Online News Media Bias Analysis using an LDA-NLP Approach

Article · May 2012

CITATIONS
2

READS
607

2 authors:

Sarjoun Doumit
University of Cincinnati
15 PUBLICATIONS **128** CITATIONS

SEE PROFILE

Ali A. Minai
University of Cincinnati
144 PUBLICATIONS **1,989** CITATIONS

SEE PROFILE

Online News Media Bias Analysis using an LDA-NLP Approach

Sarjoun Doumit and Ali Minai
School of Electronic & Computing Systems, College of Engineering, University of Cincinnati, Ohio 45221-0030, U.S.A. (email: doumitss@mail.uc.edu & ali.minai@uc.edu).

It is widely recognized that every media outlet has its own "spin" on news, and this bias has been described in many ways and at many levels. In political news for example, the bias can be liberal, conservative, moderate, corporate, etc. In addition, recent research has focused on the 'sentiment dimension' to further identify and categorize news bias. This is achieved through analysis of the adjective and adverb terms found in the news texts. The accuracy and generality of these models depend on the evaluation methods used to appraise the intensity and emotional weights of the adjectives and adverbs, thus rendering the results open to controversy. In this paper we propose a unifying system to extract information from political news texts and analyze it within a cognitive network. We view the different news media sources as agents with unique personalities, which we assume are latent within their texts. We use a combination Latent Dirichlet Allocation (LDA) and natural language processing (NLP) methods to identify the different agents' personality traits with respect to various topics or concepts. An agent's personality traits affect its inclination to word a certain event in a specific way. Using the common concepts stored in the cognitive network, our system can compare the different agents on a unified and normalized platform.

1 Introduction

Consider a finite collection of news sources or media outlets that provides news coverage to everyone at all times. We can represent these news sources as agents in a large environment which is composed of semantic constructs such as ideas.

The agent news sources or media outlets use a *stimulant* such as a factual event, to construct and generate its interpretation in the form of new ideas (i.e., news stories), using existing ideas as building blocks. Together, all this defines an *ecology* of sources and ideas, and just as in natural ecosystems, the agents and phenomena in this ecology can be identified, characterized, classified and connected using the tools of mathematical, statistical and computational analysis. In this study, we consider the specific problem of identifying ideas generated by a set of news sources and using them to characterize and classify these sources based on the clustering of their generated stories in a semantic space.

This is especially important for political news analysts and policy makers who try to understand and track the sentiment or *bias* found in news stories. There are many metrics one could propose to measure this bias, which is latent but yet palpable in each media outlet. Defining what bias actually means when dealing with news of a political nature is a critical step for quantitatifying it. Many statistical tools exist for analyzing and categorizing natural texts such as news reports, but so far these tools give little additional insight into deeper subtexts. This is due to the diversity and the complex *semantic* nature of these texts. Also when using statistics, one has to be careful to take into account data skews, because some media outlet are much more active and produce more news than others, and therefore can saturate the semantic environment with their own preferred terms.

Even more importantly, coming up with a unifying and equitable system that measures the bias across different media outlets on an equal footing is critical to obtaining any useful results. This can be done - albeit with difficulty - by using complex natural language processing techniques. The use of NLPs allows for the analysis of adjectives and adverbs which are usually associated with *sentimental bias*, in conjunction with specific keywords related to specific events. Such approaches using adjective and adverb-based sentiments are still subject to the weaknesses mentioned earlier because of the great number and diversity of the political news, especially since in these approaches, the sentiment factors are weighted by a group of experts and are not *emergent* from the system.

We applied *latent Dirichlet allocation* (LDA) [2], a probabilistic topic-modeling tool, to extract latent topics from newsfeed data. In LDA, each document (news item) is considered to be generated from a distribution of topics, where a topic is a distribution over words. We also employed Antelope[15], a natural-language processing tool to analyze the documents' semantic structure in combination with LDA. Our approach, applying LDA and Antelope together to a semantic framework that incorporates sentiment, allows us to achieve relevant insights that neither system can achieve on its own. The aim of this study is to test the model for establishing media-outlet personality signatures based on the semantic structure of its news articles, which offers its user a more robust framework for comparison and analysis

The rest of this paper is organized as follows: Section 2 reviews similar systems. Section 3 gives an overview of the LDA model, followed by a discussion in section 4 of our proposed model. In section 5 the simulations and results are

presented and finally in section 6 the conclusions are presented.

