3-2014

Applicability of Latent Dirichlet Allocation to Multi-Disk Search

George E. Noel
Air Force Institute of Technology

Gilbert L. Peterson
Air Force Institute of Technology

Recommended Citation

Noel, G. E., & Peterson, G. L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. Digital Investigation, 11(1), 43–56. https://doi.org/10.1016/j.diin.2014.02.001

Applicability of Latent Dirichlet Allocation to Multi-Disk Search

George E. Noel[a,*], Gilbert L. Peterson[a]

[a]*Department of Electrical and Computer Engineering, Air Force Institute of Technology, 2950 Hobson Way, Wright Patterson AFB, OH, USA*

Abstract

Digital forensics practitioners face a continual increase in volume of data they must analyze, which exacerbates the problem of finding relevant information in a noisy domain. Current technologies make use of keyword based search to isolate relevant documents and minimize false positives. Unfortunately, selecting appropriate keywords is a complex and challenging task. Latent Dirichlet Allocation (LDA) offers a possible way to relax keyword selection by returning topically similar documents. This research compares regular expression search techniques and LDA using the Real Data Corpus (RDC). The RDC, a set of over 2,400 disks from real users, is first analyzed to craft effective tests. Three test are executed with the results indicating that, while LDA search should not be used as a replacement to regular expression search, it does offer benefits. First, it is able to locate documents when few, if any, of the keywords exist within them. Second, it improves data browsing and dealing with keyword ambiguity by segmenting the documents into topics.

Keywords: Latent Dirichlet Allocation, topic models, query by document, data mining, text mining, document search

1. Introduction

The rapid expansion of data size, digital devices, and cloud storage threatens to vastly complicate digital investigations. Referred to as the "big data problem" [28], it results in poor decision making, duplicated efforts, lost sales, and low productivity [16]. For the forensics or intelligence analyst, these can translate to lost opportunities, failure to present incriminating or exonerating evidence or, in extreme cases, loss of life.

According to Feldman [16], not finding information is estimated to cost a company with a thousand knowledge workers over $15 million annually. This problem is compounded as devices both increase in storage capacity and quantity per user. In addition, a data store, such as a network server or shared computers, often support multiple users. This complicates digital search, both from a technical aspect and from the human analyst's viewpoint [27].

Traditional digital investigations using keyword or regular expression searches rapidly approach the cognitive limitations of a human analyst [4]. A successful result is often a product of proper keyword selection that requires experience and luck. The human limit is compounded by the fact that two humans will select the same word to describe an object less than 20% of the time [18]. If a target document does not include any of the search keywords chosen, it will not be returned. Keyword searches must either include a variety of potential keywords and increase the false positive rate or constrain their keywords, decreasing the recall.

Latent topic models, such as the Latent Dirichlet Allocation (LDA) [9], offer a potential solution towards reducing practitioner overhead in two ways. First, they automatically extract hidden topics from a document corpus, providing a summary of the contents with minimal human intervention. This also provides structure for the user to browse contents without having to open each file. Second, they relax the requirement to match a keyword to a word in the document for each search. As long as keywords co-cluster with document words during model training, the document will be returned.

While topic models have been demonstrated successfully on large topically narrow file corpus [23] [31] [37] , they have not been tested on a set of real-world storage devices. This paper makes two primary contributions. First, it calculates statistics on user-generated data in the Real Data Corpus (RDC) [19] to identify areas suitable for search evaluation. Second, it uses the RDC to compare one of the most widely used latent topic models, LDA, to traditional keyword search using three tests. The first test compares the ability of LDA and keyword search to find specific documents within a corpus. The second tests the ability of LDA to extract a major category of files and break those into sub-categories by latent topic. The final test compares LDA against regular expression search in a noisy domain with high topical overlap.

Results demonstrate that, while LDA is not a replacement for regular expression search, it relaxes the importance of keyword selection. It also offers the ability to automatically segment a corpus into topics and sub-topics, facilitating data discovery.

2. Background

Digital investigation practitioners must deal with a wide variety of media in varied formats. A number of techniques exist, designed to either improve visualization [10] [11] or condense large quantities of data into more relevant information [24] [37]. While these techniques provide improvements over

*Corresponding author
Email address: george.noel@afit.edu (George E. Noel)

existing methods, they are seldom tested on data from real-world hard drives.

This section discusses the problem and existing search techniques for large data corpus. It then describes the challenges in acquiring a realistic corpus for testing and ways the Real Data Corpus (RDC) addresses these issues. Finally it discusses one potential solution using Latent Dirichlet Allocation (LDA).

2.1. Existing Forensic Search Techniques

Digital forensics involves complicated processes designed to maintain data integrity and facilitate discovery. Most computers contain between 10,000 to hundreds of thousands of files [12]. Looking at every file would create a significant burden for the analyst, so the search space is reduced. Through a combination of exclusions, hash analysis, and filtering, a much smaller set of files can be analyzed by specialized search techniques.

Keyword and regular expression search are popular techniques, though, the false positive rate can be high. This can be improved using filters and conditions [26], which include proximity measures and relevancy rankings using weighted terms. While adjusting relevance rankings leverage techniques that have been under use for many years [33], recent research is still attempting to increase its potential, from improved weighting [14] to neural-net clustering of string search results [4].

Relevancy improvement techniques, however, are useless if poor keywords are chosen. Selection of keywords is important to ensure at least one matches the documents being sought. This is a difficult task without wide experience in both investigations, storage technology, and the law [21]. One solution is to expand keywords in a more intelligent manner. Du, et al. [13], use WordNet to prune high-noise keywords using Latent Semantic Indexing. Text clustering provides another option, using probabilistic Latent Semantic Analysis (pLSA) or other clustering methods to identify semantically or topically similar documents [3]. These techniques have recently expanded into forensics analysis to cluster query results using self-organizing maps [5].

2.2. Forensics Test Data

Most of the research in improving digital investigations use single drive corpus or homogeneous data, such as the Wikipedia corpus. Testing with large, cross-drive real-world data is complicated since mixing evidence between cases is often not allowed. To address this problem, Garfinkel developed a corpus of hard drives purchased off the open market with a variety of real user data. Garfinkel used a similar data set with a Cross-Drive Analysis technique [20] leveraging lexigraphic data to identify real-world drives heavy with financial information, e-mails, or other information with defined formats.

2.3. Latent Dirichlet Allocation

This work evaluates another approach to large corpus search using Latent Dirichlet Allocation (LDA) [9] to identify latent topics spanning drives. Latent Dirichlet Allocation (LDA) [9] is a generative probabilistic model commonly used for identifying latent topics within a text corpus. Based off pLSA, LDA improves fit by assuming multinational probabilities have been generated by Dirichlet distributions. Unlike SOM, however, it demonstrated resistance to overfitting during comparison testing and required fewer parameter adjustments. The LDA model's posterior probability permits corpus browsing and search that relaxes requirements for precise keyword selection.

