Algorithms **2012**, *4*, 469-489; doi:10.3390/a5040469

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Contextual Anomaly Detection in Text Data

Amogh Mahapatra *, Nisheeth Srivastava and Jaideep Srivastava

Department of Computer Science, University of Minnesota, 200 Union St SE, Minneapolis 55455, USA; E-Mails: nsriva@cs.umn.edu (N.S.); srivasta@cs.umn.edu (J.S.)

* Author to whom correspondence should be addressed; E-Mail: mahap022@umn.edu; Tel.: +61-2669-2974.

Received: 20 June 2012; in revised form: 10 October 2012 / Accepted: 11 October 2012 / Published: 19 October 2012

Abstract: We propose using side information to further inform anomaly detection algorithms of the semantic context of the text data they are analyzing, thereby considering both divergence from the statistical pattern seen in particular datasets and divergence seen from more general semantic expectations. Computational experiments show that our algorithm performs as expected on data that reflect real-world events with contextual ambiguity, while replicating conventional clustering on data that are either too specialized or generic to result in contextual information being actionable. These results suggest that our algorithm could potentially reduce false positive rates in existing anomaly detection systems.

Keywords: context detection; topic modeling; anomaly detection

1. Introduction

The ability to detect anomalies in streams of text data finds broad applicability in a number of web applications. For example, it can be used to detect the occurrence of important events from Twitter streams, the occurrence of fraudulent communication on email networks and even faulty descriptions in maintenance logs. An important emerging application of fault detection rests in the domain of organizational security, where it is critical for the integrity of the organization that sensitive information not be leaked. In all these cases, anomalies are detected through statistical comparisons with typical data and subsequently evaluated by human analysts for relevance and accuracy. The necessity of this latter step arises from the insight that, for the detection of interesting events in media streams, an aggressive

approach, leading to significantly many false positives, is often more useful than a conservative approach that promotes false negatives [1, Sections:3.1–3.2]. This observation holds the most strongly for security applications, where false negatives, while rare, can carry enormously higher costs as compared to the cost of weeding out irrelevant false positives.

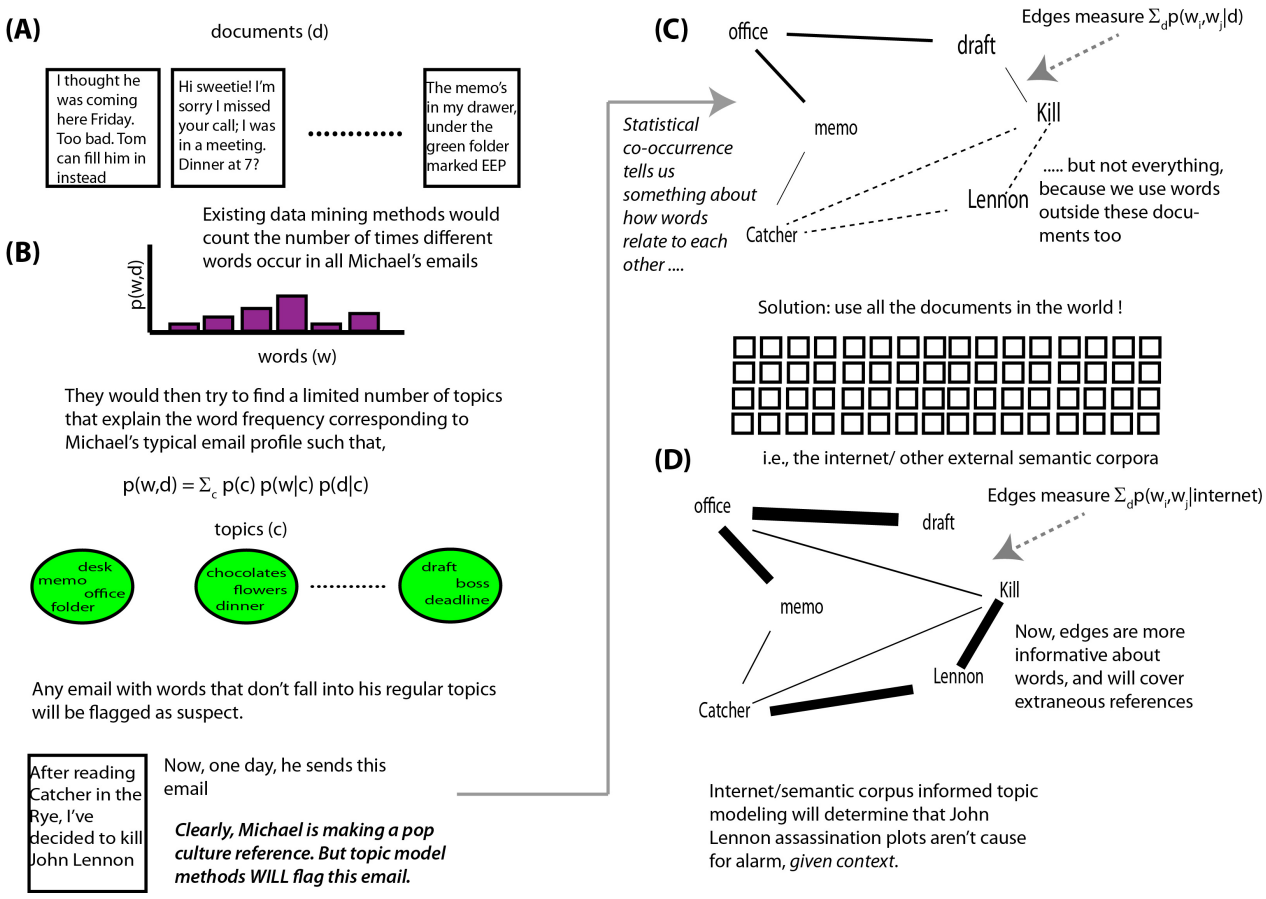
It is therefore desirable, in many application settings, to be conservative in rejecting data samples as anomalies and then post-processing detected samples with human analyst input to identify real anomalies. However, the cognitive limitations of human analysts restrict the scalability of any such proposals. It is, therefore, useful to consider the possibility of augmenting data-driven statistical anomaly detection with a post-processing filtering step that automatically determines the relevance of detected anomalies using alternative criteria. For text data, an intuitive alternative criterion for such an evaluation immediately presents itself. Current statistical text analysis techniques tokenize text data into generic categorical objects. Hence, the semantic content of text data is ignored. Reintroducing semantic information in text data analysis creates the possibility of evaluating the relevance of anomalies detected using background semantic context from a much larger corpus of data.

In this paper, we show how information about the semantic content of text data allows us to identify occurrences of words and topics that are statistically infrequent, but contextually plausible for the monitored system, and hence, not anomalous. We do this by bootstrapping a novel context-detection algorithm that operates on an external corpus of general semantic relationships (WordNet and Internet) to an LDA-based text clustering algorithm for anomaly detection that operates on particular datasets of interest. Clusters of co-occurring words (topics) that are rated highly anomalous by both modules are considered truly anomalous, while those that are flagged by statistical analysis, but considered typical through evaluating semantic relatedness are considered explainable data artifacts, see (Figure 1).

In order to further display the effectiveness of using contextual similarity measures to solve the problem of anomaly detection in textual data, we conducted a set of experiments on tags generated by users on popular social multimedia network, YouTube. Each video contains about 4–12 tags. We empirically demonstrate that the use of contextual measures could help us do anomaly detection at word-level as well, *i.e.*, we can detect anomalous topics and go even further and also detect anomalous words in a given topic.

The rest of this paper is organized as follows. In Section 2, we discuss the state-of-the-art in current anomaly detection schemes for text data and in systems for evaluating semantic context. We describe the rationale behind our approach, the technical details of its individual components and its algorithmic implementation in Section 3. A description of our empirical evaluation strategy and results follows in Section 4, following which we conclude with a short discussion of the implications of our results in Section 5.

