Analyzing Sentiment Markers Describing Radical and Counter-Radical Elements in Online News

Hasan Davulcu, Syed Toufeeq Ahmed, Sedat Gokalp
Dept. of Computer Science and Engineering
Arizona State University, Tempe, AZ
hdavulcu@asu.edu, toufeeq@asu.edu, sgokalp@asu.edu

M'hamed H Temkit, Tom Taylor
Department of Mathematics
Arizona State University, Tempe, AZ
htemkit@mathpost.asu.edu, tom.taylor@asu.edu

Mark Woodward
Department of Religious Studies and CSRC
Arizona State University, Tempe, AZ
mataram@asu.edu

Ali Amin
Center for Religious and Cross Cultural Studies
Gadjah Mada University, Yogyakarta, Indonesia
aleejtr77@yahoo.com

Abstract— **In this study, we aim to obtain "natural groupings" of 151 local non-government organizations and institutions mentioned in a news archive of 77,000 articles spanning a decade (May 1999 to Jan 2010) from Indonesia. One of our goals is to enhance our understanding of counter-radical movements in critical locations in the Muslim world. We present information extraction techniques to recognize entities, and their beliefs and practices in text as a step towards identifying socially significant scales with explanatory power. Then, we proceed to cluster organizations based on these scales. We present experimental results, and discuss challenges in reasoning with the complex interactions of many simultaneous beliefs, practices and attitudes held by the leaders and followers of various organizations.**

Keywords-component; Web information extraction, markers, spectral clustering, scales, hierarchical clustering, organizations.

I. INTRODUCTION

Many social network analysis [22] tools depend on a single kind of affinity relationship, such as friendship, kinship, or warfare among its actors. However, sociologists [23] assume that, until proven otherwise, actors' alliances are shaped by complex interaction of many simultaneous beliefs, practices and attitudes. In this study, we attempt to obtain "natural groupings" of 151 non-government local organizations and institutions mentioned in a news archive from Indonesia spanning a decade (May 1999 to Jan 2010). One of our goals is to enhance our understanding of counter-radical movements in critical locations in the Muslim world. By design our study leaves the meaning of moderate or counter-radical open-ended—except for a baseline understanding that moderates reject violence as a means to political or social objectives. In consultation with social scientists on our team, first we identified a preliminary collection of 357 social markers as significant attributes of an organization's reported beliefs and activities. These markers correspond to most frequently occurring terror and violence related keywords (such as bombings, kidnappings, and other violent crimes), keywords related to social, legal, political activities (such as call for reforms, protests, criminal and corruption related activities) and lists of religious sects, beliefs and practices. The information extraction methods for identifying organizations and sentiments and practices their leaders and followers are presented in Section 3. Section 2 presents related work. Section 4 presents our efforts for dimensionality reduction of attributes in order to (i) overcome the sparsity problem encountered while text processing, and (ii) group markers into meaningful and socially significant categories with some explanatory power. In Sections 4 and 5, we present various similarity metrics and clustering techniques we considered and, analysis of the clustering results. Social scientists on our team observed that only four out of the eight clusters we identified correspond to pure groups of radical or counter-radical organizations. Pure radical clusters were easily identified due to high similarity among their violent practices. Similarly, pure moderate clusters were identified due their strong reactionary opposition to violent practices through protests and rhetoric. We had four other mixed clusters. Upon analyzing the profiles of the organizations found within these mixed (radical and moderate) clusters, we identified that attempting a binary labeling as radical or counter-radical as a measure of cluster purity do not capture the overlap, movement and interactivity among these organizations. Specifically, counter-radicalism can manifest in many forms; very few are secular or pro-Western, most support flexible interpretations of Islamic texts, most support religious tolerance, most support the empowerment of women, most consider democracy to be a permissible form of government and most reject religious practices that are strongly associated with radical Islam. However, it is difficult to categorize many organizations due to differences of opinion and practices among their various factions and due to the utilization of non-violent, but still radicalizing (non-secular, repressive, and non-tolerant) practices and rhetoric of some others.