2 Background

There exist many research and commercial systems today that analyze and cluster textual news employing methods that range from the purely statistical to graphical models. It is up to the news analyst or user of the system to organize the output according to his or her own specific needs to benefit from the result. For example, WEIS[10, 16] and CAMEO[5] are both systems that use *event analysis*, i.e. they rely on expert-generated dictionaries of terms with associated weights, and parse the text to match the words from the news event to those in the dictionary. They can then categorize the information again into a set of expert-defined categories with respect to sentiment intensity values. Other systems, such as Oasys2.0 [3] use another construct called *opinion analysis*, which depends on user feedback rather than on experts in order to determine the intensity value of an opinion. The Oasys2.0 approach is based on aggregation of individual positive and negative references identified [1]. Yet other systems, RecordedFuture [4] and Palantir [14], rely on experts and have at hand massive amounts of data, with inference and analysis tools that uses data correlation techniques to produce results in response to specific keywords in user queries. More recently, topic chain modeling [7, 13, 8] has been suggested to track topics across time using a similarity metric based on LDA to identify the general topics and short-term issues in the news. It is important to notice that all the methods mentioned above except topic chain models adopt query-driven approaches to produce results.

3 The Latent Dirichlet Allocation Model

There has been great interest in Latent Dirichlet Allocation or LDA ever since the publication of the seminal paper by Blei, Ng and Jordan[2]. It is a machine learning technique (shown in extended version in Fig.1), that evolved from a previous model called *Probabilistic Latent Semantic Analysis* [6] (pLSA) for reducing the dimensionality of a certain textual corpus while preserving its inherent statistical characteristics. LDA assumes that each document in a corpus can be described as a mixture of multiple latent topics which, in turn, are distributions over words found in the documents of the corpus. LDA assumes that documents are made of a list words where the order of the words is not important i.e. a bag-Of-words approach. LDA is a generative model in that it can generate a document from a set of topics, but it can also be used as an inference tool to extract topics from a corpus of documents. To generate a corpus of D documents, where each document has N_d words, and for a total of T topics, LDA's generative algorithm is:

1. Pick a document size N_d

2. Pick a set of topics $\theta \sim Dirichlet(\alpha)$
3. For each of the N_d words w_n found in document d:
   (a) Draw a topic $t_{w_n} \sim Multinomial(\theta)$
   (b) Draw a word $w_n \sim Multinomial(\phi_{t_{w_n}})$, where $\phi \sim Dirichlet(\beta)$.

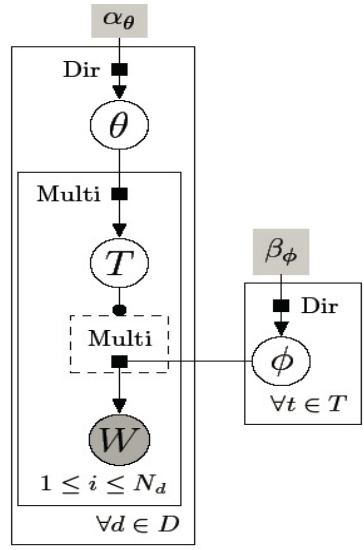

Figure 1: Directed factor graph for LDA. ϕ is the words-topic distribution, θ represents the topics-document distribution, α is the *Dirichlet* hyperparameter for θ, and finally β is the *Dirichlet* hyperparameter for ϕ

The probability equation is the following:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{t=1}^{T} P(\phi_t; \beta) \times \prod_{d=1}^{D} P(\theta_d; \alpha)... \\ \times \prod_{w=1}^{N_d} P(T_{d,w}|\theta_d) P(W_{d,w}|\phi_{T_{d,w}}) \quad (1)$$

The probability $P(T_{d,w}|\theta_d)$ is the probability of drawing topic T for word w from document d, given that that topic's distribution for that specific document d is θ_d. On the other hand, $P(W_{d,w}|\phi_{T_{d,w}})$ is the probability of drawing word W for the wth word from document d assuming it is drawn from the distribution for topic $T_{d,w}$.

