Springer Lecture Notes in Computer Science (LNCS) series -- IEEE International Conference on Intelligence and Security Informatics.

Analyzing Entities and Topics in News Articles using Statistical Topic Models

David Newman[1], Chaitanya Chemudugunta[1], Padhraic Smyth[1], and Mark Steyvers[2]

[1] Department of Computer Science,
UC Irvine, Irvine, CA
{newman, chandra, smyth}@uci.edu
[2] Department of Cognitive Science,
UC Irvine, Irvine, CA
msteyver@uci.edu

Abstract. Statistical language models can learn relationships between topics discussed in a document collection and persons, organizations and places mentioned in each document. We present a novel combination of statistical topic models and named-entity recognizers to jointly analyze entities mentioned (persons, organizations and places) and topics discussed in a collection of 330,000 New York Times news articles. We demonstrate an analytic framework which automatically extracts from a large collection: topics; topic trends; topics associated with single entities; and topics that relate entities.

1 Introduction

The ability to rapidly analyze and understand large sets of text documents is a challenge across many disciplines. Consider the problem of being given a large set of emails, reports, technical papers, news articles, and wanting to quickly gain an understanding of the key information contained in this set of documents. For example, lawyers frequently need to analyze the contents of very large volumes of evidence in the form of text documents during the discovery process in legal cases (e.g., the 250,000 Enron emails that were made available to the US Justice Department [1]). Similarly, intelligence analysts are faced on a daily basis with vast databases of intelligence reports from which they would like to quickly extract useful information.

There is increasing interest in text mining techniques to solve these types of problems. Supervised learning techniques classify objects such as documents into predefined classes [2]. While this is useful in certain problems, in many applications there is relatively little knowledge *a priori* about what the documents may contain.

Unsupervised learning techniques can extract information from sets of documents without using predefined categories. Clustering is widely used to group documents into K clusters, where the characteristics of the clusters are determined in a data-driven fashion [3]. By representing each document as a vector

of word or term counts (the "bag of words" representation), standard vector-based clustering techniques can be used, where the learned cluster centers represent "prototype documents." Another set of popular unsupervised learning techniques for document collections are based on matrix approximation methods, i.e. singular value decomposition (or principal components analysis) of the document-word count matrix [4, 5]. This approach is often referred to as latent semantic indexing (LSI).

While both clustering and LSI have yielded useful results in terms of reducing large document collections to lower-dimensional summaries, they each have their limitations. For example, consider a completely artificial data set where we have three types of documents with equal numbers of each type: in the first type each document contains only two words *wordA, wordB*, in the second type each document contains only the words *wordC, wordD*, and the third type contains a mixture, namely the words *wordA, wordB, wordC, wordD*.

LSI applied to this toy data produces two latent topic vectors with orthogonal "directions": *wordA + wordB + wordC + wordD* and *wordA + wordB - wordC - wordD*. These two vectors do not capture the fact that each of the three types of documents are mixtures of two underlying "topics", namely *wordA + wordB* and *wordC + wordD*. This limitation is partly a reflection of the fact that LSI must use negative values in its basis vectors to represent the data, which is inappropriate given that the underlying document-word vectors (that we are representing in a lower-dimensional space) consist of non-negative counts. In contrast, as we will see later in the paper, probabilistic representations do not have this problem. In this example the topic model would represent the two topics *wordA + wordB* and *wordC + wordD* as two multinomial probability distributions, $[\frac{1}{2}, \frac{1}{2}, 0, 0]$ and $[0, 0, \frac{1}{2}, \frac{1}{2}]$, over the four-word vocabulary, capturing the two underlying topics in a natural manner. Furthermore, the topic model would correctly estimate that these two topics are used equally often in the data set, while LSI would estimate that its first topic direction is used approximately twice as much as its second topic direction.

Document clustering techniques, such as k-means and agglomerative clustering, suffer from a different problem, that of being forced to assume that each document belongs to a single cluster. If we apply the k-means algorithm to the toy data above, with $K = 2$ clusters it will typically find one cluster to be centered at $[1, 1, 0, 0]$ and the other at $[\frac{1}{2}, \frac{1}{2}, 1, 1]$. It is unable to capture the fact that there are two underlying topics, corresponding to *wordA + wordB* and *wordC + wordD* and that that documents of type 3 are a combination of these two topics.