Figure 1. This figure shows a notional application scenario of our algorithm, highlighting its benefits over the state-of-the-art techniques. (**A**) shows a set of emails populating Michael's inbox; (**B**) (above) shows how the topic models would extract latent themes, based solely on statistical frequencies of word co-occurrences; (**B**) (below) shows that a pop culture reference made by him is flagged anomalous due to the statistical rarity of the used words in the given corpus; (**C**) shows the relational-language graph used by topic models where edges only capture the statistical co-occurrence between the words in the corpus; (**D**) shows the contribution made by us where the edges now also capture the semantic distance between the words tuned over a very large set of external documents and are hence more contextually informed, which eventually flags Michael's email as harmless.



2. Related Work

Anomaly detection techniques in the textual domain aim at uncovering novel and interesting topics and words in a document corpus. The data is usually represented in the format of a document to word co-occurrence matrix, which makes it very sparse and high dimensional. Hence, one of the major challenges that most learning techniques have to deal with while working on textual data is being able to deal with the curse of dimensionality. Manevitz *et al.* [2] have used neural networks to classify positive documents from negative ones. They use a feed forward neural network which is first trained on a set of positive examples (labeled) and then in the test phase the network filters out the positive documents from the negative ones. In another work which also uses the principle of supervised learning,

Manevitz *et al*. [3] have used one class SVMs to classify outliers from the normal set of documents. They show that it works better than techniques based on naive Bayes, nearest neighbor algorithms and performs just as good as neural network based techniques. The above approach might not be very useful in an unsupervised setting, whereas our approach can work in both supervised and unsupervised settings. This problem has been studied in unsupervised settings as well. Srivastava *et al*. [4] have used various clustering techniques like k-means, Sammons mapping, von Mises–Fisher Clustering and Spectral Clustering to cluster and visualize textual data. Sammons mapping gives out the best set of well separated clusters followed by von Mises–Fisher Clustering, Spectral clustering and K-means. Their technique requires manual examination of the textual clusters and does not talk about a method of ordering the clusters. Agovic *et al*. [5] have used topic models to detect anomalous topics from aviation logs. Guthrie *et al*. [6] have done anomaly detection in texts under unsupervised conditions by trying to detect the deviation in author, genre, topic and tone. They define about 200 stylistic features to characterize various kinds of writing and then use statistical measures to find deviations. All the techniques we describe above rely entirely on the content of the dataset being evaluated to make their predictions. To the best of our knowledge, ours is the first attempt that makes use of external contextual information to find anomalies in text logs. Since the topics detected in statistical content analysis strip lexical sequence information away from text samples, our efforts to reintroduce context information must look to techniques of automatic meaning or semantic sense determination.

Natural language processing techniques have focused deeply on being able to identify the lexical structure of text samples. However, research into computationally identifying the semantic relationships between words automatically is far sparser, since the problem is much harder. In particular, while lexical structure can be inferred purely statistically given a dictionary of known senses of word meanings in particular sequences, such a task becomes almost quixotically difficult when it comes to trying to identify semantic relations between individual words. However, a significant number of researchers have tried to define measures of similarity for words based on, e.g., information-theoretic [7,8] and corpus overlap criteria [9,10]. Cilibrasi and Vitanyi have shown [11], very promisingly, that it is possible to infer word similarities even from an uncurated corpus of semantic data, *viz.* the Web accessed through Google search queries. This observation has been subsequently developed and refined by [12] and presents possibilities for significant improvements to current corpus-based methods of meaning construction. Semantic similarity measures have been used in the past to accomplish several semantic tasks like ranking of tags within an image [13], approximate word ontology matching [14] *etc*., and thus hold the promise of further simplifying cognitively challenging tasks in the future.

3. Methods

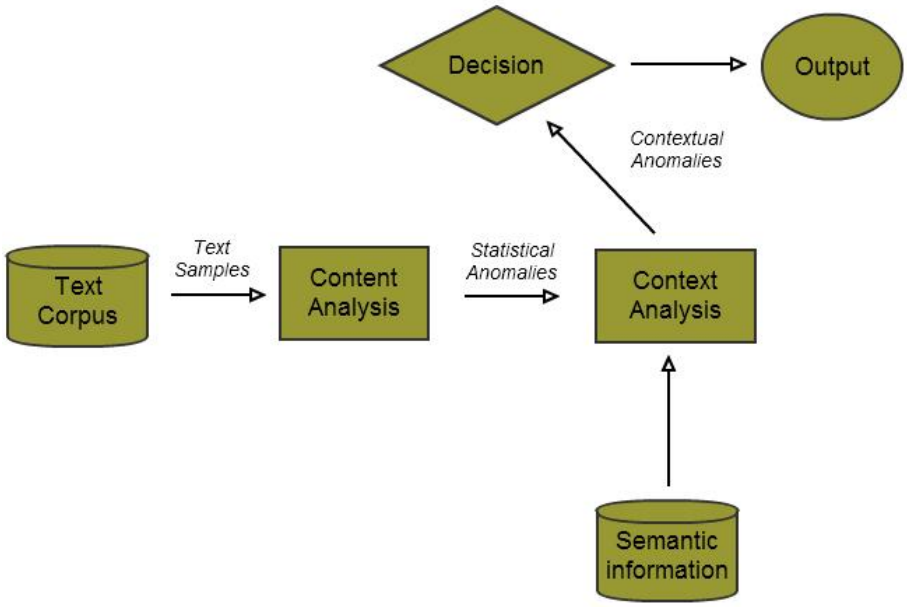
We would like to detect the occurrence of abnormal/deviant topics and themes in large scale textual logs (like emails, blogs *etc*.) that could help us in inferring anomalous shifts in behavioral patterns. Our system should then use two inputs: a text log whose content is of immediate interest and an external corpus of text and semantic relationships to derive contextual information. The output would be the set of final topics considered anomalous by our system. It is possible to subsequently evaluate the topics discovered by our dataset manually and flag a certain number of topics as anomalous. The efficacy of

our approach is measured as the ratio of number of anomalies correctly detected by the system to the number of anomalies flagged through manual inspection.

Consider all words to live on a large semantic graph, where nodes represent word labels and edges are continuous quantities that represent degrees of semantic relatedness to other words. Some subset of these words populates the documents that we evaluate in clustering-based text analysis schemes. A document can then be described by an indicator vector that selects a subset of the word labels to populate an individual document. Traditional document clustering creates clusters of typical word co-occurrences across multiple documents, which are called topics. However, such an approach throws away all information embedded in the semantic network structure. The goal of a semantically sensitive anomaly detection technique is to integrate information potentially available in the edges of the semantic graph to improve predictive performance (Figure 1).

Rather than attempting to include semantic information within a traditional clustering scheme by flattening the relatedness graph into additional components of the data feature vector, we attempt to introduce context as a post-processing filter for regular topic modeling-based anomaly detection techniques. By doing so, we simplify the construction of our algorithm and also ensure that the results from both unfiltered and filtered versions of the algorithm are clearly visible, so that the relative value added by introducing semantic information is clearly visible (Figure 2). Since our approach involves combining two separate modes of analyzing data—one content-driven and one context-driven—we now describe both these modalities in turn, and subsequently, our technique for combining them.

Figure 2. Flowchart of the decision engine.



3.1. Statistical Content Analysis

Statistical topic models based on Latent Dirichlet Allocation (LDA) (Blei *.et al.* 2003 [15]) have been applied to analyze the content of textual logs in order to identify the various topics/concepts that exist in them at different levels of thematic abstraction. LDA represents each topic as a distribution over words

and allows for mixed memberships of documents to several topics. LDA is very effective in identifying the various thematically coherent topics that exist in document collections and producing a soft clustering solution of the documents. In addition, LDA's iterative nature allows it to scale to document collections containing millions of documents (Newman *et al.* 2007 [16]).