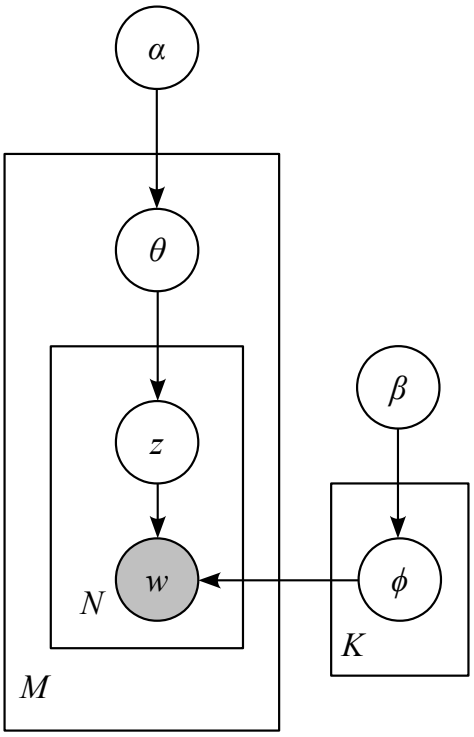

Figure 1: Latent Dirichlet Allocation [9].

LDA models each item of a collection as a finite mixture over a latent set of topics. Theoretically, a generative model produces documents and words according to its distribution. In reality, the parameters are unknown and only the words and documents are known. Using Expectation-Maximization or Markov Chain Monte-Carlo simulations, these parameters are estimated, allowing inferences to be made about new data.

The model is illustrated by the plate diagram in Figure 1. The inner plate represents the N words within the corpus while the outer plate represents M documents. Each document, utilizing the vector space model [34], is represented as a vector histogram of word frequency. The prior probability Dirichlet distributions with user-set hyper-parameters α and β generate per-document and per-word multinomial parameters.

Gibbs sampling offers an iterative approach to LDA parameter estimation using the Markov Chain Monte-Carlo (MCMC) technique. Variables are chosen, conditioned on other variables in the distribution. Reducing the overall probability distribution down to its proportional equivalent and integrating out ϕ and θ results in Equation 1. The first ratio utilizes word frequency over total words in a topic $\hat{\phi}_j^{(w)}$, while the second word-topics in a document by total words in that document $\hat{\theta}_j^{(d)}$, leaving out the

current assignment of z_i. This iterates until the log likelihood stabilizes [23].

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \quad (1)$$

The variety of uses for LDA has grown rapidly since its inception. Latent topic text co-occurrence from the original LDA paper is augmented by topic co-clustering [35], allowing for consecutive sentence and word-level clustering. Others expand it into multimedia clustering [8] [6], attempting to cluster word and image segment co-occurrence. Beyond words and images, it is used to model life patterns using cell phone data [15] twitter topic classification [32] and hierarchical models [7].

3. Corpus Analysis and Methodology

This section details the procedures used to extract data from the Real Data corpus. It then provides an analysis of the user-generated text documents and images within the corpus.

3.1. Disk Image Extraction

The Real Data Corpus (RDC) contains a wide range of disks extracted from storage devices purchased on the open market. The devices range in size from 8MB to 480GB and includes data from phones, flash cards, USB drives, and multi-partition hard drives. These drives are stored as EnCase files, Advanced Forensics Format (AFF) files, or in raw form. A total of 2,435 disks from twenty five different countries are used in testing.

This research is interested in the user-generated data. Of the 2,435 drives, 920 had user-generated files. These user-generated files are identified by searching for a list of commonly-used file extensions. To help prune out template or system files, common system directories are ignored and a SHA1 hash is used to identify duplicate files.

Each document type is prepared for analysis using a variety of techniques. Microsoft Word documents are processed using the LibreOffice API. Text is extracted to a MySQL database and embedded images are extracted to the disk, along with their location in the document. Raw text tended to be one of four types, logs, application data, readme files, or e-mails. Log data would create noise for a topic model so are excluded. Template data, readme files and application data would likely create an irrelevant topic. It was, however, difficult to eliminate log files and other template text data files completely. A regular expression scan is used to detect e-mail addresses. These files are loaded into the database and the rest ignored. Finally, images are scanned with a graph detector and face detector. Extensible image file format (EXIF) information is extracted from image files and stored in the database.

Web browser cache files are extracted separately from the rest of the files. Browsers supported include Internet Explorer, Netscape and Firefox caches on Microsoft Windows machines.

Table 2: Image Statistics.

	Regular	Embedded
Image Count	205,389	34,553
Percent Graphs	35%	47%
Percent with faces	25%	11%
Percent with camera info	32%	N/A
Images with GPS Info	324	N/A

4. Corpus Statistics

Of the 920 disk images with user-generated files, the top ten disk images with respect to file count hold 36% of the total number of files. The mean number of files per image is 977, while the median is 77, with the largest drive holding 109,938 of user-generated interest files. This indicates that most computer users don't generate a significant amount of data on their computers, while a few users generate a large number of files.

Of the 25 countries represented within the corpus, Israel has the largest number of user-generated interest files, as indicated in Table 1. Though it has a number of files in Hebrew, many are in English, making it the largest disk set in terms of searchable English data. The Indian collection also holds a number of files from a variety of disciplines, including scientific, business, and political topics. China has the largest number of drives, however, they hold few files beyond operating system and application files. Microsoft Word documents that do appear are mostly written in Mandarin. Likewise, the Mexico corpus includes a large number of user-generated interest files, however, most are in Spanish. While this may prove useful for other research, this paper constrains the scope to English documents.

4.1. File Types

Figure 2 shows file extensions of extracted files by frequency in descending order. Images make up a significant portion of the overall file count, with JPEGs being the most common file. A large portion of these images were personal photographs taken with digital cameras. Other images included those downloaded from web pages or unique images that were part of an installation. The hard drives contain over 350,000 MP3s, predominantly music.

The HTML files listed in Figure 2 include only files outside the web browser cache. These files predominantly consist of web pages saved to disk or user's manuals for software applications. Most of these documents are not generated by the user and, therefore, are not relevant for this research. Each of the html files include an assortment of images that create noise against user-generated images. Browsing through the extracted GIF files reveals large quantities of simplistic graphics that are used for web page borders or have other decorative purposes. This high number of non-user-generated files demonstrates that, even with excluding common directories and pruning duplicates, one must assume a certain level of noise.

This research extracted 205,389 images directly from disk. Another 34,553 images are found embedded in Microsoft Word documents, as shown on Figure 2. To understand the potential of this data set for image processing research, several tests

Table 1: Drive and File Statistics by Country.

Country (code)	Total Drives	Drives Loaded	Drives Failed	Total Interest Files
Israel (il)	297	283	14	146,297
India (in)	672	566	106	90,390
Mexico (mx)	175	175	0	85,479
Palestinian State (ps)	140	126	14	76,566
Serbia (rs)	8	7	1	23,184
Serbia and Montenegro (cs)	16	15	1	16,850
Canadian (ca)	18	11	7	14,173
Singapore (sg)	34	23	11	8,579
Turkey (tr)	10	10	0	7,302
Panama (pa)	17	15	2	6,434
China (cn)	808	808	0	5,141
Unknown (nnn)	109	109	0	4,072
Egypt (eg)	7	7	0	3,734
Ukraine (ua)	57	50	7	3,577
Pakistan (pk)	85	82	3	2,372
Ghana (gh)	21	20	1	1,506
Germany (de)	4	2	2	1,444
United Arab Emirates (ae)	39	34	5	624
Japan (jp)	13	13	0	423
Hong Kong (hk)	4	4	0	320
Morocco (ma)	11	10	1	44
Bangladesh (bd)	57	54	3	18
Switzerland (ch)	2	2	0	0
Hungary (hu)	22	2	20	0
Thailand (th)	4	4	0	0
Total	2,630	2,432	198	98,529

are accomplished. Graphs, such as line graphs, pie charts, and line drawings, often require different analysis than photographs. Each is analyzed to determine if it looked like a graph using the Efficient Graph-Based Segmentation [17] with Mean Threshold [30] and the mean feature vector placed into a multi-dimensional histogram. The graph detection algorithm looked for high frequency histogram spikes with small variance. These should indicate a majority of similarly-colored pixels with little variance. The more shading or color variations an image has, the less likely it will be detected as a graph. Many line drawings included shading or color shifts and resulted in a false negative. For this reason, the numbers listed in Table 2 are likely lower than the actual number of graphs.