A more realistic example of this effect is shown in Table 1. We applied k-means clustering with $K = 100$ and probabilistic topic models (to be described in the next section) with $T = 100$ topics to a set of 1740 papers from 12 years of the Neural Information Processing (NIPS) Conference[3]. This data set contains a total of $N = 2,000,000$ word tokens and a vocabulary size of $W = 13,000$ unique words. Table 1 illustrates how two different papers were interpreted by both the cluster model and the probabilistic topic model. The first paper dis-

[3] Available on-line at http://www.cs.toronto.edu/~roweis/data.html

cussed an analog circuit model for auditory signal processing—in essence it is a combination of a topic on circuits and a topic on auditory modeling. The paper is assigned to a cluster which fails to capture either topic very well—shown are the most likely words in the cluster it was assigned to. The topic model on the other hand represents the paper as a mixture of topics. The topics are quite distinct (the highest probability words in each topic are shown), capturing the fact that the paper is indeed a mixture of different topics. Similarly, the second paper was an early paper in bioinformatics, again combining topics that are somewhat different, such as protein modeling and hidden Markov models. Again the topic model can separate out these two underlying topics, whereas the clustering approach assigns the paper to a cluster that is somewhat mixed in terms of concepts and that does not summarize the semantic content of the paper very well.

The focus of this paper is to extend our line of research in probabilistic topic modeling to analyze persons, organizations and places. By combining named entity recognizers with topic models we illustrate how we can analyze the relationships between these entities (persons, organizations, places) and topics, using a large collection of news articles.

Table 1. Comparison of topic modeling and clustering.

Abstract from Paper	Topic Mix	Cluster Assignment
Temporal Adaptation in a Silicon Auditory Nerve (**J Lazzaro**) Many auditory theorists consider the temporal adaptation of the auditory nerve a key aspect of speech coding in the auditory periphery. Experiments with models of auditory localization ... I have designed an analog integrated circuit that models many aspects of auditory nerve response, including temporal adaptation.	**[topic 80]** analog circuit chip current voltage vlsi figure circuits pulse synapse silicon implementation cmos output mead hardware design **[topic 33]** auditory sound localization cochlear sounds owl cochlea song response system source channels analysis location delay	**[cluster 8]** circuit figure time input output neural analog neuron chip system voltage current pulse signal circuits networks response systems data vlsi
Hidden Markov Models in Molecular Biology: New Algorithms and Applications (**P Baldi, Y Chauvin, T Hunkapiller, M McClure**) Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs ... that are trained to represent several protein families including immunoglobulins and kinases.	**[topic 10]** state hmm markov sequence models hidden states probabilities sequences parameters transition probability training hmms hybrid model likelihood modeling **[topic 37]** genetic structure chain protein population region algorithms human mouse selection fitness proteins search evolution generation function sequence sequences genes	**[cluster 88]** model data models time neural figure state learning set parameters network probability number networks training function system algorithm hidden markov

2 A Brief Review of Statistical Topic Models

The key ideas in a statistical topic model are quite simple and are based on a probabilistic model for each document in a collection. A topic is a multinomial probability distribution over the V unique words in the vocabulary of the corpus, in essence a V-sided die from which we can generate (in a memoryless fashion) a "bag of words" or a set of word counts for a document. Thus, each topic t is a probability vector, $p(w|t) = [p(w_1|t), \ldots, p(w_V|t)]$, where $\sum_v p(w_v|t) = 1$, and there are T topics in total, $1 \leq t \leq T$.

A document is represented as a finite mixture of the T topics. Each document d, $1 \leq d \leq N$, is assumed to have its own set of mixture coefficients, $[p(t = 1|d), \ldots, p(t = T|d)]$, a multinomial probability vector such that $\sum_t p(t|d) = 1$. Thus, a randomly selected word from document d has a conditional distribution $p(w|d)$ that is a mixture over topics, where each topic is a multinomial over words:

$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d).$$

If we were to simulate W words for document d using this model we would repeat the following pair of operations W times: first, sample a topic t according to the distribution $p(t|d)$, and then sample a word w according to the distribution $p(w|t)$.

Given this forward or generative model for a set of documents, the next step is to learn the topic-word and document-topic distributions given observed data. There has been considerable progress on learning algorithms for these types of models in recent years. Hofmann [6] proposed an EM algorithm for learning in this context using the name "probabilistic LSI" or pLSI. Blei, Ng and Jordan [7] addressed some of the limitations of the pLSI approach (such as the tendency to overfit) and recast the model and learning framework in a more general Bayesian setting. This framework is called Latent Dirichlet allocation (LDA), essentially a Bayesian version of the model described above, and the accompanying learning algorithm is based on an approximation technique known as variational learning. An alternative, and efficient, estimation algorithm based on Gibbs sampling was proposed by Griffiths and Steyvers [9], a technique that is closely related to earlier ideas derived independently for mixture models in statistical genetics [10]. Since the Griffiths and Steyvers paper was published in 2004, a number of different groups have successfully applied the topic model with Gibbs sampling to a variety of large corpora, including large collections of Web documents [11], a collection of 250,000 Enron emails [12], 160,000 abstracts from the CiteSeer computer science collection [13], and 80,000 news articles from the 18th-century Pennsylvania Gazette [14]. A variety of extensions to the basic topic model have also been developed, including author-topic models [15], author-role-topic models [12], topic models for images and text [16, 7], and hidden-Markov topic models for separating semantic and syntactic topics [17].