Algorithm 1 The Semantic Anomaly Detection Algorithm

Input: Documents D, number of topics n, partition parameter m, anomaly threshold k
Output: Set of anomalous topics
B=ComputeBagOfWords(D) {Computes the statistical frequencies of word co-occurrences in the corpora, Section 3.1}
T=LDA(B,n) {Cluster bag-of-words into n topics, Section 3.1}
$T1$=Rank(T) {Rank topics based on document co-occurrence, Section 3.1}
$Test$ = Partition($T1$,m) {Partition into typical and test topic sets, Section 3.3}
FindContext($Test$) {Find context from semantic measures, Section 3.2}
R=Decision($Test$,k) {Pick k lowest context score topics, Section 3.3}
Output R

LDA's generative model is illustrated in Figure 3. Topics are assumed to be distributions over words. A Dirichlet distribution with parameter α is assumed as the prior over all documents. From Dirichlet(α) each document is drawn as a multinomial distribution θ. The dimensionality of this distribution is determined by the number of topics. For each word within a document we draw a topic from this multinomial distribution. We subsequently go to the corresponding topic distribution and draw a word from it. Ultimately a document is assumed to be a mixture of topic distributions. In training the model the objective is to learn the topic distributions along with the underlying multinomial distributions θ representing the documents. In LDA the resulting topics can be represented by the most likely words occurring in them. For example, Figure 4 shows some of the most important terms identified from the top four topics from the Enron email dataset, which can be used as a summary of the identified topic.

Figure 3. LDA's generative model illustrated with 3 topics and four words.

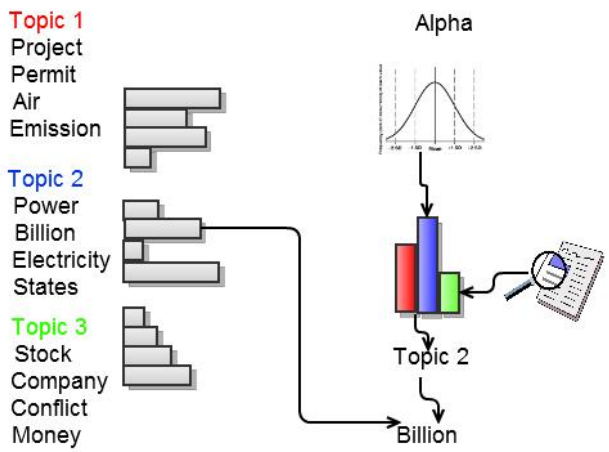
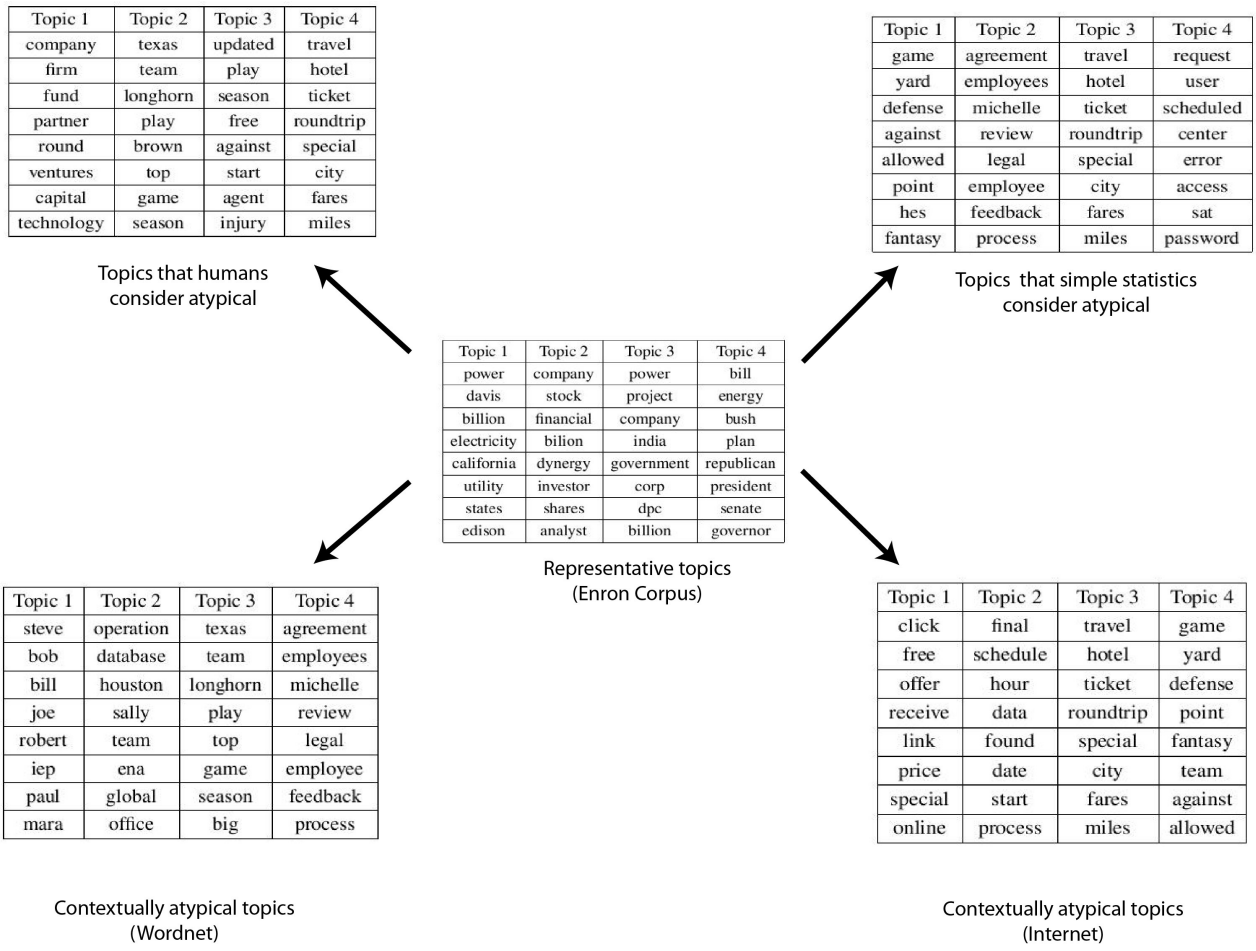


Figure 4. Results from Enron dataset. Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on WordNet (left) and Internet (right), respectively.



The documents in the text logs were converted to the bag of words format after tokenization, noise removal and removal of stop words. Then the following two steps were carried out:

1. Clustering: The text log was divided into k-Clusters using the LDA model. The values of k and other parameters (θ, α, *etc.*) were decided based on the size of the dataset and our understanding of the nature of the dataset. The top 10 most likely words were extracted as a representative summary of each topic.
2. Ranking: The topics were ordered based on their co-usage in documents. LDA assumes every document to have been created by a mixture of topic distributions. We obtain an ordering of topics based on the assumption that topics that appear together are similar to each other and should have a low relative difference in their rankings. First, the topic distributions for each topic was calculated. Then, for each pair of topics *i.e.*, (P, Q), the symmetrized KL divergence between topic distributions P and Q was calculated. Equation 1 shows the divergence measure between the probability distributions P and Q. Equation 2 shows the symmetrized KL divergence measure (henceforth SD), which has the properties of being symmetric and non-negative (Equations 3 and 4). The symmetrized KL divergence was subjected to dimensionality reduction and the first

dimension was used to rank the topics [17]. We made use of the "Topic Modeling Toolbox" to conduct out experiments.

$$D(P,Q) = \sum_{i=1}^{n} P(i) ln \frac{P(i)}{Q(i)} \tag{1}$$

$$SD(P,Q) = D(P,Q) + D(P,Q) \tag{2}$$

$$SD(P,Q) >= 0 \tag{3}$$

$$SD(P,Q) = SD(Q,P) \tag{4}$$

3.2. Computing Contexts

Determination of semantic context is, in general, a difficult question. For example, suppose we are analyzing a user's behavior on a social network over a period of time using his activity logs. Just the measurement of deviation of a user's behavior from his past behavioral patterns might not be indicative enough to make accurate predictions about his anomalous behavioral patterns. The context regarding his behavioral changes in this case could be derived from his peers' activity logs, local demographic information, *etc*., which could help in taking a more informed decision. Additionally, the trade-off between adding extra contextual information to improve predictive power and the resultant computational complexity is also important to examine.