II. RELATED WORK

Large-scale information and sentiment extraction from web is an active research area. Web information extraction systems such as KnowItAll [1] extract facts and relationships from text using set of extraction rules for each class and relation based on a set of generic, domain independent templates. Pasca et al. [2] utilized generalized contextual extraction patterns to extract large amount of facts (of type person-born-in-year) from web. Turney [3] used large number of co-occurrences of entities on

web as a measure to recognize synonyms. To understand and extract meaningful structure from document streams as they arrive continuously over time, [4] presented an approach to model "burst of activity" in a document stream. Seminal work [5], [6] in event detection and tracking explored clustering algorithms like agglomerative clustering augmented Group Average Clustering. Well-known idf-weighted cosine coefficient metric method [7] was also used to detect and track topics, [8] used both text content and date information of news articles. A real-time news event extraction system [9] extracts violence and disaster events by processing the news article using extraction grammars on each document. By first extracting signature of the event, [11] tracked and attempted to link related events with mentions of same event signature in other incoming news articles, thereby forming a thread that links all the news articles referring to specific events.

III. Extracting Sentiment Features from News

Wealth of information locked in online daily news media makes it a valuable resource to mine associations, and competing sentiments and rhetoric about any reported entity or topic. Web text mining presents a useful tool to mine and analyze these associations at a large scale and quantify them for further study or observations. In our task of mining sentiment on social markers, we implemented a robust and scalable web text mining pipeline illustrated in Figure 1. Steps of this pipeline are explained below.

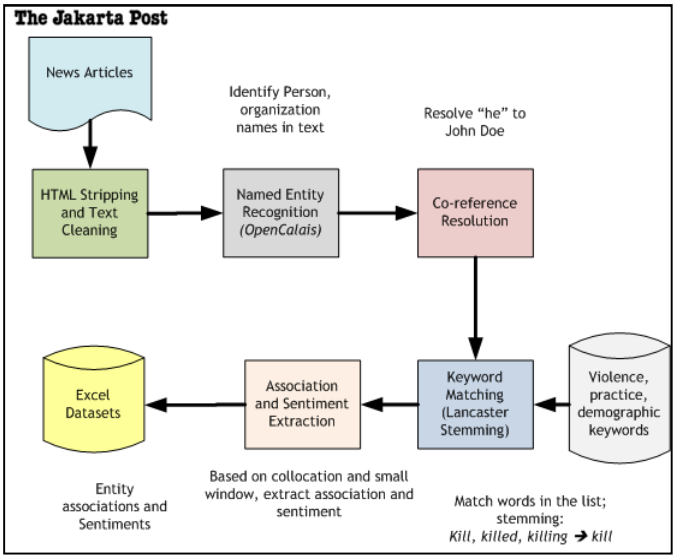


Figure1. Entity-marker association and sentiment extraction pipeline.

A. Data Collection and pre-processing

We crawled archives of the premier English online news source in Indonesia, *the Jakarta Post*[1] from May 1999 to Jan 2010, and collected 77,000 articles listed under the following categories picked by our social scientists: National, Jakarta, Bali, Archipelago, Islam, Opinion, and Violence. These articles were converted to plain text by stripping HTML tags.

[1] The Jakarta Post: www.thejakartapost.com

B. Recognizing entities and practices in text

Before we can extract entity-marker association information and sentiment (i.e. whether an organization is for or against a certain practice or belief), we need to first identity named entities in unstructured text. Named Entity Recognition (NER)[10] is a well researched area, and tools using latest machine learning algorithms like Conditional Random Fields have shown high accuracy in identifying named entities in text. Typically in news articles, entities of interest are dates, people, places, and organizations. For our task, we choose Open-Calais [2] web API, one of the state-of-the-art NER systems. Let us consider an example sentence[3]:

Example sentence *S1*:

> *National Police chief Gen. Bambang Hendarso Danuri on Tuesday said all Aceh-based fugitives were on the police's wanted list, for their possible involvement in several bombings in Jakarta and Bali.*

The sentence after Named Entity Recognition:

> *National Police chief <PERSON>**Gen. Bambang Hendarso Danuri*** *</PERSON> on Tuesday said all Aceh-based fugitives were on the police's wanted list, for their possible involvement in several bombings in <LOCATION>**Jakarta**</LOCATION> and<LOCATION> **Bali**</LOCATION>.*

Tags <PERSON> marks the beginning of the entity name and </PERSON> marks the end of the entity name. In this study we focused on three types of entity names <PERSON>, <LOCATION> and <ORGANIZATION>.