Faces are detected using a boosted cascade of Haar-like features [36]. Images are shrunk to 400 pixels wide for consistency and to improve speed, maintaining their aspect ratio. The face detector within the OpenCV API library [1] is used to detect the number of faces within the picture. Sampling of the Microsoft Word documents indicated the vast majority were business-related documents. Most embedded images were of graphs, diagrams, or close-ups of hardware, as indicated by Table 2, while images on disk tend to have a higher number of faces.

The images on disk are also examined for Exchangeable Image File (EXIF) format information. While many images have EXIF information that indicated the editing software used to create or manipulate them, the personal photographs tend to have camera information. Fully one third of all images have camera information while a small number of those included Global Positioning System information from the location where the picture was taken.

Table 3: Embedded Image Analysis.

Number of documents	55,443
Perc. docs with embedded images	16%
Perc. docs with viable SWLDA images	8.5%
Disks with embedded image files	253
Avg embedded image file per disk	35

This research considered attempting to automatically annotate images with local context using [29]. Unfortunately, as Table 3 indicates, only 16% of the documents have embedded images with at most half those with images that are not graphs. Considering that 253 disks have documents with embedded images, there were only 35 viable documents per image with no guarantee for consistent embedded topics. The low number of images would result in a sparse super-word matrix and coarse probabilistic granularity. Due to the small likelihood of success, automated image annotation using topical context is not attempted, however, this may provide an avenue for further im-

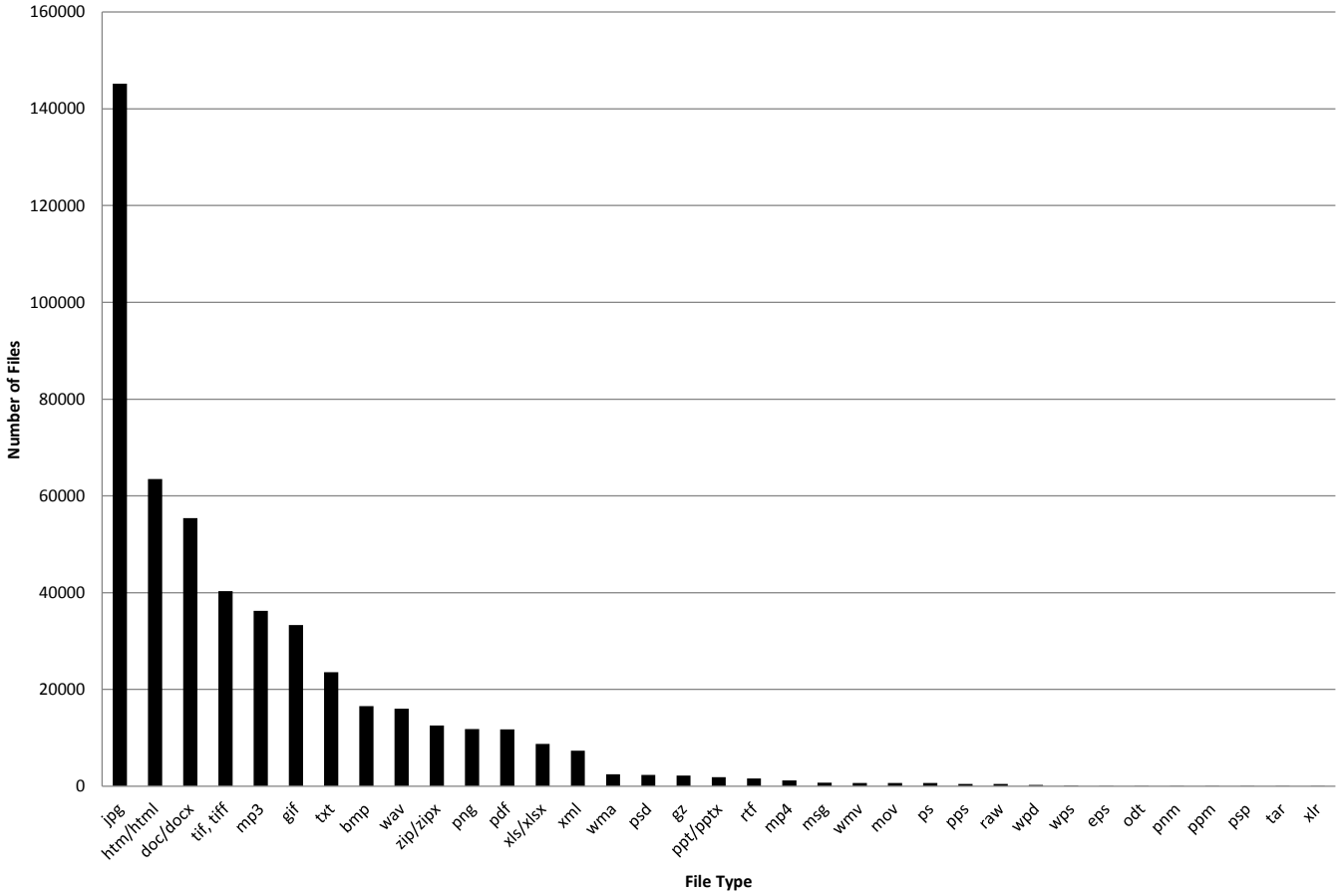

Figure 2: File types in the corpus.

age annotation research.

4.2. E-Mail and Web Caches

E-mails are initially analyzed within the corpus. Text documents were searched for e-mail addresses and, if found, were clustered using LDA. Unfortunately, due to the significant portion of advertisements and e-mail text from internet service providers, the topics do not reveal many distinct areas.

The web caches are also analyzed similarly. Most of the resulting topics appear to be clusters of text from advertisements or mail clients. Adult websites do cluster into their own topics, but other than that it was difficult to draw many inferences from the results. Better results may be gained by finding a means to filter out advertisements and standard templates for websites before clustering with LDA.

4.3. Document Search

Traditional forensics media search relies primarily on keyword search to extract useful information [25]. While there are some best practices when crafting keywords [26], most rely on experience and intuition to build a keyword list likely to produce relevant query results. Large corpus compound the search problem since high levels of false positives can quickly use up available manpower.