4 The Proposed Model

We start by defining how we view the process of political news story generation and how bias gets embedded into the very fabric of the original factual news. Let E represent a factual event, and $\bar{E}$ the smallest set of factual clauses, c_k, that can summarize event E, where $k = 1, ..., K^c$. A factual clause contains the basic semantic elements to describe the event or parts of the event without any use of adjectives or adverbs which carry sentiment, so $\bar{E}$ represents the most unbiased text report of the event. Let $\mathbb{O}$ represent the existing media outlets as a population of agents such that $\mathbb{O} = \bigcup_1^n O_i$, where O_i is an identifiable media outlet with identity i and n is the total number of media outlets in the environment. Each O_i is characterized by a set of *subjects* $\mathbb{S}_i$, where each subject S_{ij} has a *bias* B_{ij} represented as a set L_{ij} of K_{ij}^d dependent clauses d_k, which could be either an adjective clause or an adverb clause. Thus, each dependent clause carries a sentiment, and the pattern of sentiment over L_{ij} defines agent i's bias towards subject j. We assume that every agent has a bias for every subject in its repertoire, though agents may vary in their subjects.

When reporters who work for a specific media outlet O_i come across a factual event E, they first determine what subject S_{ij} this event is relevant to. Then, armed with the specific bias B_{ij} for S_{ij}, they initiate the process of creating a news story using the bias in conjunction with the unbiased description $\bar{E}$ to create a biased version of E. This is denoted by E^i, i.e., source i's (biased) report of event E. The average news reader receives this biased story as the final product. Our aim is to analyze the biased news story and try to isolate the biases or the L_{ij} which includes the adjectives and adverbs in order to compare them on a deeper semantic/cognitive level instead of making a purely lexical-weighted comparison.

We visualize the process of *weaving* the bias into the factual event as a combination of three sub-processes which we call 'Actor-assignment', 'Action-assignment' and 'Sentiment-assignment'. These processes work together as single coordinated dynamical units. We call these *meme-synergies* (Fig.2) in analogy with the muscle synergies that underlie motor control in animals. Muscle synergies are coordinated musculoskeletal degrees of freedom that are jointly controlled as a unit, and form the primitives from which more complex movement is constructed. Just as muscle synergies combine to move one's arm to hit a tennis ball in a particular style, meme-synergies combine effortlessly to generate coherent textual news products with a certain bent. In our model, we consider a media's biased clauses or *embedded memes* to be preconfigured to trigger or produce text-generation in accordance with those of the original event E, especially amongst those that share a similar construct such as a concept or sentiment. This correspondence of thought and action represents a deeply embodied view of human cognition [11].

In the model, every synergy combines *Actors* represented by proper nouns and other nouns (concepts and features), *Actions* represented by verbs, and *Sentiments* represented by adjectives and adverbs. For example, the proper noun

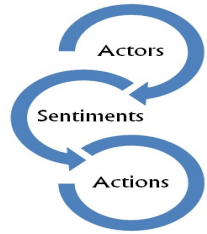

Figure 2: Textual/Meme-Synergies

"Obama" is an Actor which is strongly correlated with another noun "president" in most media irrespective of bias. On the other hand, based on context and bias, a controversial concept or Actor such as "Tamil Tigers" could be related to the sentiment adjectives "terrorist", "separatist", or "heroic" by different media sources.

We now briefly describe how we build our semantic framework. We use LDA to extract the set of topics from our database that covers a specific month in order to have a time frame to compare to. We analyze news articles originating from individual sources and then do a general LDA topic extraction using all the data across all sources for that month. We generate 10 topics for each news source and characterize each one by a *topic signature* comprising the top 10 words for the topic. These words are then fed to Antelope and with the help of a wrapper written as part of our model, identify the semantic properties of these words and all the words in the semantic chunks that they are associated with. These constructs become the *anchors* for the formation of more complex *concepts* which, in turn, are organized to form newsMemes. These concepts thus become the semantic basis for sentiments, actors and actions. This allows us to create a hierarchical network of coherent and relevant sets of words and phrases which we can then use to identify memes. We can identify all the role that an actor can assume in all the news stories and their attached sentiments. For example the rigged "elections" in Ivory Coast had very different sentiment values than those promised for Egypt after their revolution. At the very center of a large group of similar topic-oriented news stories is a cluster of memes that are organized based on their distance in terms of phrases. Visualizing this as a network, it is possible to see different groups of equally connected phrases, with the ones with the highest values at the very center of the network and the rest distributed around the center.

When organizing our synergies, we simply reinforce the stronger coherent memes and inhibit the weaker ones. This results in the emergence of more compact phrases that truly explain the reigning meme or dominant semantic topic from the LDA-generated topics. The ambiguity when viewing the traditional bag-Of-words topics from LDA arises because a simple distribution of words representing different topics does not explain much and can be hard to interpret. In our approach, reinforcing the strongest memes within each LDA topic

and across all topics gives a more robust process to identify the memes behind news from specific media outlets in a more human-understandable fashion. In many cases, we discovered that actually the same semantic topic can have multiple synergies that are related but not identical, as will be shown in the next section 5. We also study the distributions and *shapes* of the media's meme-synergies in a qualitative way to see if we can find any correlation between the bias of the sources and the specific subject-topics.

