Using Group Membership Markers for Group Identification

Jean Mark Gawron, Dipak Gupta, Kellen Stephens,
Ming-Hsiang Tsou, Brian Spitzberg and **Li An**
Gawron@mail.sdsu.edu
San Diego State University

Abstract

We describe a system for automatically ranking documents by *degree of militancy*, designed as a tool both for finding militant websites and prioritizing the data found. We compare three ranking systems, one employing a small hand-selected vocabulary based on *group membership markers* used by insiders to identify members and member properties (*us*) and outsiders and threats (*them*), one with a much larger vocabulary, and another with a small vocabulary chosen by Mutual Information. We use the same vocabularies to build classifiers. The ranker that achieves the best correlations with human judgments uses the small *us-them* vocabulary. We confirm and extend recent results in sentiment analysis (Paltoglou and Thelwall 2010), showing that a feature-weighting scheme taken from classical IR (TFIDF) produces the best ranking system; we also find, surprisingly, that adjusting these weights with SVM training, while producing a better classifier, produces a worse ranker. Increasing vocabulary size similarly improves classification (while worsening ranking). Our work complements previous work tracking radical groups on the web (Chen 2007),which classified such sites with heterogeneous indicators. The method combines elements of machine learning and behavioral science, and should extend to any group organized for collective action.

1 Introduction

For a variety of reasons, the problem of identifying web documents produced by particular groups has become a central concern for law-enforcement organizations, corporations, NGOs, social scientists, and public health agencies. This may be because a group espouses an agenda calling for large-scale social change (hate groups, or political movements), or it may be that there is an issue around which members have rallied for collective action (the anti-vaccine movement, global warming, cancer and HMOs). In many cases the set of websites that promote a particular idea are only loosely linked together or not linked at all. Understanding what is going on with such groups – who they are attracting, what their successes are, where they are succeeding, their demographics – requires data. The focus of this paper is on how to collect such data, using the domain of white militant hate groups as an example.

The problem of finding relevant data is not trivial. Keyword searches on standard search engines turn up large numbers of false postives, because of the ambiguity of many crucial terms, because the sites are by their nature rare, and because they are not regarded as authoritative by the usual search engine ranking criteria (much hate group action business is conducted on blogrings). For example, a search on the keyword "ZOG" (Zionist Occupational Government) turned up Wikipedia pages (discussing hate group acronyms), game sites, commercial sites, and a zydeco band. A more fruitful approach, pursued on the pioneering Dark Web project as well as by others (Chen 2007), is to crawl seed sites (provided by such resources as the ADL and SPLC websites and filtered search engine searches), download data, and do link and text analysis to find more data. This method can produce vast quantities of data; the Dark Web case study on U.S. domestic extremist groups reports 400,000 pages downloaded.

In this paper we focus on the problem of building militancy ranker, a system that ranks pages by *degree of militancy*. One reason for this is that this is the right criterion of relevancy: The most militant documents are the most interesting. Another is that it provides a way of getting a handle on what may be massive amounts of data, one kind of problem encountered on the Dark Web project, and a general problem with web search (rarely is the issue too little data).

The key challenge for both a classifier and a ranker is identifying randomly crawled pages as the products of known groups. This *group identification problem* is closer to the problem of identifying a particular author or a sentiment than it is to the problem of identifying a particular document topic: A variety of subjects might arise on relevant pages, from music to electoral politics to online gaming. When we add the issue of ranking by degree of militancy the problem becomes even more like one of the problems of sentiment analysis, since we are trying to locate a point on a scale.

The task of tracking the sites of a particular group organized around a set of ideas is closely linked to the problem of identifying orientation pro- or con-, a difficult problem. We suggest that bringing group identification into the mix may provide information complementary to and beneficial to markers of a particular topic: knowing what the pro- or con- orientation is can be much easier if we know who’s speaking. Experience in the domain of sentiment analysis teaches us that where subtler linguistic discriminations are called for, multiple independent sources of information make for more reliable classification (Malouf and Mullen 2007; Prabowo and Thelwall 2009). General social science motivations for linking group identity markers to degree of group membership are outlined below. (Section 2). The significant result is that ideas based on what we know about the psychology of groups can be integrated into a machine learning framework to produce an effective automatic ranking system.