In this paper all of the results reported were obtained using the topic model outlined above with Gibbs sampling, as described originally in [9]. Our descrip-

tion of the model and the learning algorithm is necessarily brief: for a more detailed tutorial introduction the reader is recommended to consult [18].

3 Data Set

To analyze entities and topics, we required a text dataset that was rich in entities including persons, organizations and locations. News articles are ideal because they have the primary purpose of conveying information about who, what, when and where. We used a collection of New York Times news articles taken from the Linguistic Data Consortium's English Gigaword Second Edition corpus (www.ldc.upenn.edu). We used all articles of type "story" from 2000 through 2002, resulting in 330,000 separate articles spanning three years. These include articles from the NY Times daily newspaper publication as well as a sample of news from other urban and regional US newspapers.

We automatically extracted named entities (i.e. proper nouns) from each article using one of several named entity recognition tools. We evaluated two tools including GATE's Information Extraction system ANNIE (gate.ac.uk), and Coburn's Perl Tagger (search.cpan.org/ acoburn/Lingua-EN-Tagger). ANNIE is rules-based and makes extensive use of gazetteers, while Coburn's tagger is based on Brill's HMM part-of-speech tagger [19]. ANNIE tends to be more conservative in identifying a proper noun. For this paper, entities were extracted using Coburn's tagger. For this 2000-2002 period, the most frequently mentioned people were: George Bush; Al Gore; Bill Clinton; Yasser Arafat; Dick Cheney and John McCain. In total, more than 100,000 unique persons, organizations and locations were extracted. We filtered out 40,000 infrequently occurring entities by requiring that an entity occur in at least ten different news articles, leaving 60,000 entities in the dataset.

After tokenization and removal of stopwords, the vocabulary of unique words was also filtered by requiring that a word occur in at least ten different news articles. We produced a final dataset containing 330,000 documents, a vocabulary of 40,000 unique words, a list of 60,000 entities, and a total of 110 million word tokens. After this processing, entities occur at the rate of 1 in 6 words (not counting stopwords).

4 Experiments

In this section we present the results from a $T = 400$ topic model run on the three years of NY Times news articles. After showing some topics and topic trends, we show how the model reveals topical information about particular entities, and relationships between entities. Note that entities are just treated as regular words in the learning of the topic models, and the topic-word distributions are separated out into entity and non-entity components as a postprocessing step. Models that treat entity and non-entity words differently are also of interest, but are beyond the scope of this paper.

4.1 Topics and Topic Trends

Upon completion of a topic model run, the model saves data to compute the likelihood of words and entities in a topic, $p(w|t)$ and $p(e|t)$, the mix of topics in each document, $p(t|d)$, and z_i the topic assigned to the i^{th} word in the corpus.

For each topic, we print out the most likely words and most likely entities. We then review the list of words and entities to come up with a human-assigned topic label that best summarizes or captures the nature of the topic. It is important to point out that these topic labels are created *after* the model is run; they are not *a priori* defined as fixed or static subject headings.

Unsurprisingly, our three-years of NY Times includes a wide range of topics: from renting apartments in Brooklyn to diving in Hawaii; from Tiger Woods to PETA liberating tigers; from voting irregularities to dinosaur bones. From a total of 400 diverse topics, we selected a few to highlight. Figure 1 shows four seasonal topics which we labeled Basketball, Tour de France, Holidays and Oscars. Each of these topics shows a neat division within the topic of *what* (the words in lowercase), and *who* and *where* (the entities in uppercase). The Basketball topic appears to focus on the Lakers; the Tour de France topic tell us that it's all about Lance Armstrong; Barbie trumps the Grinch in Holidays; and Denzel Washington most likely had a good three years in 2000-2002.

Figure 2 shows four "event" topics which we labeled September 11 Attacks, FBI Investigation, Harry Potter/Lord of the Rings, and DC Sniper. This Sept. 11 topic – one of several topics that discuss the terrorist attacks on Sept 11 – is clearly about the breaking news. It discusses what and where, but not who (i.e. no mention of Bin Laden). The FBI Investigation topic lists 9/11 hijackers Mohamed Atta and Hani Hanjour. The Harry Potter/Lord of the Rings topic combines these same-genre runaway successes, and the DC Sniper topic shows specific details about John Muhammad and Lee Malvo including that they were in a white van.

What year had the most discussion of the Tour de France? Is interest in football declining? What was the lifetime of Elian Gonzalez story? These questions can be answered by examining the time trends in the topics. These trends are easily computed by counting the the topic assignments z_i of each word in each time period (monthly). Figure 3 uses the topics already presented plus additional topics to show some seasonal/periodic time trends and event time trends. We see from the trends on the left that Basketball gets 30,000 in May; discussions of football are increasing; 2001 was a relatively quiet year for the Oscars; but 2001 had the most buzz over quarterly earnings. The trends on the right on the other hand shows some very peaked events: from Elian Gonzalez in April 2000; thru the September 11 Attacks in 2001; to the DC Sniper killing spree and the collapse of Enron in 2002.