In this paper, as we are dealing with large scale text logs, we decided to derive the contextual information from two well-known external semantic corpora namely WordNet [18] and the Internet. We describe the technical details of our procedure below:

3.2.1. Computing Semantic Similarity Using WordNet

WordNet is a lexical database for English language which has been widely used so far in many natural language processing and text mining applications. WordNet puts similar English words together into sets called synsets. WordNet has organized nouns and verbs into hierarchies of is-a relationships (example dog-is-a-animal). It provides additional non-hierarchical relations like has-part, is-made-of, *etc*., for adverbs and adjectives. Hence, two words could be considered similar if they are derived from a common set of ancestors or share similar horizontal relationships. As a result, various kinds of similarity and relatedness measures (path, wup, lch, gloss, vector, *etc*. [19]) have been defined to quantify the semantic similarity/relatedness of two words present in the database.

Pederson *et al.* [19] have enumerated 9 measures to compute the similarity/relatedness between two words in WordNet. These measures can be broadly classified into two categories. The first ones are those that quantify the similarity/relatedness between concepts based on different kinds of network distances between the concepts in the network. The second ones are those that use occurrence overlap between glossary texts associated with the concepts to measure their similarity. We have used one of each of these measures, to account for both classes of distances.

1. Path Measure: Path is a network distance measure. It is simply the inverse of the number of nodes that come along the shortest path between the synsets containing the two words. It is a measure between 0 and 1. The path length is 1 if the two words belong to the same synset.

2. Gloss Vectors: Gloss is a relatedness measure that uses statistical co-occurrence to compute similarity. Every set of words in WordNet is accompanied by some glossary text. This measure uses the gloss text to find similarity between words. Relatedness between concepts is measured by finding the cosine between a pair of gloss vectors. It is also a measure between 0 and 1. If two concept are exactly the same then the measure is 1 (angle between them is 0° and $Cos(0)$ is 1).

We take the average of those two measures for any given word pair, reflecting the fact that we have accounted for both semantic relatedness (from the first measure) and real-world statistical co-occurrence of two words (from the second measure) in our decision making process. We have used WordNet 2.05 in our experiments.

3.2.2. Computing Semantic Similarity Using the Internet

In spite of the apparent subjective accuracy of the WordNet corpus, the static nature of this corpus lacks the rich semantic connectivity that characterizes natural language interactions in social settings. Words, phrases and concepts change meanings during the course of their everyday usage and neologisms (e.g., lol, rotfl) and colloquialisms are added everyday to the human lexicon. Therefore, for its ability to offer higher conceptual coverage than any known semantic network or word ontology, we base our semantic meaning extraction on the entire World Wide Web by using the Normalized Google Distance for this purpose, as we describe below.

Given any two concepts, (a, b), the Normalized Google Distance (NGD, henceforth) between them is given by Equation 5. $f(a)$ and $f(b)$ denote the number of pages containing a and b as returned by Google. $f(a, b)$ is the number of pages containing both a and b and N is the total number of pages indexed by Google.

$$NGD(a, b) = \frac{max(\log f(a), \log f(b)) - \log(f(a, b))}{\log N - min(\log f(a), \log f(b))} \tag{5}$$

The range of this measure is (0,∞). The value 0 would indicate that the two concepts are exactly the same and ∞ would indicate they are completely unrelated. Normalized Google Distance is a non-metric and hence triangle inequality (Equation 6) does not always hold. If a, b, c are any three random words, then:

$$NGD(a, c) \not<= NGD(a, b) + NGD(b, c) \tag{6}$$

Normalized Google Distance is symmetric.

$$NGD(a, b) = NGD(b, a) \tag{7}$$

Equation 5 implicitly assumes that the ratio of the total number of pages returned by Google for a given term divided by the total number of pages indexed by Google is equal to the probability of that search term as actually used in the society (at any given point of time). Normalized Google Distance, which is based on the idea of normalized information distance and Kolmogorov Complexity [11], exploits this contextual information hidden in the billions of web-pages indexed by Google to generate a sense of semantic distance between any two concepts. NGD based natural language processing experiments have shown up to 87% agreement level with WordNet [11].

3.3. The Decision Engine

The content based anomaly detection engine clusters and ranks the topics as described earlier. The decision engine, based on a certain threshold value set by the operator (in our case $m = 2/3$, reason discussed later) divides the topics into two fractions, the first fraction consisting of two-thirds of the highest ranked topics that we assume to be normal and the second fraction consisting of the rest of the topics that are initially assumed to be potentially anomalous. Each topic is represented by a tuple consisting of its top ten words. Each anomalous topic is compared with each of the normal topic in the following manner:

1. We find the semantic distance between the first word in an anomalous topic and the first word in one of the normal topics. We aggregate it over all words in the given normal topic and aggregate that over all the words in the given anomalous topic. Let $S^m(i, j)$ denote the similarity between the i^{th} word in a test topic and j^{th} word in the m^{th} typical topic. I and J stand for the total number of words in a test topic and a typical topic respectively (10 in our case). Then relatedness of the test topic with one typical topic is measured as follows. The semantic distances used by us were the ones based on the WordNet and Normalized Google Distance as described above.

$$r = \sum_{j=1}^{J} \sum_{i=1}^{I} S(i, j) \tag{8}$$

2. We then aggregate that over all the normal topics, which finally gives us a measure of similarity of the given anomalous topic with all the normal topics. M stands for the total number of typical topics. Hence, relatedness between the given p^{th} test topic and all the typical topics is measured as follows:

$$r_{total} = \sum_{m=1}^{M} \sum_{j=1}^{J} \sum_{i=1}^{I} S^m(i, j) \tag{9}$$

3. We repeat the two steps above for all potentially anomalous topics.

Based on the newly available contextual information, the anomalous topics are sorted in ascending order of the context score obtained. Topics with the k lowest scores are considered anomalous. In our current setup, m is determined via user input, since we are modeling anomaly detection in unsupervised settings. This technique can easily be adapted to supervised settings, where a white list of allowable or typical topics informs our judgment of the m threshold in the decision engine. It is also possible to adaptively use analyst input to change these parameters to refine the accuracy of our system during operation. For example, based on past statistics or domain knowledge, if the operator decides that a certain kind of topic should not be considered anomalous, then he can incorporate it in the list of normal topics, so that topics with scores close to this one are no longer considered anomalous. Similarly, if the operator decides that a certain topic should always be considered anomalous, the threshold m can be revised upwards.

3.4. Anomaly Detection in Tag Space

Text corpora consisting of a collection of structured textual units like blogs, emails, abstracts *etc*., have a closer resemblance to the natural language interactions prevalent between human subjects in

the society. Hence, a question worth investigating is if the idea of using contextual information from semantic networks to accomplish the task of anomaly detection could also work at the word level and not just at the topic level. We investigated this question by conducting the anomaly detection experiments on tags collected from YouTube videos.

We use the "Normalized Google Distance" as our semantic measure in this case due to the noisy/colloquial nature of user supplied tags. Our dataset consisted of 50 sets of tags with one anomalous word each. Table 1 shows five such sets with the anomalous word in bold. The tag labeled anomalous was the one that the majority of our human reviewers (N = 10) agreed upon to be anomalous. The effectiveness of our technique is judged based on the number of times the most anomalous tag adjudged by the human reviewers is matched by the one flagged by our technique.