2 Group Identity Markers

The “rational actor” hypothesis, arguably the most widely accepted assumption in the social sciences, explains patterns of human behavior as the natural result of individuals acting in their own individual interest. In contrast, the growing field of social psychology starting, inter alia, with the seminal work of Tajfel (1978) is busily accumulating evidence of the importance of groups and group-identification in our decision-making process. Our group or collective identity can even supersede our individual identity, in the sense that we may embark upon courses of action detrimental to our personal economic wellbeing, liberty, and life itself. This possibility is strongest in groups in which the sense of group-identification is strongest.

Our starting assumption is that groups organized for large-scale collective action – from nationalism (Anderson 2003) to terrorism (Gupta 2008) – will have members with strong senses of group identification. This is because such groups require a clear articulation not only of who “we” are, but also who“they” – the outsiders, the other, the unbelonging, often, the enemies – are. An essential part of the process of dividing us from them is developing a group sublanguage. This may have a complex array of linguistic components, ranging from phonological to syntactic features, but an essential feature is evaluative language referring to us and to them, as well as language referring to properties of us and properties of them. For well-established groups with a longer history the language includes a complex set of references to heroes, leaders, victims, and artists, as well as to subgroups, key events, key dates, and key writings and key works of art, including music and games.

Although group formation requires identification of “us” and ”them,” the mobilization of a large number of people for collective action requires a third factor: a clear articulation of an impending existential threat (Gupta 2008). In accord with this idea, our us-them analysis will target language articulating threats as well as language referring to the enemy. In the white militant example, the in-group is members of the white race, the out group or enemies are the non-white population, including Jews and, depending on the group, the Catholics, but significantly, also a group of white people who are traitors to the race. The general existential threat is the degradation and pollution of pure white stock, but there are many more specific instantiations because degradation has many aspects.

We refer to the elements of the group sublanguage referring to us and them and to properties and products of us and them and to existential threats to us as the **us-them language**. Our hypothesis is that the elements of the us-them language are strong markers of group identity. Morever, the us-them language is largely learned, with more experienced speakers using it more fluently and more frequently. Speakers/writers who control a significant subset of this language are likely to be well-established in the group, and identifying a significant set of such markers in a text provides strong evidence of core group membership, in our case, a high degree of militancy.

As a first stab at implementing these ideas we took a representative sample of 74 militant web pages we judged to be extremely militant and extracted from them all proper names and all noun groups referring either to us-groups or them-groups including generic references to the groups as an entirety (*white people* or *ZOG*, for example). We also extracted Verb Phrases and Adjective Phrases referring to properties clearly identifiable as properties of us and them, as well as nominalizations referring to actions by us and them, and Noun phrases referring to threats or to symbolic or artistic products of us and them (such as *Collosians 2:8* and *African liberation flag*). Examples are given in Table 1.

Us	Them
Actions, products, threat	
be a hero	back gay marriage
be prepared	hate Jesus Christ
fight for white rights	deceive
protect your children	promote homosexuality
spread the good news	suppressing the truth
freedom	crimes against humanity
hatred for the federals	cruise missiles
home schooling	cultural communists
personal responsibility	African liberation flag
Collosians 2:8	AIDs plague
People, orgs	
Klansmen	ACLU
our revolution	CIA
leaderless resistance	ADL
our white brothers	ATF
Adolf Hitler	Bill Clinton
Ian Stuart Donaldson	Colin Powell
James Madison	Jesse Jackson
Branch Davidian	Jewry
family	democratic elite
folk	black community
heterosexual whites	liberals

Table 1: Sample White militant Us-Them phrases

3 The ranking problem

Given some feature space $\mathbb{X}$, we seek a learning algorithm L that provides a **ranking function** γ that assigns a real number score to each element of $\mathbb{X}$:

$$\gamma : \mathbb{X} \rightarrow \mathbb{R}.$$

We assume a fixed set of classes $\mathbb{C}$ (which we will take to be $\{-1,\ 1\}$) and a set of labeled training documents $\mathbb{D}$,

$$\mathbb{D} \subset \mathbb{X} \times \mathbb{C}.$$

The ranking function γ is trained on D. That is, given D, the learning algorithm L produces γ.