4.2 Topics Associated with Entities

The topic model can also infer the likelihood of a particular topic given an entity. Table 2 shows the most likely topics associated with frequently discussed people

Basketball	
team	0.028
play	0.015
game	0.013
season	0.012
final	0.011
games	0.011
point	0.011
series	0.011
player	0.010
coach	0.009
playoff	0.009
championship	0.007
playing	0.006
win	0.006
LAKERS	0.062
SHAQUILLE-O-NEAL	0.028
KOBE-BRYANT	0.028
PHIL-JACKSON	0.019
NBA	0.013
SACRAMENTO	0.007
RICK-FOX	0.007
PORTLAND	0.006
ROBERT-HORRY	0.006
DEREK-FISHER	0.006

Tour de France	
tour	0.039
rider	0.029
riding	0.017
bike	0.016
team	0.016
stage	0.014
race	0.013
won	0.012
bicycle	0.010
road	0.009
hour	0.009
scooter	0.008
mountain	0.008
place	0.008
LANCE-ARMSTRONG	0.021
FRANCE	0.011
JAN-ULLRICH	0.003
LANCE	0.003
U-S-POSTAL-SERVICE	0.002
MARCO-PANTANI	0.002
PARIS	0.002
ALPS	0.002
PYRENEES	0.001
SPAIN	0.001

Holidays	
holiday	0.071
gift	0.050
toy	0.023
season	0.019
doll	0.014
tree	0.011
present	0.008
giving	0.008
special	0.007
shopping	0.007
family	0.007
celebration	0.007
card	0.007
tradition	0.006
CHRISTMAS	0.058
THANKSGIVING	0.018
SANTA-CLAUS	0.009
BARBIE	0.004
HANUKKAH	0.003
MATTEL	0.003
GRINCH	0.003
HALLMARK	0.002
EASTER	0.002
HASBRO	0.002

Oscars	
award	0.026
film	0.020
actor	0.020
nomination	0.019
movie	0.015
actress	0.011
won	0.011
director	0.010
nominated	0.010
supporting	0.010
winner	0.008
picture	0.008
performance	0.007
nominees	0.007
OSCAR	0.035
ACADEMY	0.020
HOLLYWOOD	0.009
DENZEL-WASHINGTON	0.006
JULIA-ROBERT	0.005
RUSSELL-CROWE	0.005
TOM-HANK	0.005
STEVEN-SODERBERGH	0.004
ERIN-BROCKOVICH	0.003
KEVIN-SPACEY	0.003

Fig. 1. Selected seasonal topics from a 400-topic run of the NY Times dataset. In each topic we first list the most likely words in the topic, with their probability, and then the most likely entities (in uppercase). The title above each box is a human-assigned topic label.

September 11 Attacks	
attack	0.033
tower	0.025
firefighter	0.020
building	0.018
worker	0.013
terrorist	0.012
victim	0.012
rescue	0.012
floor	0.011
site	0.009
disaster	0.008
twin	0.008
ground	0.008
center	0.008
fire	0.007
plane	0.007
WORLD-TRADE-CTR	0.035
NEW-YORK-CITY	0.020
LOWER-MANHATTAN	0.005
PENTAGON	0.005
PORT-AUTHORITY	0.003
RED-CROSS	0.002
NEW-JERSEY	0.002
RUDOLPH-GIULIANI	0.002
PENNSYLVANIA	0.002
CANTOR-FITZGERALD	0.001

FBI Investigation	
agent	0.029
investigator	0.028
official	0.027
authorities	0.021
enforcement	0.018
investigation	0.017
suspect	0.015
found	0.014
police	0.014
arrested	0.012
search	0.012
law	0.011
arrest	0.011
case	0.010
evidence	0.009
suspected	0.008
FBI	0.034
MOHAMED-ATTA	0.003
FEDERAL-BUREAU	0.001
HANI-HANJOUR	0.001
ASSOCIATED-PRESS	0.001
SAN-DIEGO	0.001
U-S	0.001
FLORIDA	0.001

Harry Potter/Lord Rings	
ring	0.050
book	0.015
magic	0.011
series	0.007
wizard	0.007
read	0.007
friend	0.006
movie	0.006
children	0.006
part	0.005
secret	0.005
magical	0.005
kid	0.005
fantasy	0.005
fan	0.004
character	0.004
HARRY-POTTER	0.024
LORD OF THE RING	0.013
STONE	0.007
FELLOWSHIP	0.005
CHAMBER	0.005
SORCERER	0.004
PETER-JACKSON	0.004
J-K-ROWLING	0.004
TOLKIEN	0.004
HOGWART	0.002