5 Results

We have been collecting and building an extensive database covering 33 online world-wide news media sources through their RSS feeds [9] to test our assumptions and model. We collect all news articles from these media sources around the clock at specific intervals. Obviously, this produces significant redundancy because an important news article can "stick" on the list of top news longer than others, or might "evolve" with time as more information and analysis becomes available regarding its event. This pattern of lingering is important to our model and we capture it so that we can measure the intensity and importance that a news item represented for its parent media source. This represents another aspect of bias, which is not covered by the conventional systems that only look for keywords such as adjectives. This aspect of bias captures the *preferential alignment* towards a certain direction or subject, which is another *latent* way of influencing the readers based on a specific agenda. Repetitiveness is used as reinforcement that strengthens the bonds across the synergetic functions. As previously mentioned we used a smoothed-LDA method based on the work of Newman [12] in addition to a wrapper program around Antelope in order to use it efficiently for our NLP analysis. In the following tables and graphs, we describe the different aspects of bias characterization that our model has to offer with respect to actual events in the news.

We start by showing in table 1 the smoothed-LDA topic results for a sample of media sources – The New York Times, BBC, CNN, USA Today and CBS – for the month of December 2010. After getting the 10 topics for each media source, we chose 3 general 'themes' because judging from the bag-Of-words result alone, that is the best a human reader can discern. The themes were ***China***, ***WikiLeaks*** and the ***Koreas***. As a reminder, in the month of December there was an international issue with China protesting the winning of a Nobel peace prize by one of its citizens, the Wikileaks event with the publication of secret documents, and the incidents between North and South Korea with the North firing at the South. Table 1 shows for each theme the resultant topic for those media sources and also indicates the table number where the meme results for that theme are shown. We should state that the comparison of the different news sources will yield more results when multiple media outlets are compared across many themes so that the pattern and differences become more apparent.

The results shown in table 2 are from the NY Times, BBC and CNN for the topic involving the story about China. Note that this theme did not appear

Media	*Theme*	Bag-of-Word Top 10 Topic Words	Table ref.
NY Times	China	china chinese lives peace prize secretary forces russian friday president	Table 2
BBC	China	election china minister host government internet peace presidential prize help	Table 2
CNN	China	peace prize thursday christmas nobel country death chinese christian explosion	Table 2
NY Times	WikiLeaks	wikileak american cables diplomatic continues site update leak reader blower	Table 3
BBC	WikiLeaks	police wikileak afghanistan country president founder court minister high assange	Table 3
	WikiLeaks	wikileak government cables attack manchester united league minister released reveal	Table 3
USA Today	WikiLeaks	wikileak police christmas julian assange country house british friday million	Table 3
CNN	WikiLeaks	wikileak charges court assange authorities julian president police thursday london	Table 3
	WikiLeaks	minister wikileak government prime website sunday diplomatic cables told latest	Table 3
NY Times	Koreas	afghan attack north south leader american tension korea killed killing	Table 4
	Koreas	afghan attack north south leader american tension korea killed killing	Table 4
BBC	Koreas	south korea north according film award children figures suggest tiger	Table 4
CBS	Koreas	president obama north korea report prince attack afghanistan south leader	Table 4
	Koreas	military report china glor jeff korea show north speed drill	Table 4
USA Today	Koreas	korea military attack north south friday sunday month bomb terrorist	Table 4
	Koreas	korea north south island saturday attack korean monday official british	Table 4
	Koreas	afghanistan korea killed killing north attack taliban suicide eastern monday	Table 4

Table 1: LDA Topics for the New York Times, BBC and other sources for the three themes

in the top 10 global themes that were common to all the 33 media outlets but was a theme that appeared in the top 10 topics for the individual media outlets. What we first notice is that *in general*, the highest ranking meme has fewer words than lower-ranking ones, and that is normal because it is supposed to be more general. We see that for the NY times the important results were *Nobel Peace Prize*, *Liu Xiabo dissident* and *block foreign news* which clearly explains the story which is about a dissident Chinese citizen who won the Nobel Peace prize and that China is blocking the foreign news sources from this subject. The BBC has similar content but did not mention that foreign news where being blocked, though one can still discern the gist of the news. CNN, on the other hand, seemed to have at least the same information as the BBC but has also connected to so many other concepts that the story became more confusing. Another interesting point was that the system uncovered 2 memes for the NY Times.