OpenCalais also provides co-reference resolution. Resolving pronouns and anaphora in an article is crucial, since after the first occurrence of a proper noun *(say "John Doe"),* the entity name is usually referred to by its pronouns *("He")* or anaphora *("Mr. Doe").* For the above example, if we follow few sentences in the same article we find a sentence where a pronoun *"his"* and a first name is used to refer to the person entity *"Gen. Bambang Hendarso Danuri",* as shown below:

Example sentence *S2*:

> *"Some of [the terrorist suspects] were also involved in the Indonesia Stock Exchange [BEI] and Bali bombings," **Bambang** said during **his** visit to Aceh Besar.*

After co-reference resolution:

> *"Some of [the terrorist suspects] were also involved in the Indonesia Stock Exchange [BEI] and Bali bombings," **Gen. Bambang Hendarso Danuri** said during **Gen. Bambang Hendarso Danuri** visit to Aceh Besar.*

[2] OpenCalais: www.opencalais.com

[3] Example article: www.thejakartapost.com/news/2010/03/16/aceh-terrorists-involved-hotel-bombings-says-national-police.html

This co-resolution step makes it possible to extract further associations from all the sentences and their clauses.

C. Extracting (collocation) associations around the entities

For every entity name (and its co-referent) we extracted their associated markers corresponding to various beliefs and practices. Using a windowing technique of size of seven words to the left and right side of the entity name, close occurrence information is used to establish connections between named entities and markers. If there is a connection, then we associate the marker with the closest entity. Alongside <PERSON>, <LOCATION> and <ORGANIZATION> types, we also created a list of keywords for <DEMOGRAPHIC> groups, such as students, farmers, workers, police forces etc. The list for demographic groups was created starting with a seed list of keywords provided by social scientists on our team and expanding the list through simple text patterns and relaxation labeling techniques [24] suitable for entity identification in text. Our final list for demographics currently includes 1,528 phrases. Let us consider the example sentence S1 again with tagged markers and demographics:

Sentence *S1*:

> *National Police chief Gen. Bambang Hendarso Danuri on Tuesday said all Aceh-based <DEMOGRAPHICS>**fugitives**</DEMOGRAPHICS> were on the<DEMOGRAPHICS>**police's** </DEMOGRAPHICS> wanted list, for their possible involvement in several <MARKER>**bombings** </MARKER> in Jakarta and Bali.*

An example sentence showing a practice:

> *A recent survey also found <MARKER> **polygamy**</MARKER> was a significant factor behind the country's rising divorce rate.*

D. Sentiment Extraction

We also extracted sentiment toward a belief or practice by the associated named entity. Starting with two seed lists of hand-build lexicon for support and dissent indicating phrases, we expanded our lists by finding their synonyms/antonyms in WordNet [4] and recursively adding them into corresponding categories. Our final lists of support/dissent indicators include 2,457 and 2,735 keywords correspondingly. We extracted sentiment polarity either as support, dissent or no data based on the co-occurrence of a sentiment phrase within close proximity of an entity-marker association. Consider the following sentence:

> "The <*DEMOGRAPHICS*> **residents** </*DEMOGRAPHICS*>, who <DISSENT> **opposed** </DISSENT> the <MARKER>**eviction plan**</MARKER>, said it would be an insult for Muslims who see the place as sacred as it used to be the grave of the …."

[4] WordNet: wordnet.princeton.edu

IV. CLUSTERING ORGANIZATIONS USING MARKERS

The main objectives of this paper are twofold: (i) to reduce the large number of markers corresponding to various beliefs and behaviors by constructing socially significant scales with explanatory power, and (ii) utilizing these scales for clustering organizations to identify their common traits and groupings.

A. Methods

The final dataset or Excel spreadsheet was comprised of 151 organizations by 357 markers. The columns were comprised of two categorical variables, the name of the organizations and a preliminary labeling of the organizations by the social scientists as either primarily radical(R) or counter-radical(C). According to this labeling there are 67 radicals and 84 counter radical organizations in our dataset. The 357 markers for each organization were filled with integer values based on a simple formula assessing the difference between the number of supportive associations between an organization and a marker and the number of opposing associations. This formula yielded with 0's to denote no recorded associations, positive values indicating support and negative values indicating opposition. We observed that the eventual 151 by 357 matrix between organizations and markers had a high frequency of 0's, which became a challenge during the analysis. Next we discuss dimensionality reduction through clustering of the markers to obtain new data scales and subsequently clustering of the organizations based on these scales.