DC Sniper	
sniper	0.024
shooting	0.019
area	0.010
shot	0.009
police	0.007
killer	0.006
scene	0.006
white	0.006
victim	0.006
attack	0.005
case	0.005
left	0.005
public	0.005
suspect	0.005
killed	0.005
car	0.005
WASHINGTON	0.053
VIRGINIA	0.019
MARYLAND	0.013
D-C	0.012
JOHN-MUHAMMAD	0.008
BALTIMORE	0.006
RICHMOND	0.006
MONTGOMERY-CO	0.005
MALVO	0.005
ALEXANDRIA	0.003

Fig. 2. Selected event topics from a 400-topic run of the NY Times dataset. In each topic we first list the most likely words in the topic, with their probability, and then the most likely entities (in uppercase). The title above each box is a human-assigned topic label.

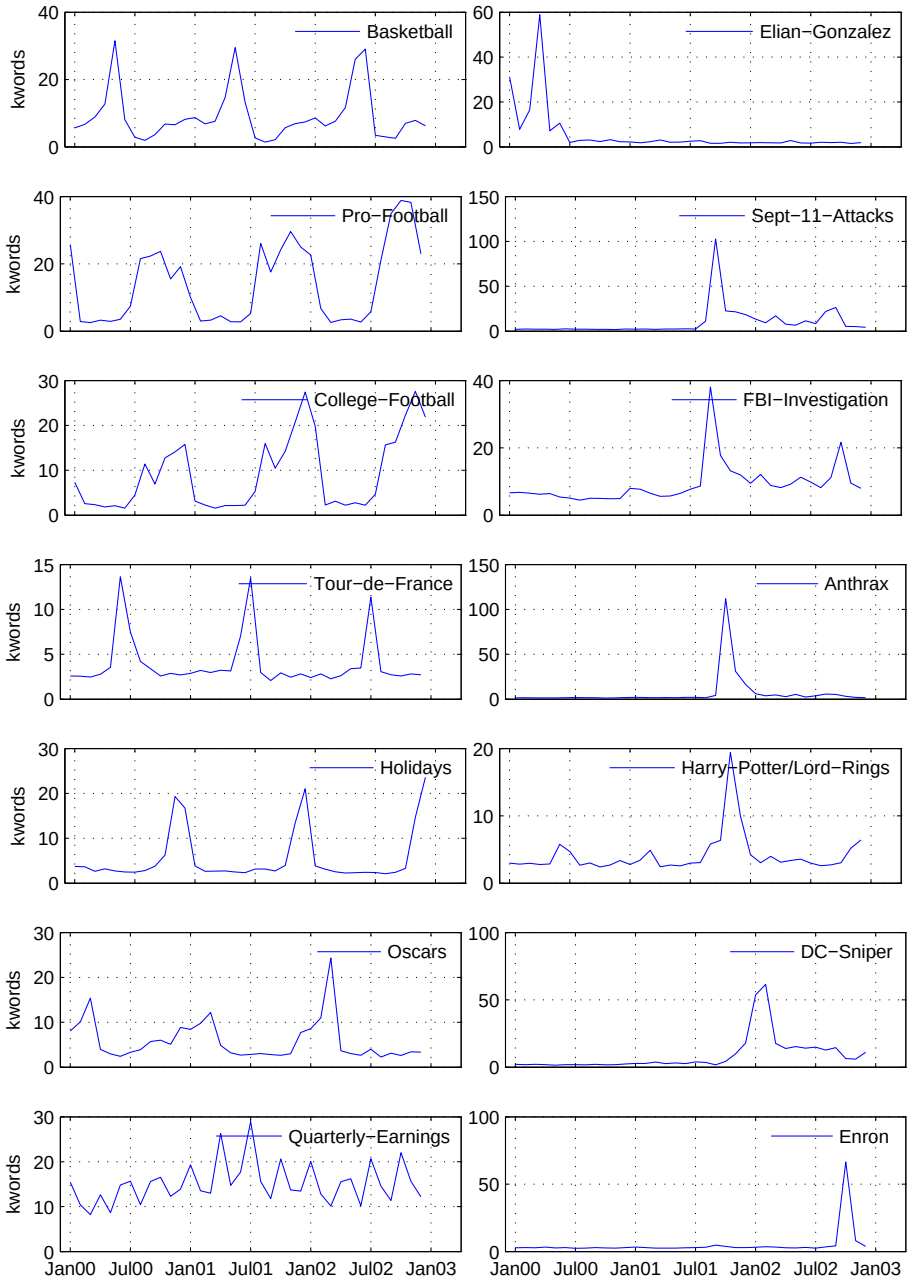

Fig. 3. Selected topic-trends from a 400-topic run of the NY Times dataset. Seasonal/periodic topics are shown on the left, and event topics are shown on the right. Each curve shows the number of words (in thousands) assigned to that topic for the month (on average there are 9,000 articles written per month containing 3 million words, so if the 400 topics were equally likely there would be 8 kwords per topic per month). The topic words and entities for Basketball, Tour de France, Holidays and Oscars are given in Figure 1, and Sept 11 Attacks, FBI Investigation, Harry Potter/Lord of the Rings and DC Sniper are given in Figure 2.