Table 1. Example of Tag-Sets (Visually determined anomalous tag In Bold).

Tag Set Number	Tag Set
1	Education, Tutorial, Teacher, Class, **Somersault**
2	Yoga, Health, Exercise, **Shocking**
3	Lonely, Island, Holiday, Adventure, **Rap**
4	Cute, Babies, Funny, Laughter, **Blood**

Suppose a video V is accompanied by a set of n tags $I = i_1, i_2, i_3, ..., i_n$. In order to find the most atypical tag in this set, we used a modified version of the K-farthest neighbor algorithm [1], as described below. We calculated the distance of a tag from every other tag in the set and the sum of these distances would result in a score that quantifies the overall divergence of a tag from the rest of the set. Based on value of the parameter k (1 in our case), the top k most divergent tags are flagged as anomalous. We believe that an approach similar to this could be applied in ontologies, taxonomies, concept hierarchies *etc*., to detect the presence of anomalous content.

Algorithm 2 The Semantic Anomaly Detection Algorithm For Tags

Input: Tag-Set T, number of anomalies k
Output: Set Of Anomalous Tags
Initialize the array $Scores$ to zero {This array will contain the divergence of each tag from the rest of the set.} {Find the Score for each tag.}
while i=1 to n **do**
  while j=1 to $n-1$ **do**
    D=$NGD(i,j)$
    $Score(i)$+=D {Find the Score for one tag.}
  end while
end while
Sort $Score$ in Descending Order
Output Top k tags

4. Results

We conducted experiments on three standard text datasets: Enron emails, DailyKos blogs and NIPS abstracts, which we detail below. The underlying methodology was as follows: we clustered each dataset into 50 topics. Then the topics were ranked by the content engine. The bottom one-third (in our case 16) of the topics were flagged as potential anomalies. Thereafter, each of the topics was compared with the top 34 topics by the context engine and a score was generated. Finally, based on a new threshold, some of the lowest ranked topics were declared to be anomalous.

The heuristic assumption of considering the top two-thirds of the topics normal is based on two important reasons. Firstly, because anomalies are rare, topics that are statistically rarer and hence ranked lower in our list are more likely to be semantically atypical. Secondly, during our preliminary empirical investigations we experimented with different values of the threshold and found the above assumption most effective in terms of removing anomalies. We would like to emphasize that this threshold parameter could be set using a domain expert's insights rather than relying completely on empirical tuning like in our case.

Each of the next three subsections has the following format. We start with a brief description of the dataset and then show a diagram consisting of 5 tables, (Figures 4–6, respectively). The table in the center shows the top 4 most representative topics of the dataset that reflects the general tenor of the corpus. The table in the upper right corner shows the top 4 most anomalous topics in the dataset based on content based ranking, which basically shows the most statistically "rare" topics in the given corpus. The table in the upper left corner shows the top 4 topics identified as interesting or anomalous by human evaluators. For human evaluation, we have used a majority voting scheme using 5 human subjects, who were asked to rate the anomalousness of bottom one-third of the topics generated by our technique on a scale of 1–10 (10 being the most anomalous). The human subjects were graduate students in the department of Computer Science at the University Of Minnesota, who were informed about the general tenor of these three corpora before conducting the poll. The inter-rater agreement was 0.54, which is considered to be in the range of good agreement. The tables in the bottom left and right corners show the top 4 most anomalous topics in the dataset after context matching using WordNet and Normalised Google Distance respectively. We then conclude each subsection with a discussion on the quality of our obtained results.

Table 2 shows all the quality metrics obtained by us on these three datasets. We report precision, recall, F-score, sensitivity and specificity scores of the classification of the anomalous class. The range of all these values is within $(0, 1)$ (1 being optimal). Our evaluation measure was F-score as it captures both the accuracy and sensitivity of a model. A brief description of these measures is provided below. Let tp, fp, fn, tn denote the number of true-positives, false-positives, false-negatives and true-negatives respectively, obtained from a classification task. Then, $Precision = tp/(tp + fp)$, $Recall = tp/(tp + fn)$ and $F = Harmonic\_mean(Precision, Recall)$. We also report sensitivity and specificity scores as they are quite popular in the anomaly detection literature. $Sensitivity = tp/(tp + fn)$. $Specificity = tn/(tn + fp)$.

Table 2. Results on all the datasets with and without the addition of contextual information. Notice the increase in specificity (False Positive Rate = 1 − Specificity) and hence the decrease in false positive rate upon augmenting contextual information. The F score increases with the addition of context. WordNet performs slightly better than NGD.

Dataset	Precision	Recall	F Measure	Sensitivity	Specificity
Enron without context	0.62	1	0.77	1	0
Enron with WordNet	0.889	0.8	**0.84**	0.8	0.83
Enron with NGD	0.77	0.7	0.73	0.7	0.67
NIPS without context	0.25	1	0.4	1	0
NIPS with WordNet	1	0.8	**0.88**	1	0.9167
NIPS with NGD	0.5	0.5	0.5	0.5	0.833
Kos without context	0.43	1	0.60	1	0
Kos with WordNet	0.8	0.57	**0.67**	0.57	0.89
Kos with NGD	0.75	0.42	0.54	0.4286	0.88

4.1. Enron Email Dataset

The corpus consists of a set of email messages. This dataset after de-identification of private fields was made public during a legal investigation. The original dataset contained 619,446 email messages from 158 users. A cleaned version of this corpus contains 200,399 messages from 158 users with an average of 757 messages per user [20]. We have used a bag of words version of this cleaned dataset, which contains 28,102 unique words and approximately 6,400,000 total words in the entire collection. Please note that a lot words are filtered in the process of tokenization, stop-word removal, *etc*.

Expert evaluation flagged 10 topics as deviant/interesting/noisy from the set of 50 topics in this dataset. These were topics like discussion about fantasy football, travel discussions, football teams, legal agreements, spam messages, system maintenance emails, discussion among new MBA graduates *etc.*, see Figure 4.

Our system based on the WordNet similarity measures flagged 9 topics as anomalous and 8 of those matched with the ones picked out via human selection. The two topics it could not predict correctly were (1) a topic with spam keywords and (2) a discussion thread among new MBA graduates. However, it was able to identify most noisy clusters really well. It had just one false positive (which happened to be a topic about marketing). It was able to flag a topic about legal agreements correctly as anomaly, which was ranked 35 by content ranking. It was able to filter out topics (like topic 1 in the upper right table) that are contextually uninteresting although statistically interesting, see Figure 4.

Our system based on Normalized Google Distance flagged 9 topics as anomalous and 7 of those matched with the ones picked out via human selection. The three topics it could not predict correctly were (1) a topic about fax and phone communication issues, (2) a topic about legal agreements and (3) a topic about discussions among new MBA graduates. Again it was able to identify most of the noisy clusters really well. It had two false positives, a topic about football and a noisy cluster. It was able to flag the topic with spam keywords as anomalous, unlike WordNet based measures that were unable to do so.

It should be noted that the topic about discussion among new MBA graduates was the only topic which was considered anomalous by human evaluators but not flagged by either WordNet based measures or Normalised Google Distance. As summarized in Table 2, F-score increases with the augmentation of context. Another thing worth noting is the increase in the specificity of the model, see Figure 4.