B. Spectral clustering of the markers

Spectral clustering [16] has become one of the most modern clustering algorithms and has very often outperformed traditional clustering algorithms such as K-means [21]. Spectral clustering is based on similarity graphs, the graph Laplacians and K-means algorithm to cluster the observations or points represented in the new space formed by the specified number of eigenvalues obtained from the graph Laplacian derived from the graph similarity. Spectral clustering is very well suited when the data is sparse and suitable for detecting clusters in non-convex regions. After using spectral clustering [17] on the 357 markers and looking at their plot of within-cluster-sum-of-squares versus the potential number of predicted clusters we observed that 20 clusters form the best number of groups of markers. We further investigated the quality of these 20 groups of markers by using graphical methods such as scatter plots, their correlation to social scientists' classifications in the linear sense (Pearson[13]) and in the general sense according to (Spearman[14]). Finally we also used the Cronbach alpha coefficient [12] to assess the internal consistency of these groups as scales. As a rule of thumb a Cronbach alpha coefficient >= 0.70 is indication of a good scale. Out of 20 clusters we had 11 good scales according to Cronbach alpha coefficient. Next we shared these 20 groups of markers with social scientists. Social scientists were able to organize 19 clusters containing 151 markers into 9 core clusters with very few (only 7) edits. We also had one big group of markers with 206 entries which we did not have many matches in our corpus and they were aggregated together. The nine categories identified by social scientists

were related to: religiosity/piety, politics, war, corruption, terror, violent crimes, protest and moderate attitudes. We named these nine clusters of markers as our golden scales. Figure 2 below shows a sample of the organization of the 20 marker clusters (integers corresponding to cluster id's of each marker) into 9 scales (capitalized) and the suggested edits (red keywords highlighted in yellow) by domain experts.

	WAR		TERROR		PROTEST		VIO. CRIME
8	conflict	19	bomb	9	clashing	11	assaulted
8	fighting	19	bombing	9	contention	11	hit
8	firing	19	corps	9	riots	11	crime
8	damaging	19	explore	9	hostage	11	killing
8	rockets	19	hatred	9	eruption	11	vicious
8	battle	19	hiding	9	turmoil	11	torture
8	ceasefire	19	threat	9	opposition	11	victimize
8	bombardment	19	extremists	9	block	11	embezzle
	0.93		0.76		0.92		0.66
	CORRUPTION		RADICAL		MODERATES		RELIGIOUS
15	abuse	13	struggle	4	seminar	5	halal
11	embezzle	13	justice	4	peaceful	5	prayers
15	pillage	13	jihad		0.008	5	quran
15	charge	13	separatists	6	charity	5	ramadan
15	misuse		0.4	6	pardon	5	sunnah
15	grave	10	training	6	equality	5	tabligh
15	offense	10	militant	6	consensus	5	tahir
15	bribe	10	combat	13	justice	18	fasting
	0.63	10	death	6	permitted	14	fatwa
		10	terrorizing	6	pluralism		0.73
			0.61		0.58		

Figure2. Golden Scale organization of markers suggested by domain experts

This process reduced the dimensionality of the data and aggregating the corresponding marker scores into golden scale scores by adding them up, rendered the data less sparse by deflating the rate of 0 scores.

C. *Hierarchical Clustering and clustering the organizations*

To meet our second objective, we tried clustering the organizations through several metric distances and assessed the quality of the resulting clustering by checking their purity against the a priori labeling provided by the domain experts. Purity [26] measures how pure is the cluster *i* by weighting each cluster proportional to its size and according to the numbers of different organizations in it belonging to different categories. In our clustering analyses we tried several metrics based on the original data and recoded data to ordinal measurement {-(negative sentiment), 0(neutral sentiment), +(positive sentiment)} which was one of the ways to address the discrepancy in ranges between the scales, as well as the use of normalizing techniques on the original scores. The clustering analyses used the package Hclust in R[19], which is based on hierarchical clustering based on several choices of distances such as Hamming, Euclidean and Minkowski and several clustering methods such as single, complete or ward[20]. The clustering results using Minkowski distance with p=1 and ward method using the ordinal data fared better than other distances and methods. Hamming distance did not take into account the ordinal nature of the data and Euclidean distance was more sensitive to outliers.