We consider a class of linear models in which the n-ary ranking vector for the n documents in D is computed as follows:

$$s = w^{\mathrm{T}} \cdot \mathrm{D},$$

where w is an m-ary weight vector containing weights for each of m features. Thus learning a ranking function γ is learning the weights in w. For feature choice we again follow a standard assumption, the bag of words model: Each feature in a document vector represents a word. Thus m is the size of some fixed vocabulary V and feature choice is vocabulary choice. This formulation allows us to investigate using linear classsifiers as rankers. In particular, if we use maximum margin classifiers (SVMs), the margin of a test example can be taken to determine "degree of militancy". As we will see below, this assumption is not unproblematic.

We focus here on the problem of vocabulary choice. Though vocabulary (or feature) choice is in principle less significant for a maximum margin model like an SVM, because learned weights can devalue less significant words, the problem of feature choice re-emerges once we train a ranking system rather than a classifier.

To train a ranking system, we could in principle train a system on data sorted into multiple classes. But it is much more difficult to get annotators to agree on what militancy score to assign documents than it is to get them to agree on whether they are or are not militant. We thus chose to train a standard binary classifier on data classified as either positive or negative (1 or -1) using its margin as our ranking score.

The downside of this approach is that such a classifier isn't trained to assess degree of militancy; although our maximum margin classifier assigns a confidence score, high confidence in a militant rating is not the same as belief in a high *degree* of militancy. Using all vocabulary features or features selected by mutual information, a standard classifier which performs excellently on the classification task can perform quite poorly on the ranking task.

Our hypothesis is that if we focus on features that all have the property that they signal group identification (whatever they denote), the presence of more such features will reliably indicate greater militancy.

The motivation for attending to feature choice then is quite similar to that in sentiment detection where noise-eliminating strategies such as word and phrase selection based on semantic orientation or subjectivity have proven to be of help (Turney 2002).

4 The experiment

We built 6 systems, testing them both as rankers and as classifiers, as well as testing them with 3 different feature sets. The 6 ranking models are the results of variation in two dimensions. In one dimension, we build models that used syntactic information, versus syntactic combined with TFIDF weights, versus TFIDF weights alone. The syntactic model is described in Gawron, et al. 2012. On the other dimension, we built SVMs versus simple weighted systems that use linear combinations of similarity scores and/or TFIDF weights to compute a document score. To turn the simple scoring systems into classifiers, we chose a decision threshhold based on optimizing the F-score on the training set.

We describe the SVMs first. The table below contains the feature value computed for each word w_i in d_k. Sim computes the cosine similarity of the distributional vector of w_i in d_k with the syntactic model distributional vector for w_i.

SVM Sim	$\mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$
SVM TFIDF	$\frac{\mathrm{count}(w_i, d_k)}{\sum_j \mathrm{count}(w_j, d_k)} * \log \frac{N}{\mathrm{doc\_freq}(w_i)}$
SVM Sim + TFIDF	$\mathrm{TFIDF}(w_i, d_k) \cdot \mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$

The simple weighted systems all use the same document vectors as the 3 SVM systems, but rather than learning weights for the vector components and doing a weighted sum to compute a ranking score, the vector components are simply summed:

	$\mathrm{score}(v^k)$
Sim + TFIDF	$\sum_j \mathrm{TFIDF}(w_j, d_k)\mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$
Sim	$\sum_j \mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$
TFIDF	$\sum_j \mathrm{TFIDF}(w_j, d_k)$

For all 6 system designs, we built ranking systems using 3 vocab sets as features:

1. The full vocabulary used in all our collected militant docs (full vocab), minus stopwords.
2. A vocabulary chosen by sorting the entire militant vocabulary by its mutual information with militant-class document and choosing the top 3500 words (MI vocab).
3. A vocabulary consisting of all the nonstop words that showed up in the phrases of our group-marker us-them analysis, minus stopwords (us them vocab).

For the ranking experiment we collected 22 hand-selected sites ranging from totally non militant to militant with a sample of 3 web pages from each site. We instructed 5 human subjects rank sites on a scale from 1 to 10 for militancy, using promoting the superiority of the white race and advocating violent means to achieve racial separation as criteria.[1] To evaluate our ranking systems, we had the systems compute militancy scores for each of the 66 pages, and then rank each website according to its highest scoring page (this seemed to reflect how our subjects judged militancy: one very militant page out of 3 was enough to rank a site highly). We then computed the average Spearman correlation rank coefficients of each of the ranking systems with the human subjects.