In addition to proper keyword selection, both regular expression search and LDA search require intelligent parameter tuning. Regular expression search can benefit from term weighting since some terms often have higher relevance for a topic than others. Ideally, relevant documents will contain a variety of keywords with high frequency. For this research, the document list is sorted first by the number of unique keywords found, then by the keyword score. The score s is calculated as $s = \sum_{i=1}^{X} w_i k_i$ where X is the set of keywords, k_i is the frequency of keyword i and w_i is the weight of keyword i. The 'unknown' target documents are identified and used to compare each algorithm based on its position in the query results.

The LDA tests are designed to be as close to the regular expression search as possible. Once the LDA model parameters have been estimated, they can be used to draw inferences about new data. For this research, the LDA model is trained using Gibbs sampling with iterations adjusted based on corpus size. For instance, models formed using the Israeli corpus were trained using 1,400 iterations due to its large size while models trained on the Serbian corpus only required 1,000 iterations. LDA fixed prior Dirichlet parameters $\alpha = 0.05$ and $\beta = 1.0$ are used based on previous comparison testing using 1,504 categorized documents from the U.S. Patent and Trademark Office's database [2]. Unless otherwise specified, twenty-five topics are

used to provide acceptable granularity while maintaining reasonable algorithmic speed. Equation 2 defines the probability that a document d_i relates to the provided set of keywords X. The multinomial parameters ϕ and θ define the document-topic and word-topic probabilities for each topic K. Weights are factored in to stress certain probabilities over others and is defined as the weight w_j for keyword x_j.

$$p(d_i|X,\theta,\phi) = \sum_{k=1}^{K}\left(p(d_{i,k}|\phi)\frac{\sum_{j=1}^{X} w_j, p(x_j|\theta)}{\sum_{j=1}^{X} w_j}\right) \quad (2)$$

Word documents are used since they have the largest quantity of user-generated text. Adobe PDF documents tend to be user's manuals or other professionally created documents downloaded from the internet. To avoid creating unnecessary noise in the experiment, PDFs are excluded. Similarly, Microsoft Excel documents and Powerpoint lacked sufficient text to assist in latent topic extraction.

5. Corpus Topic Extraction Results.

This section discusses the three tests using weighted regular expression search and search using the Latent Dirichlet Allocation (LDA) topic model. This section first provides initial inferences drawn from the corpus using LDA, then describes the three tests conducted on the data. The first test compares regular expression search and LDA in an information retrieval task for three different topics. The second test uses LDA to automatically extract and dissect topics into subtopics, then discusses the difficulties with using regular expression to accomplish this task. The final test investigates noisy data with high topical overlap and compares LDA versus regular expression search in this domain.

5.1. LDA on Country Corpus

To initially identify topics for testing, LDA is used to extract common topics among various country subsets within the corpus. Table 4 illustrates the results of an LDA analysis on the India corpus. Many topics, including 10-12 in the example, appear related to business. Topic 10 has a project management, personnel, or resumé tone, also mentioning the Indian city of Pune. Topic 11 contains words related to finance or market analysis. Topic 12 also refers to the city of Pune, but appears to focus more on ordering computer systems. Topic 13 and 14 identify very different topics, the first related to science and the second ancient Egyptian gods. The final topic listed only has four files associated with it that include a number of Roman numerals.

Initial analysis of the Indian corpus indicates the main topics center around business and science. The Israeli corpus includes business topics, but also has a large number of Jewish religious and cultural topics. It also includes several unique topics, such as photography, Japanese historical and political analysis, as well as some educational topics.

Table 5 defines a cursory estimate of the major topics that emerge when LDA is performed on the corpus for each country. Most topics from the Mexican word documents are in Spanish and the only English topic appeared to be from software user manuals. Some, like the hard drives labeled from Panama, contained fairly uniform topics related to medical research. A brief analysis of the actual documents indicates significant amount of Arabic letters, hinting that these drives were likely mislabeled and came from Pakistan. The researchers who collected the data were aware of this and had plans to correct the labels.

The inconsistent labels, however, provide a valuable data point for this research. Two LDA results for the disks labeled 'pk' and those labeled 'pa' provide some clues as to how they were collected. While the 'pk' disks include large quantities of English words and mostly business topics, the hard drives labeled 'pa' contains significant amount of Arabic letters. Where there are English words, the topics are mainly focused on medical research and academics. Processing some of the Arabic text through a translator revealed more medical topics. This hints that the drives marked 'pa' were likely drawn from similar businesses while the ones marked 'pk' were drawn from a medical university.

Country corpus with fewer than 200 documents tend to cluster poorly, producing confusing topics. The drives marked 'ua' had a number of documents related to business in the United Arab Emirates from four separate hard drives. A fifth hard drive contains mostly French words. Even though most documents are written in English, while only one was written in French, the models include topics with a mix of English and French. This indicates the model has clustered poorly since different languages should cluster into highly segregated topics. Likely, this was caused by the low number of documents and would benefit from parameter tuning. The Chinese hard drives are mostly Mandarin symbols and out of scope for this research.

5.2. Test 1: Information retrieval

The first test compares information retrieval results using regular expression search versus LDA search. Documents are pruned that have fewer than ten words and words are pruned that have fewer than 10 instances in the corpus. Initial testing demonstrates this caused no decrease in performance of the regular expression search and is required for adequate performance of LDA search.

Since LDA is a stochastic algorithm, results can vary depending on the initial order of words and documents. For this reason, the LDA algorithm is run 30 times and the average result and standard deviation recorded. Posterior probabilities are used to obtain the top 25% of documents most likely to match the keywords. These documents are passed into a second iteration of the LDA algorithm where the process is performed again.

5.2.1. Retrieval topic 1: Passport files

The first search assumes the analyst wants to find files related to passport requests. Four documents have been pre-identified within the Israeli corpus and a list of keywords selected (Table 6) that are similar to what an analyst might choose for this purpose. After pruning, 10,999 documents from 70 disks are available for search. Three of the target documents exist in one disk while the fourth is in another.

Table 4: Sample of LDA Results from India Corpus.

Topic	File Count	Representative Words
10	583	pune project ltd work com date training management pvt experience name software skills
11	249	india years bank company market year financial policy services investment gold term business
12	511	date prices power pune supply may required installation order systems kashyo payment upon
13	21	nanotubes carbon dna memory nram nanotube sequence computer single bit flash new genome
14	8	horus earth god seven two mother egyptian great heaven amenta water human upon
15	4	iii drone vii viii xii queen nic xvii xiii bel analogy ioc polaris patni xvi maruti cipla acc

Table 5: Countries and Major Topics.

Country (code)	Total Word Files	Assessed Topics
Israel (il)	210477	Jewish culture and religion, politics, news
Mexico (mx)	14777	all Spanish words, plus one computer topic
Serbia & Montenegro (rs/cs)	7565	commerce, power generation, health, Serbian words
India (in)	4332	business, science
Singapore (sg)	3833	business
Palestinian State (ps)	1921	computer systems and applications, business
Panama (pa)	1244	medical research
Pakistan (pk)	212	business
Ukraine (ua)	170	UAE business, French words
China (cn)	134	N/A
Unknown (nnn)	50	German words
Egypt (eg)	49	syringe production
Hong Kong (hk)	34	N/A
Turkey (tr)	34	N/A
Ghana (gh)	11	N/A
Canada (ca)	10	N/A
United Arab Emirates (ae)	7	N/A
Japan (jp)	3	N/A

Table 6: Topic 1 keywords.