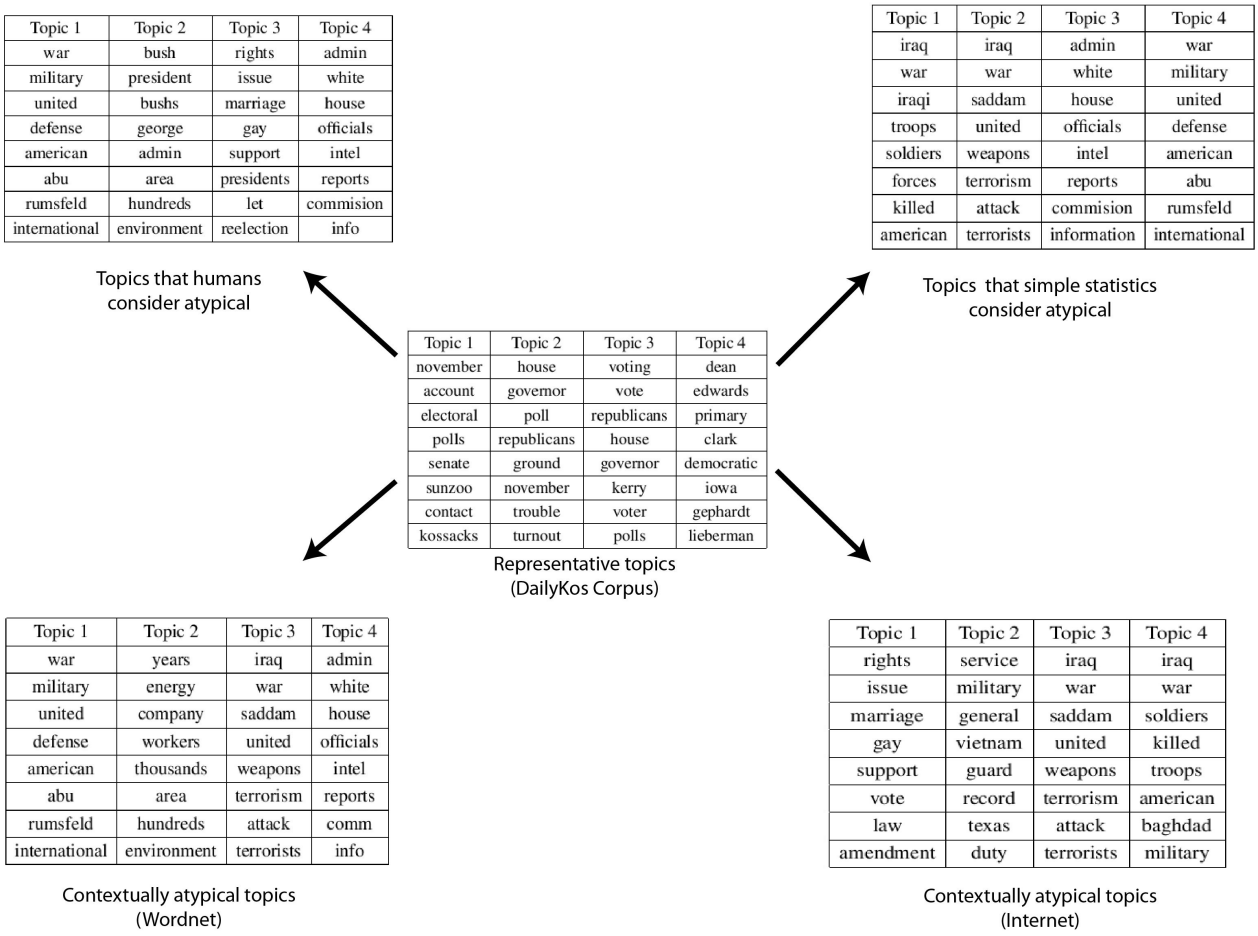
4.2. DailyKos Blogs Dataset

Dailykos.com is an American political blog that publishes political news and opinions, typically adopting a liberal stance. This dataset consists of a set of blogs taken from this website. It has 3430 documents, 6906 unique words and approximately 467,714 total words in it, in the bag of words format.

Our results on this dataset are not as clear-cut, but illuminating nonetheless. Human evaluation did not find any of the topics generated in our content analysis to be particularly anomalous (low anomaly scores assigned by humans in the scale of 1–10, 10 being the most anomalous). Our system flagged 5 topics as anomalous using WordNet similarity measures and 4 topics as anomalous while using Normalized Google Distance, none of which, however, are intuitively out of place in the DailyKos setting. The anomalous topics as shown in the bottom two tables are about Iraq war, gay rights, budget cuts, Abu-Ghraib prison incident, *etc*. Some, such as discussions of the Iraq War *etc*., are quite representative of the tenor of this website. As in the previous case, our algorithm was able to eliminate several topics in the second stage of contextual analysis which might have otherwise resulted in false positives. However, unlike in the case of Enron, where clear explanations that determine whether a particular topic is an anomaly or not are available, in this case, no such clarity is forthcoming. In particular, it appears that none of the topics determined to be anomalous in a statistical sense are not, in fact, real anomalies in the political context in which DailyKos operates. We believe this is a limitation of the external corpus that we resort to in obtaining contextual information, see Figure 5.

Recall that both WordNet and Internet are general semantic relation corpora, whereas DailyKos text is immersed in a far more sophisticated context of political discourse. As a consequence, the context inferred fails to capture the semantic associations between topics such as gay rights, Iraq war, *etc*. with typical content on DailyKos. A more dynamic approach to context detection, potentially leveraging Internet resources, would be expected to perform significantly better in this case. We believe these negative results are interesting, because they accentuate the fact that our findings in the Enron dataset are not statistical artifacts, but consequent to value added by generic semantic context.

Figure 5. Results from the DailyKos dataset. Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on WordNet (left) and Internet (right) respectively.



Topic 1	Topic 2	Topic 3	Topic 4
war	bush	rights	admin
military	president	issue	white
united	bushs	marriage	house
defense	george	gay	officials
american	admin	support	intel
abu	area	presidents	reports
rumsfeld	hundreds	let	commision
international	environment	reelection	info

Topics that humans consider atypical

Topic 1	Topic 2	Topic 3	Topic 4
iraq	iraq	admin	war
war	war	white	military
iraqi	saddam	house	united
troops	united	officials	defense
soldiers	weapons	intel	american
forces	terrorism	reports	abu
killed	attack	commision	rumsfeld
american	terrorists	information	international

Topics that simple statistics consider atypical

Topic 1	Topic 2	Topic 3	Topic 4
november	house	voting	dean
account	governor	vote	edwards
electoral	poll	republicans	primary
polls	republicans	house	clark
senate	ground	governor	democratic
sunzoo	november	kerry	iowa
contact	trouble	voter	gephardt
kossacks	turnout	polls	lieberman

Representative topics (DailyKos Corpus)

Topic 1	Topic 2	Topic 3	Topic 4
war	years	iraq	admin
military	energy	war	white
united	company	saddam	house
defense	workers	united	officials
american	thousands	weapons	intel
abu	area	terrorism	reports
rumsfeld	hundreds	attack	comm
international	environment	terrorists	info

Contextually atypical topics (Wordnet)

Topic 1	Topic 2	Topic 3	Topic 4
rights	service	iraq	iraq
issue	military	war	war
marriage	general	saddam	soldiers
gay	vietnam	united	killed
support	guard	weapons	troops
vote	record	terrorism	american
law	texas	attack	baghdad
amendment	duty	terrorists	military

Contextually atypical topics (Internet)