[1] http://bulba.sdsu.edu/SWIDSAS/militant˙eval˙welcome.shtml

5 Results

System	Vocabulary		
	Full	MI	Us Them
Sim	-0.13	-0.18	0.05
TFIDF	-0.03	-0.06	0.46
TFIDF + Sim	-0.44	-0.37	0.34
SVM Sim	-0.30	-0.28	0.01
SVM TFIDF	-0.10	-0.14	0.20
SVM TFIDF + Sim	-0.37	-0.41	0.05

Table 2: Average system correlations with human rankings. Average human-human correlation: .82. Only positive/negative diffs are significant. P-value for 0.05 v. 0.46 = 0.08

Table 2 shows the average Spearman rank correlation coefficients with humans for the 6 system designs and the 3 vocabularies. For comparison, the average human to human correlation is also given. The 3 best systems, as far as matching human correlations, are the simple TFIDF system, the same system with a usage model, and the generic SVM classifier (SVM TFIDF) without a usage model, *all with the us-them vocabulary*. Clearly most of the work at capturing human militancy judgments is being done by the choice of vocabulary combined with the simplest TFIDF weighting scheme.

Table 3 shows the classifier systems for the full vocab, which provided far and away the best featureset for classificationn. The best classifier was the SVM TFIDF model, more or less the generic SVM. As classifiers, the SVMs always outperformed the corresponding simple weighted systems.

The worst classifier is the Sim system (Acc: 53.48); simply using the syntactic model scores yields close to random performance. The same model augmented with TFIDF scores (Sim + TFIDF) has much-improved accuracy (93.00), and outperforms TFIDF, the weighting system without Sim scores (89.13). This shows the syntax is contributing some information.

6 Discussion and Related Work

The most significant finding is that features hand-selected for their use in marking group membership, weighted only by their TFIDFs, made for the best ranking system, significantly outperforming another small feature set selected by Mutual Information. This provides strong evidence that the us-them analysis is turning up something significant. Admittedly the data set is small and we have yet to show that this set of features will extend robustly to ranking more diversified sets of documents, but we suspect that this feature set is a good seed. The importance of such *feature-engineering* is well-known for a variety of applications, for example, in building good classifiers for spam detection and email filtering.

The other significant finding here was the clear separation between what makes a good classifier and what makes a good ranking system. The full vocab SVM TFIDF system was the best classifier; but the simplest possible weighted system (TFIDF only) with the hand-selected vocabulary made the best ranking system. Restricting the SVM to the hand-selected vocabulary made it much better than its full vocabulary cousin (-0.13 → 0.20), but still not as good as the weighted system with that vocabulary. The fact that TFIDF weighting played such a significnant role in the best ranker parallels the finding in Paltoglou and Thelwall (Paltoglou and Thelwall 2010) that various TFIDF weighting schemes gave better performance than binary features in sentiment analysis.

	Accuracy	Precision	Recall
Sim	53.48	33.68	90.28
TFIDF	89.13	90.28	90.28
Sim + TFIDF	93.00	95.65	91.67
SVM Sim	89.42	93.10	75.00
SVM TFIDF	95.43	92.00	95.83
SVM Sim + TFIDF	90.88	100	83.33

Table 3: Systems used as classifiers

7 Acknowledgments

This research was supported by NSF CDI grant # 1028177.

References

Anderson, B. 2003. *Imagined Communities: Reflections on the Origin and Spread of Nationalism.* New York: Verso.

Chen, H. 2007. Exploring extremism and terrorism on the web: the dark web project. In Yang, C. C.; Zeng, D.; Chaur, M.; Chang, K.; Yang, Q.; Cheng, Z.; Wang, Jue annd Wang, F.-Y.; and Chen, H., eds., *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2007*. Berlin: Springer. 1–20.

Gawron, J. M. and Stevens, Kellen. 2012. Group membership. http://www-rohan.sdsu.edu/˜gawron/group˙identity.pdf.

Gupta, D. 2008. *Understanding Terrorism and Political Violence: The Life Cycle of Birth, Growth, Transformation, and Demise*. New York: Routledge.

Malouf, R., and Mullen, T. 2007. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW)*.

Paltoglou, G., and Thelwall, M. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1386–1395. Association for Computational Linguistics.

Prabowo, R., and Thelwall, M. 2009. Sentiment analysis: A combined approach. *Journal of Infometrics* 3(1):143–157.

Tajfel, H. 1978. *Differentiation between Social Groups: Studies in Inter-Group Relationship*. London: Academic Press.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL-02*, 417–424.