Table 2. Frequently discussed people and their associated topics from the NY Times dataset. Next to each person we list the most likely topics associated with that person.

Entity	Most likely topics associated with entity
Yasser Arafat	Mid-East Conflict (81%), Mid-East Peace Effort (18%), Attack/Bombing (1%)
Tony Blair	N. Ireland Peace (84%), Iraq War (5%), European Union (3%)
George Bush	White House Policy (43%), White House Staff (42%), Election (4%), Primaries (2%)
Fidel Castro	Embargo (62%), Elian Gonzalez (34%), Latin America (4%)
Bill Clinton	Pres. Scandal (61%), Election (9%), White House Staff (6%), Peace Effort (3%)
Vicente Fox	Mexico (93%), Illegal Immigration (6%)
Al Gore	Election (86%), Recount (13%), Primaries (1%)
Saddam Hussein	Iraq War (99%)
Osama Bin Laden	Sept 11 Attack (76%), Afghanistan/Taliban (12%), Muslim/Arab (12%)
Slobodan Milosevic	War Crimes (100%)
Valdimir Putin	Russia/Soviet (96%), Missile Defense System (4%)
Ariel Sharon	Mid-East Conflict (79%), Mid-East Peace Effort (20%)

in our 2000-2002 NY Times dataset. Some people (e.g. Bush, Clinton) come up in a variety of topics, while others (Hussein, Milosevic) are associated with just one topic. Arafat and Sharon are associated with the Mid-East Conflict and Mid-East Peace Effort in the same 80/20 proportion.

4.3 Entity-Entity Relationships

We use the topic model to determine topic-based entity-entity relationships. Unlike social networks created from co-mentions – which would not link two entities that were never co-mentioned – our topic-based approach can potentially link a pair of entities that were never co-mentioned. A link is created when the entity-entity "affinity", defined as $(p(e_1|e_2) + p(e_2|e_1))/2$, is above some threshold. The graph in Figure 4, constructed using this entity-entity affinity was created in two steps. First we selected key entities (e.g. Yasser Arafat, Saddam Hussien, Osama Bin Laden, Zacarias Moussaoui, Vladamir Putin, Ariel Sharon, The Taliban) and determined what other entities had some level of affinity to these. We then took this larger list of 100 entities, computed all 10,000 entity-entity affinities, and thresholded the result to produce the graph. It is possible to annotate each link with the topics that most contribute to the relationship, and beyond that, the original documents that most contribute to that topic.

A related but slightly different representation is shown in the bipartite graph showing relationships between entities and topics in Figure 5. A link is present when the likelihood of an entity in a particular topic $p(e|t)$, is above a threshold. This graph was created by selecting 15 entities from the graph shown in Figure 4 and computing all 15×400 entity-given-topic probabilities, and thresholding the result to plot links. Again, with this bipartite graph, the original documents associated with each topic can be retrieved.