Keywords	Weight
passport(s|)	3
ambassador	3
embass(y|ies)	2
visa	2
clearance	1
application	1

Table 7: Topic 1 search results - Passport files.

		Regex Search	LDA Iteration			
			1	2	3	4
Document		10999	10999	2506	615	150
Doc 1	μ	11	1459.10	282.33	86.87	9.43
	σ	—	512.03	135.50	57.01	6.44
Doc 2	μ	58	1766.80	242.90	62.80	10.20
	σ	—	1139.85	1687.96	50.98	6.33
Doc 3	μ	53	1859.53	297.10	68.10	10.20
	σ	—	1223.67	145.89	59.94	5.53
Doc 4	μ	N/A	2749.47	459.47	91.33	32.67
	σ	—	1514.97	231.49	50.31	30.24
Prob. Of Loss			0.49	0.37	0.12	0.44

The results in Table 6 and 7 list the keywords and the average position of the four target documents within the query results. Standard deviation indicates the variation between runs of the LDA algorithm and may be an indicator of the difficulty the algorithm has categorizing the document. The final row, 'Prob. of Loss', indicates the likelihood that a document will not survive to the next iteration, assuming only 25% of the documents survive each iteration.

Regular expression keyword search outperforms a single application of the LDA query algorithm by returning the target documents in position 11, 53, and 58 of the query results. The fourth document is pruned from the list as it did not include any of the query words in the text, while with LDA, it has a 0.50 probability of being pruned out on the first iteration. Assuming the fourth document survives to the second iteration, the smaller topical complexity of the data decreased the probability that the document would be pruned to 0.37. The fourth iteration of the LDA algorithm offers more relevant query results than regular expression search. Due to the small number of files remaining, the uncertainty in the fourth document starts to re-emerge with a 0.44 probability of not surviving another iteration.

This test demonstrates an advantages LDA has over keyword search. The fourth document does not include any of the keywords so keyword search failed to identify it. It does, however,

Table 8: Topic 2 keywords - With name.

Keywords	Weight
(masked name)	3
plantiff	2
defendent	2
estate	2
deceased	1
inheritance	1
marriage	1

Table 9: Topic 2 results - Legal documents with name.

		Regex Search	LDA Iteration 1	2	3	4
Document		10999	10999	2506	613	150
Doc 1	μ	153	577.97	235.97	64.30	29.17
	σ	—	325.17	152.90	24.78	12.77
Doc 2	μ	60	852.97	486.60	158.60	55.60
	σ	—	669.63	245.72	64.59	26.66
Doc 3	μ	1	951.80	146.73	49.20	16.43
	σ	—	802.78	195.93	29.45	11.95
Doc 4	μ	4	1409.10	122.20	30.27	14.07
	σ	—	1086.64	162.38	32.70	11.93
Prob. Of Loss			0.11	0.30	0.47	0.77

include words topically similar to the keywords, so LDA has a moderate chance of returning it as a possible document.

5.2.2. Retrieval topic 2: Legal documents

The second topic search attempts to retrieve documents for a specific court case involving a dispute between two families over the will of a deceased member. In this test, the name of the deceased is first included in the keyword search (Table 8), increasing the chance that a keyword search would be successful.

Table 9 provides the results for the regular expression and LDA search using the whole Israel corpus. With the exception of the first document, keyword search returns relevant results. LDA is run for four iterations but fails to provide more relevant results. Additionally, the second document has a moderate to high probability that it will be pruned out.

Table 10 provides the results of running the same test against the Israel corpus, but removes the deceased's name from the keyword search. In this test, keyword search fails to return the first document while LDA returns it each iteration. LDA once again has a moderately high probability of pruning the second document from the list.

The two sets of results in Tables 9 and 11 again demonstrate one of the advantages of LDA over Keyword. Since LDA is a topic model, documents may be returned even if there isn't a single keyword match. All that is required is for the keywords to fall within the same topic as found within the document. One of the drawbacks to the LDA method over keyword search is that training an effective LDA model typically requires word pruning. As the number of documents in the LDA model shrank, topical words disappeared and the first two documents had a high probability of being pruned.

Table 10: Topic 2 keywords - Without name.

Keywords	Weight
plantiff	2
defendent	2
xestate	2
deceased	1
inheritance	1
marriage	1

Table 11: Topic 2 results - Legal documents without name.

		Regex Search	LDA Iteration 1	2	3	4
Document #		10999	10999	2481	614	152
Doc 1	μ	N/A	565.97	217.33	130.60	24.17
	σ	—	337.11	68.91	44.69	22.78
Doc 2	μ	523	975.37	662.77	222.47	31.60
	σ	—	515.47	240.73	92.73	15.87
Doc 3	μ	1	916.97	167.33	51.03	5.70
	σ	—	628.20	82.58	22.35	5.72
Doc 4	μ	7	1360.87	132.73	65.40	7.63
	σ	—	1051.17	48.45	25.44	5.03
Prob. Of Loss		—	0.11	0.57	0.78	0.63

Table 12: Topic 2 results - Legal documents without name, single disk.

		Regex Search	LDA Iteration 1	2	3
Document #		262	262	66	33
Doc 1	μ	20	5.83	18.17	22.70
	σ	—	2.89	15.12	6.71
Doc 2	μ	1	9.23	6.57	15.33
	σ	—	4.78	9.45	5.35
Doc 3	μ	10	3.00	1.53	1.43
	σ	—	2.25	0.50	0.50
Doc 4	μ	2	1.57	1.53	1.57
	σ	—	1.08	0.56	0.50
Prob. Of Loss		—	0.00	0.16	0.84

Table 14: Topic 3 keywords.

Keywords	Weight
electricity	3
power	2
distribut(e|ion)	2
consumption	2
generat(or|ion)	2
network	1

Since all four documents are from the same disk, Table 12 provides the keyword search and LDA results for just the documents on that disk. Using the disk that has the target documents results in far fewer documents than with the entire corpus. This improves the keyword search results, and requires fewer iterations of the LDA algorithm to return good LDA results. On the other hand, the user must search through twenty documents to find all relevant results. On average, using LDA search only requires a search through ten documents after the first iteration. In later iterations, documents 1 and 2 appear to be losing some of the keywords that their ranking relies upon since they are dropping down the list. The final iteration was only pruned 50% since the number of documents was small. Without decreasing the prune rate, document 1 would have had a high chance of being pruned.

With large, noisy data, LDA has trouble maintaining consistent topics for the four documents in question. However, as the corpus is pruned away, topical consistency becomes more distinct. This is illustrated by re-running the second test on the Israel corpus without the name, but only pruning off half of the documents each iteration. Table 13 has the results of the nine iterations, along with the probability of loss for each one. As indicated, probability of loss is close to or essentially zero for every step, demonstrating an improvement at the trade-off of processing time.