TABLE I. A RE-CODED SUBSET OF THE DATA WITH GOLDEN SCALES

ORGS	CATEGORY	war	terror	protest	vio_crime	corruption	politics	radical	moderates	religious
Org 1	R	+	0	0	+	0	0	0	0	0
Org 2	C	0	0	0	+	0	0	0	0	0
Org 3	R	+	0	0	0	0	0	0	0	0
Org 4	C	0	0	0	+	0	0	0	+	0
Org 5	C	0	0	0	+	0	-	0	0	0
Org 6	R	+	0	0	+	0	0	0	0	0

D. *Results*

In Figure 4, we provide the dendogram resulting from the hierarchical clustering of the categorical coded scale, using minowski distance with p=1 and the ward method. In the leaf nodes of the dendogram we notice the presence of several runs of C and R type organizations. To determine the optimal number of clusters we used the within cluster sum of squares criterion or, cluster ratio of within sum of squares to between sum of squares and looked at its plot against the potential number of clusters in Figure 3 below. Upon analyzing Figure 3, we can see that the within and between sum of squares ratio has a significant drop at 8 clusters.

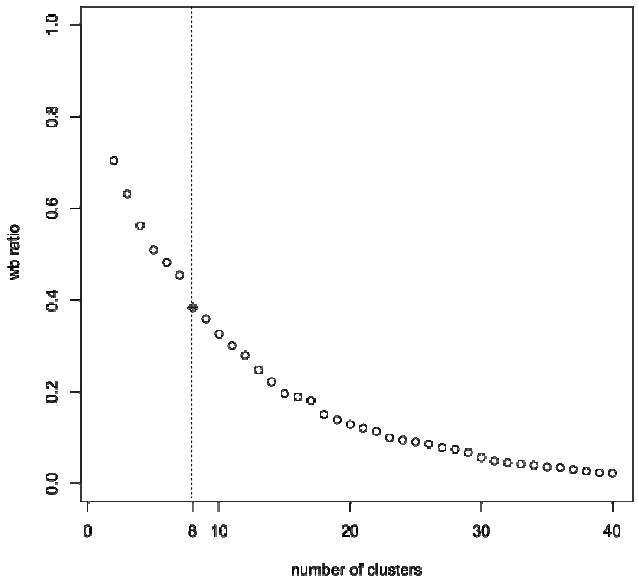
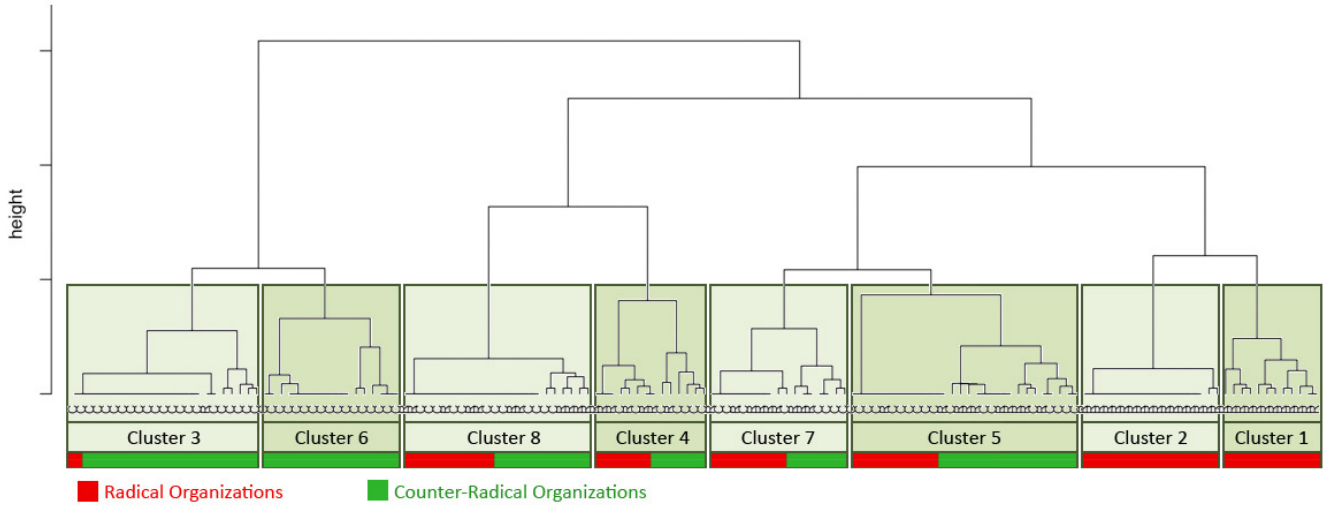