The results shown in table 3 are from the NY Times, BBC, USA Today and CNN for the topic involving the Wikileaks story. Note that this theme did appear in the top 10 global themes that were common to all the 33 media outlets. What we first notice is that CNN happens to have the same pyramid-like shape for its memes, while other media outlets kept the same shape for the previous theme as well. What is interesting is that from all news sources, it was the BBC's meme that mentioned the sexual allegations against the Wikileaks founder Julian Assange. The NY Times was again more objective with clear phrases such as *American whistle-blowers Web site* while the CNN meme mentioned all the different concepts that were affected by the leaks, such as Italy's Berlusconi and Afghanistan's Asif Rahimi etc.

Finally the results shown in table 4 are for the NY Times, BBC, USA Today and CBS for the topic involving the story about the 2 Koreas. Note that this theme did appear in the top 10 global themes. We start to see that NY Times memes always start with the main Actors that are involved and then proceed to the Action involved in the event. It is evident from other news sources that the memes have now more sentiments i.e. *defiant*, *major* etc. It is interesting that CBS uses terms such as *war track*, *remains defiant* and especially *Attack Exercises* versus NY Times' *Island Drills* at the same ranking level. The more provocative nature of the former is evident with respect to the more descriptive definition of the latter. It is also evident that comparing each media outlet's LDA's 10 top topics to the global LDA's 10 top topics (shown in figure 4) across every day in December 2010, the topics that had the most match where the topics with larger memes. In figure 3 we show in a graph how the BBC's Koreas meme looks, with the nodes with the same distance from the central node arranged on a circle and showing the semantic connections that link each node to the other.

6 Conclusion

The study reported in this paper investigated a novel approach to infer information from a large and diverse corpus of political news. We hypothesized that

#	NY Times	BBC	CNN
1.	Friday	China	Chinese
2.	Chinese ceremony	Nobel Peace Prize	dissident China
3.	Nobel Peace Prize	Nobel Chinese Dissident Liu Xiaobo	peace imprisoned Liu Xiaobo prize nobel thursday
4.	Liu Xiaobo dissident	ceremony jailed	award committee Nobel Peace Prize
5.	imprisoned	committee website winner	Pope Benedict *XVI* Middle East absentia Friday protests condemnation laureate renowned artist human rights advocate rhetoric interference internal affairs first-ever Norwegian choice latest casualty government effort no-fly list prominent guests travel ban ceremony director Nobel Institute winner Peace Prize chair troubled lands
1.	China		
2.	block foreign news		
3.	schedules Skip Ceremony Nobel		

Table 2: Chinese Dissident Nobel Prize Memes

#	NY Times	BBC	USA Today	CNN
1.	leak	Assange Julian founder Wikileaks	Julian Assange	cables diplomatic
2.	WikiLeaks	court	WikiLeaks founder	government US Prime Minister parliament WikiLeaks
3.	diplomatic	London	British	sunday Iraqi Nuri al-Maliki latest Tuesday
4.	cables	appeal allegations sexual	judge Tuesday bail	coalition cabinet rival incumbent Monday months-long political stalemate Ayad Allawi series votes leadership dispute ally summer allegations website
5.	American whistleblowers Web site	custody assault	police court appeal	online leaked island nation country deep sectarian tensions throats Zimbabwe elections process view writers minister level Agriculture Asif Rahimi bribery release pages military information United States Wednesday thousands sensitive picture corruption Afghanistan fall Berlusconi confidant cable supporters Australian

Table 3: WikiLeaks

#	NY Times	BBC	USA Today	CBS
1.	North Korea	South Korea	North Korea	president
2.	South Korea	live-fire North Korea north	military	North Korea
3.	South American North Korea island drills	largest force tensions tinderbox provocations planned military drills disputed maritime border major exercises warnings drill Women defectors envoy Russia today sacred war	attack South Korea	South Remains Defiant Attack Exercises
4.	Talks live-fire artillery Yeonpyeong		drills island north shelling	War track
5.	leader korea Sign US-China Tension Delay Taliban Afghanistan Military nuclear		jets live-fire month sunday chief defense south deadly Korea border troops	Drills Heavily Armed Border Planned Last Minute Remains Defiant review strategy troops US North Korea Lee Myung-bak