5.2.3. Retrieval topic 3: Power generation documents

The third topic search uses the corpus from Serbia (rs) and Montenegro (cs), attempting to locate a set of seven documents describing technical details of electrical power distribution. Within two iterations, LDA places some of the documents higher in the query results than the regular expression query. However, pruning to fewer than 200 documents significantly risks pruning the fourth document. After the third iteration, we decrease the document prune rate from 75% to 50% to avoid pruning relevant documents. Ordinarily, predicting when to change the amount pruned would be difficult. Pruning too many, too early increases the probability of removing relevant documents. Pruning too few increases processing time or false positive rates. Unfortunately, changing the pruning rate only helped the algorithm survive one more. The next iteration is worse with an almost guaranteed chance of losing the first two documents when pruning to 75% and very high probability at 50%. Additionally, the document order changed between iterations 3 and 4. Iteration 3 had low performance with documents 3 and 4 while iteration 4 had low performance with documents 1 and 2. This illustrates how different the models can be with a different set of documents used for training

Table 16 provides the results of running regular expression search and LDA search on only the disk containing the target documents. This disk has the majority of word documents for the Serbian and Montenegro corpus (6,075 on disk versus 6,780 for the entire corpus). With the slightly smaller corpus, keyword search performs only marginally better while LDA search returns query results slightly better. The probability of pruning a relevant document is slightly higher, mainly due to fewer documents overall.

5.2.4. Summary

The three information retrieval topics in this section demonstrate LDA can provide improved query results. In three out of seven tests, LDA returned the worst performing document at a higher position than using regular expression within the first iteration. It achieved this in all seven tests by the fourth iteration, though excessive pruning became a concern. Keyword choice is still important, however, LDA relaxes the requirement to pick at least one keyword in the target document. As long as the words in the keyword match the major topics of the document, it will likely be returned as a query result.

Iterating LDA and pruning results improve document retrieval to a point, though there does appear to be a minimum number of documents required to produce an effective topic model. While the threshold where this happens varies from model to model, it appears to be around 200 to 400 documents. Tables 7 and 9, results started to fail around 150 documents. Table 12 had performance drop between 262 and 66 documents. Table 15 lost performance around 500 documents. Detecting when to stop iterating will vary by data source, however, iterating to this point improves search reliability.

5.3. Test 2: Subtopic Discovery

Occasionally a digital forensics practitioner may not have a specific topic in mind, but wants to get an idea of the overall topics within a disk or corpus. Traditional techniques require manually viewing a sampling of the documents, or selecting a set of potential keywords to isolate expected topics. Delineating between topics within a corpus, on the other hand, is a major strength of LDA. The India corpus has a large number of scientific documents from a variety of disciplines. The tests detailed in this section measures the ability of LDA to extract scientific documents amid other topics. Next, these documents are subdivided into their natural latent sub-topics.

An LDA model is trained on the entire India corpus of Microsoft Word documents and the non-Latin alphabet characters discarded. Documents with fewer than ten words and words with fewer than ten occurrences in the corpus are pruned. The words with the highest posterior probability for each topic are analyzed and an assessment made as to the overall topic. Four topics out of thirty had scientific words defining the topic. Documents that were most likely to belong to one of those four topics were selected and passed into the second run of the LDA algorithm.

Table 13: Topic 2 results - Legal documents without name, 50% retention.

		Regex	LDA Iteration								
		Search	1	2	3	4	5	6	7	8	9
Document #		10999	10999	5476	2745	1375	687	343	172	86	43
Doc 1	μ	N/A	383.33	423.43	213.77	150.63	79.90	68.23	41.57	30.43	19.87
	σ	—	245.00	151.52	76.58	77.44	43.80	31.71	24.06	13.07	0.96
Doc 2	μ	523	1532.17	720.63	323.97	171.73	117.17	58.27	37.17	18.30	15.70
	σ	—	654.14	222.48	163.03	81.83	90.56	28.15	12.20	7.67	2.97
Doc 3	μ	1	389.50	370.40	86.37	44.87	8.90	10.43	7.40	3.00	1.87
	σ	—	249.07	164.33	40.05	26.48	17.99	7.70	5.67	2.39	1.15
Doc 4	μ	7	253.90	258.20	101.27	70.23	16.40	10.30	9.20	4.23	1.97
	σ	—	119.03	106.65	20.50	37.30	16.87	6.88	8.70	5.88	0.80
Prob. Of Loss		—	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.17	0.14

Table 15: Topic 3 results - Power distribution documents.

		Regex	LDA Iteration			
		Search	1	2	3	4
Document #		6,780	6,780	1,668	417	204
Doc 1	μ	70	139.43	97.97	98.77	148.27
	σ	—	73.27	38.82	66.24	19.48
Doc 2	μ	68	180.97	127.23	101.93	148.00
	σ	—	71.55	36.95	63.44	19.45
Doc 3	μ	560	101.67	113.97	95.23	108.47
	σ	—	41.40	37.80	45.20	27.50
Doc 4	μ	225	423.70	343.30	157.20	85.93
	σ	—	95.64	62.49	48.77	24.62
Doc 5	μ	61	394.30	346.80	154.47	85.37
	σ	—	64.39	61.34	48.33	24.93
Doc 6	μ	201	99.67	80.73	79.70	56.57
	σ	—	44.86	44.08	61.41	42.41
Doc 7	μ	179	374.70	281.07	191.27	121.67
	σ	—	74.48	45.85	48.64	30.34
Prob. Of Loss, 75% prune			0.00	0.13	0.86	1.00
Prob. Of Loss, 50% prune			0.00	0.00	0.15	0.99

After a second LDA model was created from the 828 potential scientific documents, an assessment was made as to the subject of each topic and these are given in Table 17. Each document was then assessed to see if it matched the anticipated subject. The 'Total' column provides the total number of documents that were assigned to that topic based on dominant probability. The 'Correct' column provides the number of documents that were assessed to match the anticipated subject. In some cases, two topics were assessed to have the same subject and they are linked via equal values in the 'Group' column. 'Topic Percent' indicates the number correct in the given topic while 'Category Percent' indicates the number correct in that particular group.

Many of the documents were individual pages scanned from an electronics textbook using optical character recognition. The LDA algorithm successfully divided the chapter topics into broad electrical terms (topic 0), electrical chemistry (topic 14 and 23), electromagnetic induction (topic 12) and generators (topic 18). Topic 13 was assessed to be 'Animal Science', but instead described lab procedures that happened to include handling of lab rabbits. Topic 24 was assessed to be articles about color and light but included a number of job performance reviews. Some documents were assessed and the results found to be similar with slight differences. For instance, topic 8 was assessed as chemistry, yet the documents discuss a chemical coatings product.

The test performed in this section demonstrates that LDA can be used successfully to retrieve documents from a broad topic, then that topic divided that into sub-categories. For an analyst conducting e-discovery, topics and sub-topics could be browsed similar to a directory structure. Personal information can be ignored while business or financial documents are divided into sub-categories and explored. While not perfect, LDA correctly categorized documents over 80% of the time. Using the techniques defined in the first test, keywords could then be used to retrieve specific documents from a desired sub-category using LDA models.