Figure 3. Within/Between Ratio vs. Number of Clusters

We provide the purity of the clustering based on the 8 clusters identified in the form of a confusion matrix. This is presented in the contingency table in Figure 4 by counting the numbers of R and C type organizations found within each cluster.



		Cluster 3	Cluster 6	Cluster 8	Cluster 4	Cluster 7	Cluster 5	Cluster 2	Cluster 1
Dist	Radicals	1	0	11	7	9	10	**17**	**12**
	Counter-Radicals	**22**	**17**	12	7	8	18	0	0
Golden Scales	War	- **0** +	- **0** +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +
	Terror	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- **0** +	- 0 +
	Protest	- 0 +	- 0 +	- 0 +	- 0 +	- **0** +	- **0** +	- 0 +	- 0 +
	Violence/Crime	- **0** +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +
	Corruption	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +
	Politics	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +
	Radical	- 0 +	- 0 +	- **0** +	- 0 +	- **0** +	- **0** +	- 0 **+**	- 0 **+**
	Moderate	- 0 +	- **0** +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +
	Religious	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +	- 0 +

Figure 4. Dendogram of the hierarchical clustering of organizations

V. Experimental Analysis

In this section, we will explain the characteristics of each cluster by analyzing the common traits of the organizations found with a cluster, and also by analyzing the discriminatory scales that would improve the purity of each cluster based on the polarity of the sentiments {-(opposition), 0 (zero) or +(support)} found within each cluster.

The table in Figure 4 highlights the discriminatory features of the clusters. Bold polarities correspond to sentiments shared by many members of a cluster. For example, the organizations in Cluster 3 all share a negative sentiment on war related markers.

Here, we hope to shed more light on the purity and its lack thereof within the eight clusters identified.

1) *Cluster 1: Radical*

a) *Quantitative Characteristics*

Organizations in *Cluster 1* have supportive sentiment on war, mixed sentiments on terror related activities and they support radical beliefs and practices.

b) *Semantic Characteristics*

This cluster contains extremely radical and rebellious organizations with military capabilities including some guerilla fighters that are listed in the US terrorist blacklist.

2) *Cluster 2:Radical*

a) *Quantitative Characteristics*

Organizations in *Cluster 2* support war/fighting and they support radical markers.

b) *Semantic Characteristics*

This cluster has extremist militant and political forces.

3) *Cluster 3:Counter-Radical*

a) *Quantitative Characteristics*

Organizations in *Cluster 3* have negative sentiment on war, no sentiment on violence/crime scale.

b) *Semantic Characteristics*

This cluster hosts peaceful counter-radical organizations addressing social issues.

4) *Cluster 4: Mixed*

a) Quantitative Characteristics

Mixed

b) Semantic Characteristics

Mostly political organizations

5) *Cluster 5: Mixed*

a) Quantitative Characteristics

Organizations in *Cluster 5* share an opposing sentiment on terror and radical scales.

b) Semantic Characteristics

This cluster has less active and/or localized smaller groups of organizations of intellectual and elite classes.

6) *Cluster 6: Counter-Radical*

a) Quantitative Characteristics

Organizations in *Cluster 6* have negative sentiment on war, opposing sentiments against violent crimes.

b) Semantic Characteristics

This cluster contains counter-radicals with more conservative views and a social issues agenda with emphasis on human rights.

7) *Cluster 7: Mixed*

a) Quantitative Characteristics

Organizations in *Cluster 7* have positive sentiment on war, non-positive (opposing) sentiment on protest and radical scales.

b) Semantic Characteristics

This cluster has organizations that are taking action against terror, violence and corruption. It also has groups that are promoting tolerance in Islam as well as some radical organizations.