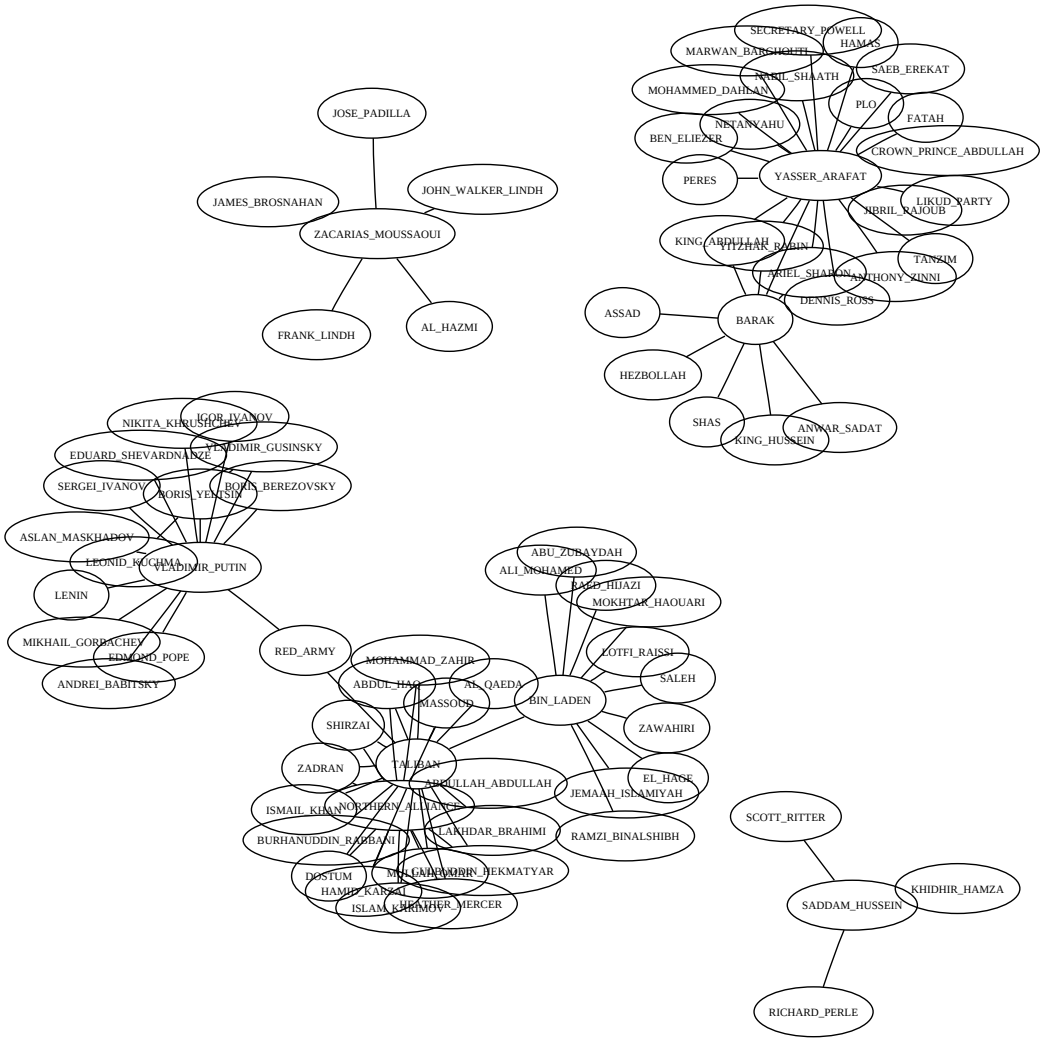

Fig. 4. Social network showing topic-model-based relationships between entities. A link is present when the entity-entity "affinity" $(p(e_1|e_2) + p(e_2|e_1))/2$ is above a threshold.