Table 16: Topic 3 results - Power distribution documents, single disk.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		6,075	6,075	1,513	387	194
Doc 1	μ	70	141.97	99.50	114.00	124.87
	σ	—	42.87	31.05	85.18	33.75
Doc 2	μ	68	177.23	127.67	116.87	126.80
	σ	—	41.38	24.39	83.15	32.75
Doc 3	μ	539	94.57	106.23	100.47	85.13
	σ	—	17.37	27.89	45.97	15.42
Doc 4	μ	224	357.47	328.13	139.47	122.83
	σ	—	60.87	79.19	44.83	24.51
Doc 5	μ	61	359.10	321.97	138.73	122.30
	σ	—	54.53	64.95	47.30	23.36
Doc 6	μ	200	76.73	78.40	88.80	68.10
	σ	—	26.33	42.55	53.38	20.58
Doc 7	μ	178	315.80	249.90	179.60	129.77
	σ	—	37.54	26.48	57.77	29.14
Prob. Of Loss, 75% prune			0.00	0.27	0.93	0.99
Prob. Of Loss, 50% prune			0.00	0.00	0.40	0.94

Table 17: LDA Science Analysis.

Topic	Estimated Topic	Group	Total	Correct	Topic Percent	Group Percent
0	Electronics	0	34	32	94.12%	94.12%
1	Earthquakes/Structures/Disaster Management	1	6	4	66.67%	—
17		1	6	5	83.33%	75.00%
2	Biology	2	7	6	85.71%	85.71%
4	Nanotubes	3	19	17	89.47%	89.47%
5	Materials/Stress	4	49	48	97.96%	—
22		4	106	106	100.00%	99.35%
6	Navigation	5	9	7	77.78%	77.78%
7	Math	6	33	31	93.94%	93.94%
8	Chemistry	7	11	9	81.82%	—
14		7	20	20	100.00%	93.55%
9	Medical	8	11	6	54.55%	—
26		8	3	3	100.00%	64.29%
10	Vehicle Safety	9	18	17	94.44%	94.44%
11	Environmental	10	5	1	20.00%	—
21		10	8	6	75.00%	—
28		10	15	15	100.00%	78.57%
12	Electrical Power	11	82	82	100.00%	—
18		11	227	227	100.00%	100.00%
13	Animal Science	12	14	3	21.43%	—
15		13	0	0	0%	21.43%
16	Industrial Paints	14	5	5	100.00%	—
29		14	78	78	100.00%	100.00%
19	Construction Science	15	13	13	100.00%	100.00%
23	Batteries	16	22	22	100.00%	100.00%
24	Color/Light	17	18	10	55.56%	55.56%
27	Engineering	18	8	8	100.00%	100.00%
				Total	84.30%	—

5.4. Test 3: Overlapping Topic Analysis

Words can have multiple meanings and the specific intent is often only revealed by examining local context. For example, sub-topics such as disease are relevant to both medicine and biological weapons. This overlap complicates both search by

Table 18: Documents in Water Topic.

Topic	Count
Water	15
Waste Management	15
Economy/Corporation/Trade	9
Environmental	8
Economy (farming)	6
Electricity	5
Agriculture	4
Transportation	3
Global Warming	3
Geography	2
Mining and Minerals	2
Seeds	2
Other	9
Total	83

keyword and LDA since overlapping topics increase the false positive rates. Some of these topical complexities can be revealed by increasing the number of topics, which may cause a broad topic like "outdoor sports" to be categorized into "hunting" and "fishing". A document about fishing, however, may be about recreation or environmental conservation. Posterior probabilities from LDA may reveal both topics, but only if those topics have a large enough representation to be revealed using model parameter estimation techniques.

This first uses a "query by document" approach to detect similar documents. It also examines topical shift as the number of topics is adjusted and compares this shift to existing topical overlap. While simpler than [22] [35], using LDA in this fashion can be used to retrieve relevant documents for tasks where keyword search is ill suited. The LDA model is tested against a regular expression keyword search, where the keywords in Table 18 are identified by performing a subjective analysis of the "known" document's topic. The document chosen for this test is a discussion of government water policy.

The entire Indian country corpus is used to estimate initial LDA model parameters, including the known document about water policy. Initially, 25 topics are assumed. The known document had 74% of its words assigned to a particular topic. Eighty-two other documents also appear in the same topic, so these are used to generate a new LDA model. These documents are assigned labels by conducting a manual document survey and are listed in Table 18. The labels in this topic tend towards environmental management and agriculture, with a few single topic documents in the 'Other' category. Still, a moderately large number of documents remain unrelated to water.

Table 18 presents the topics identified in the 82 document LDA sub-model. Table 20 provides some information about the 15 documents from the water topic and their topical clustering as the number of topics is changed from 25 to 15 and 10. The document labeled 'C8' is our query document about government water policy and documents 'C9' and 'C10' also discuss government water policy. As the number of topics in the model is adjusted, some topics combine, illustrating overlap. The most relevant 'C' series documents, however, remain assigned to the same topic.

Table 19: Water document keywords.

Keywords	Weight
water	3
water resource(s|)	3
united nations	2
international	1
management	1

Table 20: LDA Clustering of Water Topic.

	Topics			Keyword
Document	25	15	10	Order
A1	7	1	6	22
B2	7	1	2	3
B3	7	1	2	20
B4	7	1	2	19
B5	7	1	2	21
B6	7	1	2	8
B7	7	1	2	6
C8	2	3	5	2
C9	2	3	5	1
C10	2	3	5	10
D11	2	9	14	7
E12	20	9	6	24
E13	20	9	6	30
F14	20	8	6	49
F15	20	8	6	23
X1	–	–	–	28

The other document labels have been assigned to link similar documents with the same letter. Document 'A1' has similar topics as the documents marked 'B' that primarily discuss water pollution. The 'E' and 'F' series documents are scientific documents discussing the removal of impurities from water and this explains the link with document 'A1' about pollution. Document 'D11' is a scientific analysis of irrigation techniques, explaining the link between that and the 'E' series with 15 topics, but it also discusses governmental water policy as found in the 'C' series documents.

Running a regular expression query using the keywords defined in Table 19 returns 73 documents with their positions indicated by the last column in Table 20. The regular expression query returns the same documents as LDA, and included one additional document that LDA did not find. It did, however, produce 58 more false positives that must be manually inspected. The query document is returned as the second document with the other two target documents returned first and tenth. The fourth, fifth and ninth document are unrelated to water.

In this test, LDA used a "query by document" approach, where for a keyword search the practitioner must reduce the document into a keyword set. The first iteration of LDA identified a topic containing a number of water documents along with documents in fields related to water, such as agriculture

and waste management. These topics are further broken up into sub-topics containing some overlap. LDA was able to consistently group the three relevant documents where keyword search included a number of unrelated documents within the set bounded by the relevant documents.

6. Discussion and Conclusion

Traditional keyword or regular expression search is being overwhelmed by rapidly increasing quantities of data. Analyst time can be the limiting factor in an investigation and automated techniques that improve analyst efficiency are needed. Latent Dirichlet Allocation (LDA) provides one possible technique to help filter noise and isolate relevant material. This work evaluated three primary tasks and methods for incorporating LDA effectively and compared these results to regular expression search. LDA was able to produce results with higher relevance within one iteration in almost half the tests and all tests within four iterations. It was also able to return relevant documents even if no keywords were present in the document. On the other hand, LDA was much slower. Where a regular expression search on the Israeli corpus took approximately one minute, four iterations of the LDA algorithm took over eight hours. While pruning using LDA often resulted in higher precision due to smaller query results, it occasionally produced false negatives by pruning too much.