4.3. NIPS Papers Dataset

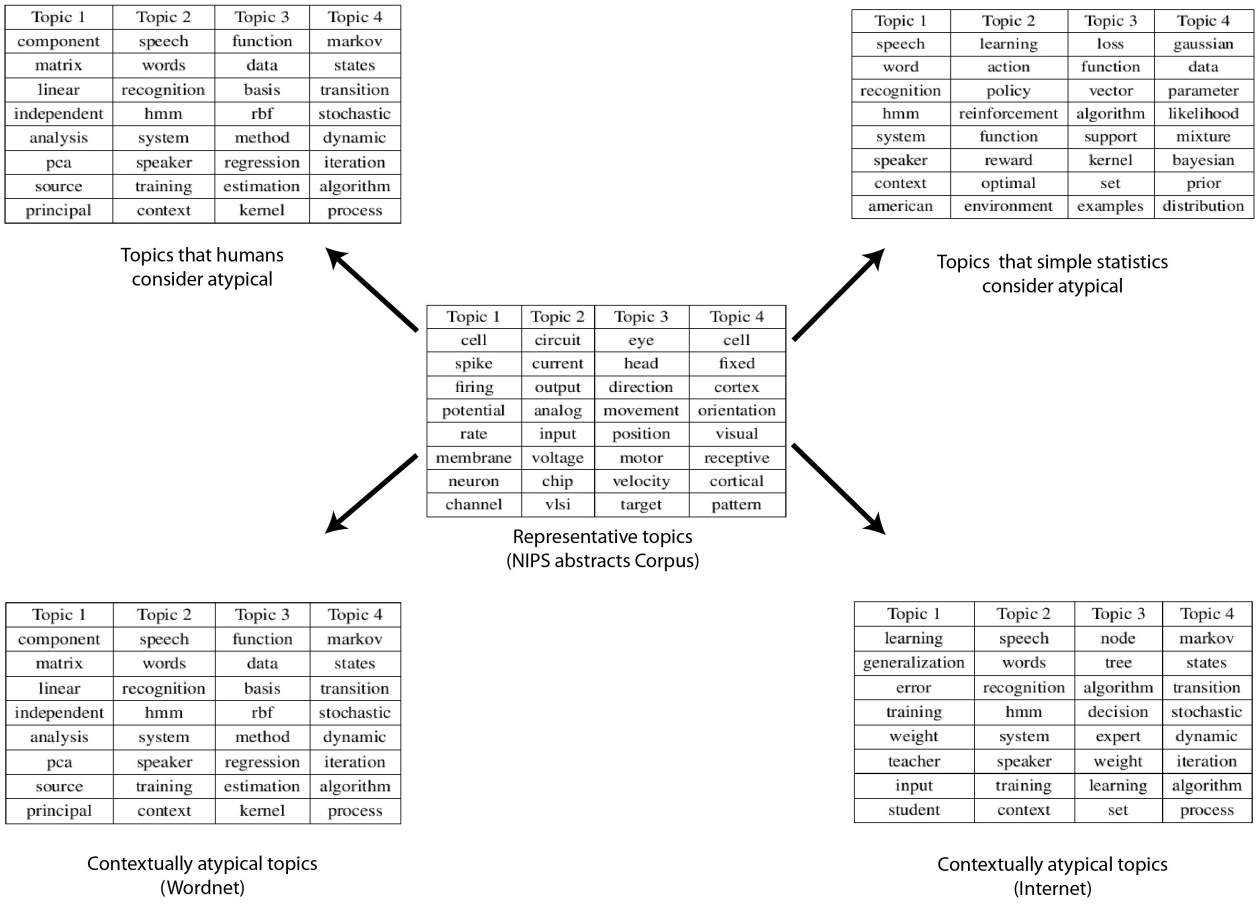
NIPS, which stands for Neural Information Processing Systems, is a conference on computational neuroscience. This dataset consists of a set of full papers taken from collection of papers published in this conference. It has 1500 documents, 12,419 unique words and approximately 1.9 million total words in it, in the bag of words format.

In comparison to the previous two datasets, this was the most difficult dataset to deal with. This was because there was very little divergence between the various topics and there was also very little noise in the dataset. Expert evaluation flagged 4 topics as deviant/interesting/noisy from the set of 50 topics. These topics were slightly less frequent than the top 34 topics and were related to acoustics, principal component analysis, regression, *etc*. Our system based on WordNet measures flagged 5 topics as anomalous and 4 of those matched with the ones flagged via human analysis. We had just one false positive (it was a topic related to support vector machines). On the contrary our algorithm based on "Normalised Google Distance" performed less well. It flagged four topics as anomalous and only two of them matched with human judgment. The amount of noise in the NIPS dataset was quite low. As

a result, the more accurate semantic network WordNet turned out to be a better match as compared with the Internet. We believe this observation could guide future decisions regarding choice of semantic measures to perform anomaly detection. The choice between WordNet and Normalized Google Distance is the choice between greater accuracy *versus* greater coverage, see Figure 6.

As this corpus consisted of a set of papers from a technical conference, it is somewhat natural to find that the text samples were heavily focused towards the central themes of the conference, which are neuroscience, artificial intelligence and machine learning. The biggest victory of our technique with this dataset was the fact that only content based analysis could have resulted in a very high number of false positives, which is something our algorithm could avoid. The number of false positives was reduced heavily as the specificity increased from 0 to 0.9167, see Table 2.

Figure 6. Results from the NIPS abstracts dataset. Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on WordNet (left) and Internet (right), respectively.



4.4. YouTube Video Tags Dataset

YouTube is a well-known community contributed video repository. Users upload videos on various themes and topics from all over the world on this website. During the process of upload, users assign certain tags which characterize the content of the given video. For example, the tag set

"*smartphone*", "*apps*", "*demo*" would indicate that the given video is likely to be a demonstration of smart-phone applications. Hence, tags can be assumed to contain a concise summary of the given video's content. Due to various reasons like presence of inherent unexpected content, user errors, use of technical jargon, cross lingual colloquials, *etc*., some videos end up having some anomalous/noisy/atypical tags. We would like to use our methodology to be able to detect those anomalous tags.

The average number of tags per video is 4–12 (This observation is made on a dataset of 42,000 videos.) The number is too low for the effective application of statistical text clustering algorithms, which rely heavily on statistical frequencies of word co-occurrences to extract meaning. Hence, we directly jump to the second step of our algorithm, which is to test for contextual atypicality. Finally, we evaluated the judgment of contextual atypicality obtained from our algorithm against the judgment of human evaluators.

Out of the 50 sets containing one-anomaly each, our algorithm was able to match the human judgment 41 times correctly (Accuracy = 0.82). Our algorithm failed to detect anomalies in tag-sets containing multiple conflicting tags or the sets where the anomalous word seemed too subtle. A few examples of the failed cases are shown in Table 3. For example, in set 3 in Table 3 the human reviewers marked the word "orange" as anomalous possibly because the words mouse, computer and monitor have strong correlation with the computer company "Apple" so the word orange seems anomalous to them, whereas the word that stands out as semantically anomalous is the word "computer", possibly because the rest of the words have some connection with biological world except this one. This experiment supports our conjecture (with reasonable amount of confidence) that contextual information could be used to accomplish anomaly detection even at word-level.

Table 3. Example of Tag Sets where the judgment of our algorithm proved inadequate. (Human adjudged anomalous tags are shown in bold. The ones detected by our algorithm are shown in italic).

Tag Set Number	Tag Set
1	fish, *net*, hook, island, **U2**
2	*motivational*, speaker, speech, inspiration, **sony**
3	mouse, *computer*, monitor, apple, **orange**
4	*angry*, birds, game, mobile, **playstation**

5. Discussion

We make three contributions to the state of existing anomaly detection techniques in this paper. First, we show how the use of external context information can reduce the false positive rate in existing systems by explaining away spurious statistical deviations. Previous research has settled the theory of anomaly detection for categorical and numeric data quite decidedly. Using similarity metrics on test data samples to estimate statistical deviance from typical system behavior proves to be a robust anomaly detection strategy across domains and datasets [1]. Because of the acontextual nature of such data objects, the use of external data to inform and contextualize anomalies detected has typically not been considered,

although [21] have recently described how contextual and functional system constraints can be used to bias anomaly detection techniques towards finding systemically meaningful anomalies. However, their approach is primarily concerned with using contextual constraints as a pre-processing step to reduce the dimensionality of the feature space to be searched for anomalies, whereas our technique uses such information as a post-processing filter.

Introducing contextual information from a different data corpus helps us find semantic explanations for anomalies and in filtering out false positives. For example, in our analysis with the Enron email dataset, the most anomalous topic after content analysis was a theme related to venture capital investment in a technology firm. This was filtered out by the decision engine after contextual analysis as a false positive, a conclusion in concordance with the evaluation of human subjects. By augmenting contextual information, we can also flag previously suspicious data points as anomalies with greater confidence.