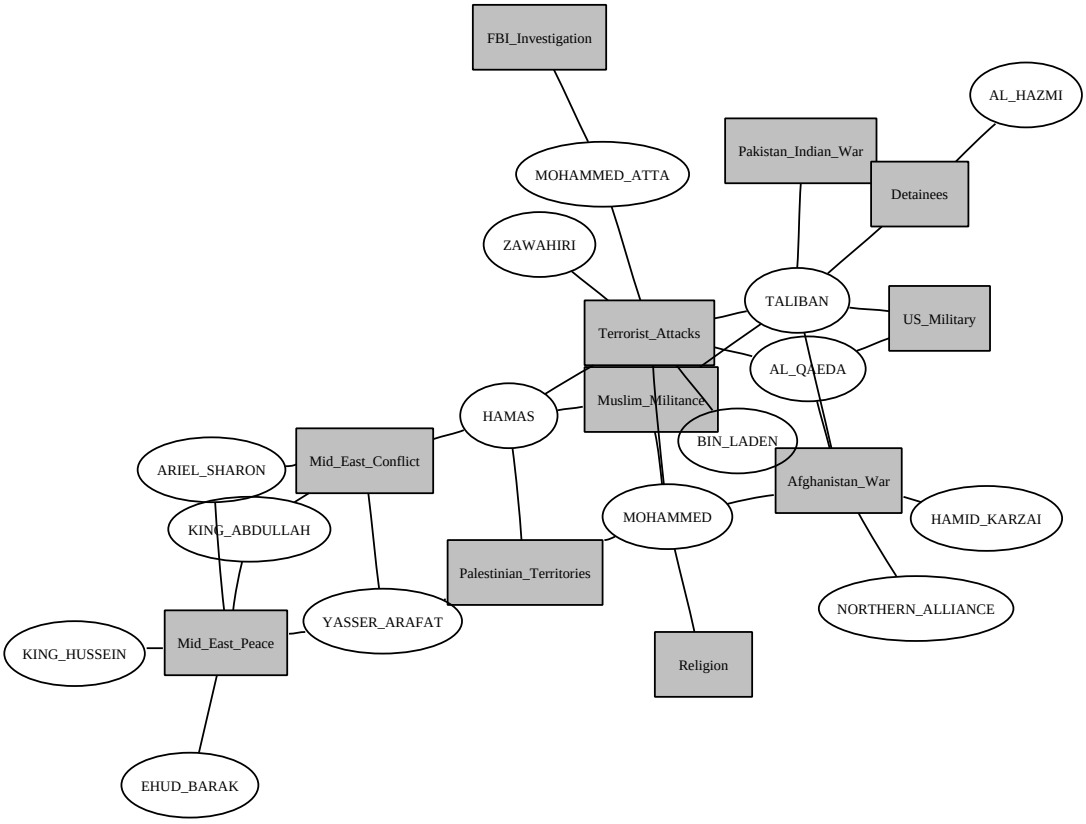

Fig. 5. Bipartite graph showing topic-model-based relationships between entities and topics. A link is present when the likelihood of an entity in a particular topic $p(e|t)$ is above a threshold.

5 Conclusions

Statistical language models, such as probabilistic topic models, can play an important role in the analysis of large sets of text documents. Representations based on probabilistic topics go beyond clustering models because they allow the expression of multiple topics per document. An additional advantage is that the topics extracted by these models are invariably interpretable, facilitating the analysis of model output (in contrast to the uninterpretable directions produced by other language models such as LSI).

In this research, we have applied standard entity recognizers to extract names of people, organizations and locations from a large collection of NY Times news articles. Probabilistic topic models were applied to learn the latent structure behind these named entities as well as other words that are part of documents, through a set of interpretable topics. We showed how the relative contributions of topics changed over time, in lockstep with major news events. We also showed how the model was able to automatically extract social networks from documents by connecting persons to other persons through shared topics. The social networks produced in this way are different from social networks produced by co-reference data where persons are connected only if they co-appear in documents. One advantage over these co-reference methods is that a set of topics can be used as labels to explain *why* two people are connected. Another advantage is that the model leverages the latent structure between the other *words* present in document to better estimate the latent structure between *entities*. In its current version, the probabilistic topic model ignores the difference between entities and other words occurring in documents. There are many other models that are yet to be defined that explicitly model different types of words, such as names of people, organizations, locations and other words. Nevertheless, this research has already shown the benefits of applying simple statistical language models to understand the latent structure between entities.

Acknowledgements

Thanks to Arthur Asuncion and Jason Sellers for their assistance. This material is based upon work supported by the National Science Foundation under award number IIS-0083489 (as part of the Knowledge Discovery and Dissemination Program) and under award number ITR-0331707.

References

1. Klimt, B., and Yang, Y.: A New Dataset for Email Classification Research. 15th European Conference on Machine Learning (2004)
2. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, Vol. 1 (1999) 67–88
3. Chakrabarti, S: Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers (2002)

4. Deerwester, S.C. , Dumais, S.T. , Landauer, T.K. , Furnas, G.W. , Harshman, R.A.: Indexing by Latent Semantic Analysis. American Society of Information Science, 41(6) (1990) 391–407
5. Berry, M.W., Dumais, S.T., O'Brien G.W.: Using Linear Algebra for Intellingent Information Retrieval. SIAM Review 37 (1994) 573–595
6. Hofmann, T.: Probabilistic Latent Semantic Indexing. 22nd International Conference on Research and Development in Information Retrieval (1999)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, **1** (2003) 993–1022
8. Minka, T., and La, J.: Expectation-Propagation for the Generative Aspect Model. 18th Conference on Uncertainty and Artificial Intelligence (2002)
9. Griffiths, T.L., and Steyvers, M.: Finding Scientific Topics. National Academy of Sciences, 101 (suppl. 1) (2004) 5228–5235
10. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of Population Structure using Multilocus Genotype Data. Genetics 155 (2000) 945–959
11. Buntine, W. , Perttu, S. , Tuulos, V.: Using Discrete PCA on Web Pages. Proceedings of the Workshop W1 on Statistical Approaches for Web Mining (SAWM). Italy (2004) 99-110
12. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and Role Discovery in Social Networks. 19th Joint Conference on Artificial Intelligence (2005)
13. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery. 10th ACM SIGKDD (2004)
14. Newman, D. J., and Block, S.: Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper. Journal American Society for Information Science and Technology (2006)
15. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents. 20th International Conference on Uncertainty in AI (2004)
16. Blei, D., and Jordan, M.: Modeling Annotated Data. 26th International ACM SIGIR (2003) 127-134
17. Griffiths, T., Steyvers, M., Blei, D. M., Tenenbaum, J. B.: Integrating Topics and Syntax. Advances in Neural Information Processing Systems, 17 (2004)
18. Steyvers, M., and Griffiths, T.L.: Probabilistic Topic Models. T. Landauer, D. McNamera, S. Dennis, and W. Kintsch(eds), Latent Semantic Analysis: A Road to Meaning: Laurence Erlbaum (2006)
19. Brill E.: Some Advances in Transformation-Based Part of Speech Tagging. National Conference on Artificial Intelligence (1994)