Table 4: South Korea and North Koreas Tensions

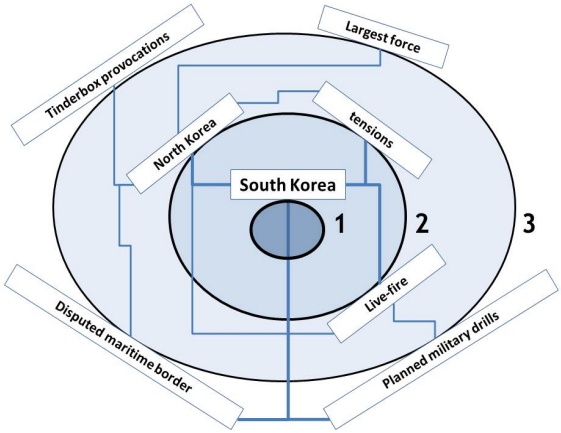

Figure 3: BBC's South and North Koreas synergy for December 2010

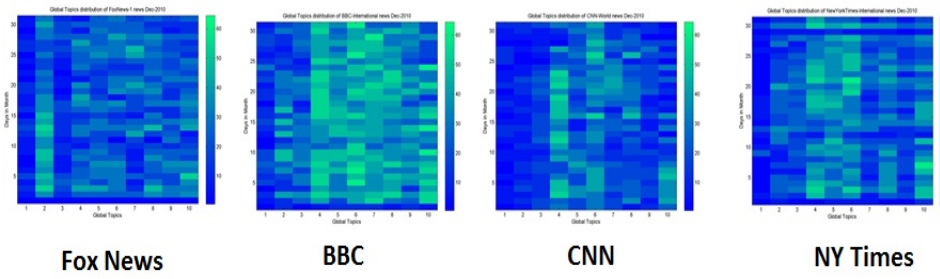

Figure 4: Comparison between 4 Media Sources for Normalized Topics

the semantic coherence of an article is a result of it being influenced by a textual synergy, and that different news sources can be evaluated by observing their corresponding synergistic memes. By organizing the standard LDA's topic distribution through NLP methods, we can achieve better understanding of the news corpus for deeper analysis and comparison.

Bibliography

[1] Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, and V.S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone", *International AAAI Conference on Weblogs and Social Media (ICWSM)* (2007), 203–206.

[2] Blei, D.M., A.Y. Ng, M.I. Jordan, and J. Lafferty, "Latent dirichlet allocation", *Journal of Machine Learning Research* **3** (2003), 993–1022.

[3] Cesarano, C., A. Picariello, D. Reforgiato, and V.S. Subrahmanian, "The oasys 2.0 opinion analysis system.", *International AAAI Conference on Weblogs and Social Media (ICWSM)* (2007), 313–314.

[4] Future, Recorded, "Recorded future - temporal & predictive analytics engine, media analytics & news analysis" (2010), [Online; accessed 22-November-2010].

[5] Gerner, D.J., R. Abu-Jabr, P.A. Schrodt, and . Yilmaz, "Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions", *International Studies Association of Foreign Policy Interactions* (2002).

[6] Hofmann, T., "Probabilistic latent semantic analysis", *Uncertainty in Artificial Intelligence, UAI99* (1999), 289–296.

[7] Kim, D., and A. Oh, "Topic chains for understanding a news corpus", *12th International Conference on Intelligent Text Processing and Computational Linguistics(CICLING 2011)* **12** (2011).

[8] Leskovec, J., L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), 497–506.

[9] Libby, D., "Rss 0.91 spec, revision 3", *Netscape Communications* (1997).

[10] McClelland, C., "World event/interaction survey", *Defense Technical Information Center* (1971).

[11] Minai, A., M. Perdoor, K. Byadarhaly, S. Vasa, and L. Iyer, "A synergistic view of autonomnous cognitive systems", *Proceedings of the World Congress on Computational Intelligence* (2010).

[12] Newman, D., "Topic modeling scripts and code", *Department of Computer Science, University of California, Irvine* (2010).

[13] Oh, A., H. Lee, and Y. Kim, "User evaluation of a system for classifying and displaying political viewpoints of weblogs", *AAAI Publications, Third International AAAI Conference on Weblogs and Social Media* (2009).

[14] Palantir, "Privacy and civil liberties are in palantirs dna" (2004), [Online; accessed 10-December-2010].

[15] Proxem, "Antelope (Advanced Natural Language Object-oriented Processing Environment)" (2010), [Online; accessed 30-April-2010].

[16] Tomlinson, R.G., "World event/interaction survey (weis) coding manual", *Department of Political Science, United States Naval Academy, Annapolis, MD.* (1993).