Second, contextual information can be used to detect previously undetected anomalies, if we allow for adaptive threshold manipulation based on operator input. If we can augment these anomalies in our contextual stream appropriately, then it could help in detecting anomalies similar to the one flagged by our operator more easily in the future. For example: during our analysis with the Enron email dataset, a topic about legal agreements was ranked higher (less anomalous) by the content analysis engine, but context analysis was able to flag it out as an anomaly. Such disparity in content-context rankings can inform operators' judgment to tune the threshold better so that such topics are not missed out in the future, or to construct better similarity metrics for the content analysis task itself. Additionally, by augmenting a dummy topic in our contextual database, in the spirit of "must-link" proposals [22] in earlier constrained clustering approaches, containing words related closely to legal agreements, we can promote subsequent detection of related topics going forward.

Third, contextual information can be used to perform anomaly detection in datasets made up of user generated exemplars (YouTube tags in our case). This shows that this idea could be implemented at various levels of abstraction as we have operationalized it at word level and topic level. A possible future work in this direction would be to be able to do this exercise at document level as well.

The overall complexity of our technique is as follows:

1. The first step that clusters the text corpus into topics has a complexity of $O(((NT)^{\tau}(N+\tau)^3))$ [23], where N is the number of words in the corpus, T is the number of topics in the corpus and τ is the number of topics in a document. This has polynomial run-time if τ is a constant.
2. The second step that performs context incorporation has a complexity of $O(m*n*k*k)$, where m is the number of training topics, n is the number of test topics and k is the number of words in a topic ($k = 10$ in our case).
3. The overall complexity is thus $O(((NT)^{\tau}(N+\tau)^3)) + O(m*n*k*k)$. The second step does not affect the overall complexity of the algorithm asymptotically.

The two major computational tasks performed by our algorithm are performing topic modeling to extract topics and computing semantic distances. We mentioned earlier that the iterative nature of LDA can scale it up to millions of documents. Calculation of semantic distances using WordNet can be performed in polynomial time. Calculation of semantic distance using the web is restrictive in practice,

as search engines throttle automated queries (e.g., Google permits only up to 1000 queries per day from an IP address). In order to tackle this challenge, we suggest the use of a dictionary which enlists the number of search queries returned for a list of common terms by a search engine. Hence, we believe that with the use of map-reduce framework to perform clustering and with the use of a pre-computed Internet semantic dictionary, it is possible to apply our technique on much larger datasets.

To conclude, in this paper we have proposed that, by augmenting topic modeling techniques with contextual information derived from semantic networks, we can improve the detection of deviant topics in large scale text logs. We were able to validate this empirically and build a system that could accommodate the human judgment of anomalies as well. Our results show both reductions of false positives and detection of previously undetected anomalies in existing datasets. Extension of our approach to online settings could significantly improve existing techniques of sentiment extraction being researched using social network feeds [24]. Also, since anomaly detection here occurs at a topic level, it is possible to implement privacy-preserving tracking of intra-organizational communication using systems built around our basic concept.

Acknowledgments

We thank all the members of the DMR lab in the Department of Computer Science at the University of Minnesota for their feedback at various stages of this project. This work has been supported by a grant from the ARL Network Science CTA via BBN TECH/W911NF-09-2-0053, and NSF grant CNS-0931931.

References

1. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58.
2. Manevitz, L.; Yousef, M. Document Classification on Neural Networks Using Only Positive Examples. In *Proceedings of the 23rd Annual International ACM SIGIR Conference Research and Development in Information Retrieval*, New Orleans, USA, 24–28 July 2000; Volume 34, pp. 304–306.
3. Manevitz, L.; Yousef, M. One-class SVMs for document classification. *J. Mach. Learning Res.* **2002**, *2*, 139–154.
4. Srivastava, A.; Zane-Ulman, B. Discovering Recurring Anomalies in Text Reports Regarding Complex Space Systems. In *Proceedings of IEEE Aerospace Conference*, Los Alamitos, CA, USA, 5–12 March 2005; pp. 55–63.
5. Agovic, A.; Shan, H.; Banerjee, A. Analyzing Aviation Safety Reports: From Topic Modeling to Scalable Multi-label Classification. In *Proceedings of the Conference on Intelligent Data Understanding*, Mountain View, CA, USA, 5–6 October 2010; pp. 83–97.
6. Guthrie, D.; Guthrie, L.; Allison, B.; Wilks, Y. Unsupervised Anomaly Detection. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, Hyderabad, India, 9–12 January 2007; pp. 1626–1628.

7. Lin, D. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, USA, 24–27 July 1998; pp. 296–304.
8. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, CA, USA, 20–25 August 1995 ; pp. 448–453.
9. Jiang, J.J.; Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan 1997; pp. 19–33.
10. Mangalath, P.; Quesada, J.; Kintsch, W. Analogy-making as Predication Using Relational Information and LSA Vectors. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, Chicago, USA, 5–7 August 2004.
11. Cilibrasi, R.; Vitanyi, P. The google similarity distance. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 370–383.
12. Bollegala, D.; Matsuo, Y.; Ishizuka, M. Measuring the Similarity between Implicit Semantic Relations from the Web. In *Proceedings of the 18th International Conference on World Wide Web, ACM*, Madrid, Spain, 20–24 April 2009; pp. 651–660.
13. Liu, D.; Hua, X.; Yang, L.; Wang, L.; Zhang, H. Tag Ranking. In *Proceedings of the 18th International Conference on The World Wide Web*, Madrid, Spain, 20–24 April 2009; pp. 351–360.
14. Gligorov, R.; Kate, W.; Aleksovski, Z.; Harmelen, F. Using Google Distance to Weight Approximate Ontology Matches. In *Proceedings of the 16th International Conference on the World Wide Web*, Banff ALberta, Canada, 8–12 May, 2007; pp. 767–776.
15. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet allocation. *J. Mach. Learning Res.* **2003**, *3*, 993–1022.
16. Newman, D.; Asuncion, A.; Smyth, P.; Welling, M. Distributed Inference for Latent Dirichlet Allocation. In *Proceedings of NIPS 2008*, Vancouver, Canada, 8–11 December 2008; MIT Press: Cambridge, MA, USA, 2008; pp.1081–1088.
17. Topic Modelling toolbox. Available online: http://psiexp.ss.uci.edu/research/programsdata (accessed on 10 August 2011).
18. WordNet. Available online: http://wordnet.princeton.edu/ (accessed on 10 August 2011).
19. Pedersen, T.; Patwardhan, S.; Michelizzi, J. WordNet: Similarity-measuring the Relatedness of Concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence*, San Jose CA, USA, 25–29 July 2004; pp. 1024–1025.
20. Frank, A.; Asuncion, A. UCI Machine Learning Repository; University of California: Irvine, CA, USA, 2010. Available online: http://archive.ics.uci.edu/ml (accessed on 5 July 2011).
21. Srivastava, N.; Srivastava, J. A hybrid-logic Approach Towards Fault Detection in Complex Cyber-Physical Systems. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, Portland, Oregon, USA, 13–16 October 2010.
22. Wagstaff, K.; Rogers, S.; Schroedl, S. Constrained K-Means Clustering With Background Knowledge. In *Proceedings of the International Conference on Machine Learning*, Williamstown, MA, USA, 28 June–1 July 2001; pp. 577–584.
23. Sontag, D.; Roy, D. *Complexity of inference in Latent Dirichlet Allocation;* NIPS: Grenada, Spain, 2011; pp. 1008–1016.

24. Petrovi, S.; Osborne, M.; Lavrenko, V. Streaming First Story Detection with Application to Twitter. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, LA, USA, 1–6 June 2010; pp. 181–189.
25. WordNet: Similarity. Available online: http://marimba.d.umn.edu/ (accessed on 10 August 2011).