The second and third tests highlight the advantages of LDA. The second test demonstrated topic browsing using a corpus-trained LDA model. LDA was used to extract all scientific documents from the corpus simply based on dominant topic keywords. This set was then sub-divided into its various disciplines. Regular expression keyword search methods require far more steps to accomplish the same task. The final test used LDA to analyze overlapping topics and demonstrated how LDA can help further separate the overlapping topics.

There are a number of potential areas for further research. First, the corpus itself provides data for research in language processing, syntactical analysis, image processing, sound file analysis, and cross-domain techniques, among many others. Due to the manpower required to develop recall information within the RDC, this paper had to limit its analysis to precision measures. Manually categorizing the documents within the corpus would determine how many false negatives are produced by document query algorithms. Additionally, many tests required subjective analysis about document topics and could be improved by using a survey of varying opinions on document topics. Amazon Mechanical Turk has been successfully employed for this purpose [30], though due to privacy concerns, would not be possible with the RDC. LDA and regular expression search each has strengths that could be combined into a hybrid technique. Finally, while this paper briefly examined image and text, a multi-modal technique such as that found in [29] may improve latent topic extraction and add a viable automated image annotation technique.

Acknowledgment

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

We would like to thank Dr. Garfinkel and his research group for the providing the data and valuable assistance with the research.

References

[1] OpenCV Computer Vision Library, C++ Library, 2011.

[2] U.S. Patent Office Data Dump, 2011.

[3] N.O. Andrews, E.A. Fox, Recent developments in document clustering, Technical Report TR-07-35, Virginia Tech, 2007.

[4] N. Beebe, G. Dietrich, A new process model for text string searches, in: Advances in Digital Forensics III, Springer, 2007, pp. 179–191.

[5] N.L. Beebe, J.G. Clark, G.B. Dietrich, M.S. Ko, D. Ko, Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies, Decision Support Systems 51 (2011) 732–744.

[6] D. Blei, J. McAuliffe, Supervised topic models, in: Conference on Neural Information Processing Systems, MIT Press, 2007, pp. 1–8.

[7] D.M. Blei, T.L. Griffiths, M.I. JOrdan, J.B. Tenenbaum, Hierarchical topic models and the nested chinese restaurant process, in: Advances in Neural Information Processing Systems, MIT Press, 2004, pp. 17–24.

[8] D.M. Blei, M.I. Jordan, Modeling annotated data, in: Special Interest Group on Information Retrieval, ACM, 2003, pp. 127–134.

[9] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[10] A.J.B. Chaney, D.M. Blei, Visualizing topic models, in: Sixth International AAAI Conference on Weblogs and Social Media, Springer, 2012, pp. 419–422.

[11] Y. Chen, L. Wang, M. Dong, J. Hua, Exemplar-based visualization of large document corpus, in: IEEE Transactions on Visualization and Computer Graphics, volume 15, IEEE, 2009, pp. 1161–1168.

[12] J.P. Craiger, Handbook of technology management, Handbook of Technology Management, John Wiley & Sons, 2010, pp. 921–930.

[13] L. Du, H. Jin, O. de Vel, N. Liu, A latent semantic indexing and wordnet based information retrieval model for digital forensics, in: IEEE International Conference on Intelligence and Security Informatics, IEEE, 2008, pp. 70 – 75.

[14] R. Fagin, Y.S. Maarek, Allowing users to weight search terms, in: Proceedings of the Computer-Assisted Information Retrieval, pp. 682–700.

[15] K. Farrahi, D. Gatica-Perez, Learning and predicting multimodal daily life patterns from cell phones, in: Proceedings of the 2009 International Conference on Multimodal Interfaces, ACM, 2009, pp. 277–280.

[16] S. Feldman, C. Sherman, The high cost of not finding information, White Paper, International Data Corporation, 2003.

[17] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59 (2004) 167–181.

[18] G. Furnas, T. Landauer, L. Gomez, S. Dumais, The vocabulary problem in human-system communications, Communications of the ACM 30 (1987) 964–971.

[19] S. Garfinkel, P. Farrell, V. Roussev, G. Dinolt, Bringing science to digital forensics with standardized forensic corpora, Digital Investigation 6 (2009) S2–S11.

[20] S.L. Garfinkel, Forensics feature extraction and cross-drive analysis, Digital Investigations 3 (2006) 71–81.

[21] D.B. Garrie, Effective keyword selection requires a mastery of storage technology and the law, Pace Law Review 32 (2012) 399–406.

[22] M.R. Gormley, M. Dredze, B.V. Durme, J. Eisner, Shared components topic models, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational LOinguistics, 2012, pp. 783–792.

[23] T. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences, PNAS, 2004, pp. 5228–5235.

[24] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.
[25] T. Larson, The other side of civil discovery, in: E. Casey (Ed.), Handbook of Computer Crime Investigation, Academic Press, 2002, pp. 17–52.
[26] L. Mueller, Crafting good keywords in EnCase and using conditions to refine results, ForensicKB, 2013.
[27] K. Nance, B. Hay, M. Bishop, Digital forensics: Defining a research agenda, in: Proceedings of the 42nd Hawaii International Conference on System Sciences, IEEE, 2009, pp. 1–6.
[28] J. Napolitano, The future of science as public service, Speech presented at the Massachusetts Institute of Technology, 2011.
[29] G.E. Noel, G.L. Peterson, Improving Document Retrieval in Multi-Modal LDA using Image-Derived Super-Words, Technical Report, Air Force Institute of Technology, 2012. Submitted to Information Sciences Journal, Dec 2012.
[30] G.E. Noel, G.L. Peterson, Context-based image annotation using imagenet, in: 26th International Florida Artificial Intelligence Research Society Conference, AAAI, 2013, pp. 462–467.
[31] A. Perotte, N. Bartlett, N. Elhadad, F. Wood, Hierarchically supervised latent dirichlet allocation, in: Twenty-Fifth Annual Conference on Neural Information Processing Systems, MIT Press, 2011, pp. 12–15.
[32] D. Quercia, H. Askham, J. Crowcroft, TweetLDA: Supervised topic classification and link prediction in twitter, in: Proceedings of the 4th ACM International Conference on Web Science, ACM, 2012, pp. 247–250.
[33] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523.
[34] G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing, Information Retrieval and Language Processing 18 (1975) 613–620.
[35] M.M. Shafiei, E.E. Milios, Latent dirichlet co-clustering, in: Proceedings of the Sixth International Conference on Data Mining, IEEE, 2006, pp. 542–551.
[36] P. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2001, pp. 1511–1518.
[37] G. Wang, D. Hoiem, D. Forsyth, Building text features for object image classifications, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1367–1374.

George E. Noel is currently working towards the PhD degree in Computer Science at the Air Force Institute of Technology. He received a BS degree in Computer Science from the United States Air Force Academy and an MS degree in Information Resource Management from the Air Force Institute of Technology. His interests are in Content-Based Image Retrieval, data mining, and statistical machine learning.

Gilbert L. Peterson is an Associate Professor of Computer Science at the Air Force Institute of Technology, and Vice-Chair of the IFIP Working Group 11.9 Digital Forensics. Dr. Peterson received a BS degree in Architecture, and an M.S and Ph.D in Computer Science at the University of Texascat Arlington. He teaches and conducts research in digital forensics, statistical machine learning, and autonomous robots.