of political blogs, each blog and each hyperlink would be assigned to a particular faction (e.g. liberal or conservative); below we will describe a method for assigning blogs to factions given a small set of seeds. To assess a faction-specific measure of authority, we define MultiRank as follows:

$$\mathbf{r_f} = (1-d)\mathbf{u} + dW_f\mathbf{r_f} \quad (1)$$

where W_{fij} is W_{ij} if the edge from i to j is in E_f, otherwise zero; and $\mathbf{u}$ is the uniform personalization vector where $u_i = 1/|V|$ and d is a constant damping factor. In this equation, $\mathbf{r_f}$ can be seen as the probability of a random walk on G if the we only follow edges belongs to faction f. In context of a political blog network, we can see this as the probability of a liberal/conservative blog surfer randomly clicking on links pointing to liberal/conservative blogs.

In order to calculate $\mathbf{r_f}$, we need E_f. We propose an iterative bootstrapping algorithm, shown in Figure 1, to gradually expand the set of edges E_f from a set of initial seed nodes S until the every edge in the entire graph has been labeled.

Input: A graph $G = (V, E)$, set of seed nodes S, an edge expansion metric on the graph $M(G, f)$ that returns a set of previously unlabeled edges and label them f
Output: Ranking vectors $r_{f=1...n}$ where f correspond to each faction
Algorithm:

- initialize E_f using S
- while $|\bigcup_{f=1...n} E_f| \neq |E|$ do
  - $e \leftarrow infinity$
  - while $e > 0$
    - $\mathbf{r_f} \leftarrow MultiRank(G, E_f)\forall f$
    - $label(v) \leftarrow argmax_f \mathbf{r_f}(v)\forall v \in V$
    - $E'_f \leftarrow \{e(x \rightarrow v) \in E : label(v) = f\}\forall f$
    - $e \leftarrow |E'_f - E_f|$
    - $E_f \leftarrow E'_f \forall f$
  - $E_f \leftarrow E_f \bigcup M(G, f)\forall f$

Figure 1: The MultiRank bootstrap algorithm (Exploratory Phase)

We tried two expansion metrics: the first metric simply label all currently unlabeled edges neighboring currently labeled edges with the same label as the common endpoint. The second metric is the same, except we *control* the expansion by limiting it to n unlabeled edges incident to the nodes with the highest combined ranking $\sum_f \mathbf{r_f}(v)$, where n is the number of nodes incident to labeled edges. We refer to the first metric as *infinite* expansion and the second as *controlled* expansion.

After the algorithm converges, we can classify the edges according to E_f, rank the nodes within factions according to $\mathbf{r_f}$, and classify the nodes according to $argmax_f \mathbf{r_f}(v)$.

We also present a second, optional phase to the algorithm

	Kale Infinite Expansion				Kale Controlled Expansion			
	Exploratory		Settling		Exploratory		Settling	
Seeds	Vertex	Edge	Vertex	Edge	Vertex	Edge	Vertex	Edge
2	0.641	0.763	0.819	0.968	0.787	0.898	0.804	0.952
4	0.698	0.876	0.804	0.952	0.770	0.912	0.819	0.968
8	0.703	0.894	0.804	0.952	0.785	0.949	0.819	0.968
12	0.700	0.893	0.804	0.952	0.827	0.953	0.804	0.952
16	0.728	0.917	0.804	0.952	0.824	0.953	0.804	0.952
20	0.757	0.952	0.807	0.966	0.780	0.959	0.804	0.965

	Adamic Infinite Expansion				Adamic Controlled Expansion			
	Exploratory		Settling		Exploratory		Settling	
Seeds	Vertex	Edge	Vertex	Edge	Vertex	Edge	Vertex	Edge
2	0.700	0.835	0.846	0.978	0.593	0.776	0.845	0.977
4	0.744	0.888	0.849	0.978	0.614	0.770	0.848	0.978
6	0.745	0.892	0.849	0.978	0.797	0.887	0.854	0.978
10	0.736	0.880	0.849	0.978	0.727	0.872	0.849	0.978
20	0.731	0.889	0.847	0.977	0.743	0.916	0.849	0.978
40	0.708	0.909	0.846	0.977	0.760	0.945	0.849	0.978

Table 1: Blog (Vertex) and link (Edge) classification accuracy on the Kale and Adamic datasets

that may further improve the output of the first phase. We will refer to the original algorithm shown in Figure 1 as the *exploratory* phase and the second extension algorithm as the *settling* phase. The settling phase again exploits the link property found in political blog network: blogs are more likely to links to blogs of the same political faction. First, we find all the nodes where the majority of the neighbors are of an different faction, changing the labeling of its incoming edges to the majority neighbor faction, and running the MultiRank algorithm on the modified graph. This is repeated until the algorithm converges when a) there are no more changes in edge labeling or b) when the algorithm revisits an old state due to cycling changes.

Experiments and Discussions

To assess the effectiveness of our algorithm, we tested it on two datasets. The first dataset is constructed in the same way as described in (Kale *et al.* 2007), where we ended up with a graph of 404 connected blogs. We will refer to this as the Kale dataset. The second dataset is constructed by simply creating a graph from (Adamic & Glance 2005) and taking the largest connected component. This dataset contains 1222 connected blogs and we refer to it as the Adamic dataset. It should be pointed out that the dataset labeling is not 100% accurate as noted in (Adamic & Glance 2005).

We run our algorithm on the two datasets varying three parameters: the number of seed nodes, the expansion metric, and the inclusion or exclusion of the optional "settling phase." In all our experiments, we pick seeds according to the top n PageRanked blogs, $n/2$ per faction. In all instances of the MultiRank algorithm the damping factor d is set to 0.85, a popular choice of damping factor which we borrowed without further tuning.

We point out some observations on the effect of the three variables. First, inclusion of the optional settling phase tends to improve upon the results of the first exploratory phase up to an almost constant point regardless of the number of seeds with the exception of controlled expansion with 12 and 16 seeds on the Kale dataset, where settling phase actually hurt the performance. Second, increasing the number of seeds improves the performance of the exploratory phase, but not with the addition of the settling phase, which works surprisingly well with only two seeds. Third, in general, controlling the expansion seems to help classification accuracy.

Another interesting property of this algorithm is that most classification errors are made on blogs with lower PageRank. If blogs are ordered by PageRank, the error rate on the top quartile of blogs is 0.05, while the error rate on the bottom quartile is 0.45 (data not shown due to space limitations).

Conclusions

We have introduced a new semi-supervised algorithm for simultaneously classifying and ranking political blogs based on link structure. We showed that this algorithm requires very few initial seeds to achieve performance above 80% on two political blog datasets of different size and link structure. This algorithm tend favor more authoritative blogs in terms of classification accuracy.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*.

Kale, A.; Karandikar, A.; Kolari, P.; Java, A.; Finin, T.; and Joshi, A. 2007. Modeling trust and influence in the blogosphere using link polarity. In *ICWSM 2007*.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1998. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.