8) *Cluster 8:Mixed*

a) Quantitative Characteristics

Organizations in *Cluster 8* have are pro-war, pro-protest tendencies and mixed sentiments on radical scales.

b) Semantic Characteristics

This cluster contains student organizations or organizations with youth concentration taking action for social rights with some mixed militarist tendencies.

VI. Conclusions and Future Work

In this paper we present the results of information extraction and data clustering techniques to obtain "natural groupings" of 151 local non-government organizations and their beliefs and practices identified in a news archive of 77,000 articles spanning a decade (May 1999 to Jan 2010) from Indonesia. The study results indicate both challenges and opportunities. Our information extraction algorithms mostly rely on collocation statistics and hence they are language independent for extracting organization-marker associations and support/opposition sentiment polarities. This enables us to expand our corpus beyond English sources to include Indonesian online news sources. An expanded corpus might enable us to fill the sentiment values for organization-marker associations with more robust values. Another opportunity is to utilize the relevant markers identified with the assistance of social scientists on our team to train classifiers for automatically detecting and extracting additional markers. Additional markers might in turn enable our clustering algorithms to yield more nuanced scales and clustering results with better purity. Our study also points to the critical role that social scientists play in guiding and gauging the efforts of computer scientists in social computing investigations. Most of the computational steps presented in this paper, such as the generation of seed lists for various information extraction tasks, and generation of scales for clustering and accessing the quality of the clustering results, were supervised or semi-supervised based on the input from social scientists. Social scientists have great insights for reasoning with the challenging dynamics of the social organizations involving complex interaction of many evolving beliefs and practices.

References

[1] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. 2004. Web-scale information extraction in knowitall: (preliminary results). In Proceedings of the 13th international Conference on World Wide Web, 2004. ACM, New York, NY, 100-110.

[2] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. In Proc. of AAAI-2006, 2006.

[3] Turney, P. D. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning* (September 05 - 07, 2001).

[4] J. Kleinberg, "Bursty and hierarchical structure in streams," Data Mining and Knowledge Discovery, vol. 7, no. 4, pp. 373–397, 2003.

[5] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in Proceedings of the DARPA broadcast news transcription and understanding workshop, vol.1998. Citeseer, 1998.

[6] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-lineevent detection," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA, 1998, pp. 28–36.

[7] J. Schultz and M. Liberman, "Topic detection and tracking using idfweighted cosine coefficient," in Broadcast News Workshop'99 Proceedings. Morgan Kaufmann, 1999, p. 189.

[8] Z. Li, B. Wang, M. Li, and W. Ma, "A probabilistic model for retrospective news event detection," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM New York, NY, USA, 2005, pp. 106–113.

[9] H. Tanev, J. Piskorski, and M. Atkinson, "Real-Time News Event Extraction for Global Crisis Monitoring," in Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008), London, UK. Springer, 2008, pp.

[10] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling." In ACL. The Association for Computer Linguistics, 2005.

[11] S. T. Ahmed, R. Bhindwale, and H. Davulcu, "Tracking Terrorism News Threads by Extracting Event Signatures ," in Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI-2009). Dallas, Texas, USA, 2009

[12] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika. 16, 297-334

[13] Fisher, R.A. (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population". Biometrika 10 (4): 507–521.

[14] Hollander, D.A. Wolfe, "Nonparametric statistical methods" , Wiley (1973)

[15] 4- Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905, August 2000

[16] Andrew Y. Ng, Michael I. Jordan, Yair Weiss, "On Spectral Clustering: Analysis and an Algorithm", Neural Information Processing Symposium 2001

[17] http://bm2.genes.nig.ac.jp/RGM2/R_current/library/kernlab/man/specc.html

[18] Murtagh, F. (1985). "Multidimensional Clustering Algorithms", in COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used).

[19] http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html

[20] Ward, J. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58:236-244.

[21] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.

[22] Wasserman, S. and Faust, K. (1994) Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press.

[23] Hanneman R. A., and Riddle M., Introduction to Social Network Methods. Online textbook.

[24] Shen, W., Li, X., and Doan, A. Constraint-Based Entity Matching. Proceedings of the National Conference on Artificial Intelligence (AAAI) (2005)

[25] M. Rosell, V. Kann, and J.-E. Litton. Comparing comparisons: Document clustering evaluation using two manual classifications. In ICON, 2004.