Topic Detection Using Independent Component Analysis

Scott Grant
Queen's University
scott@cs.queensu.ca

David Skillicorn
Queen's University
skill@cs.queensu.ca

James R. Cordy
Queen's University
cordy@cs.queensu.ca

Abstract

If the documents of a large text corpus can be modeled as the rows of a matrix, it can be shown that existing mathematical methods can be used to extract previously unseen information about their relationships. In particular, it can be shown that Independent Component Analysis offers a way of identifying threads of related conversations in a large data set such as VAST. By treating each document as a vector, with word frequencies representing the components, we can extract two interesting pieces of information from the set: a list of the topics used in each document, and a list of the documents that best fit each of these topics.

1 Introduction

Text mining is a process concerned with the extraction of relevant information from a document collection [5]. A key aspect of text mining is the identification of patterns and relationships, both between documents and in the documents themselves. This process often involves a significant amount of manual effort by users with detailed knowledge in the domain of discourse.

In order to improve the technique of analysing extremely large textual data sets, it is beneficial to find ways of improving a computer's ability to preprocess data in a way that will make it easier for humans to make sense of the entire corpus. One such goal is automating the process of identifying topics that thread their way through large numbers of documents, or even paragraphs within a document itself. By modeling independent sections of text as the rows of a matrix, where documents are represented as rows, and word frequencies are represented as columns, it becomes possible to use existing mathematical techniques to pull out this topic information without any necessary semantic knowledge.

2 Background

Given a set of documents of an extremely large size, such as the set of news files found in the VAST 2007 data set [2], identifying groupings of related text is extremely difficult and time-consuming. There are many instances where prior knowledge is needed, or where experts are required in order to make sense of which documents belong together. This overhead increases the time required to comprehend the overall set of knowledge that can be obtained from the text. In addition, some documents may be accidentally mislabeled, or even missed entirely. Reliable automation of this process would help minimize these problems, and greatly decrease the overall time required.

2.1 Topic Detection. Topic Detection is a form of text categorization, whereby a set of documents is analysed in order to identify threads of related information that are commonly used. With the increase in documents available and the need to classify the large amount of data now present, text categorization has begun to attract a great deal of attention as a research topic (for example, [10]).

The most common method of performing categorization of a large text corpus is to identify topics beforehand, and to then train a classifier using techniques such as machine learning to identify documents that best match the topic. A problem with this technique is that topics must match the data quite well, or the possibility of missing topics entirely becomes apparent. For example, outliers may be present that do not fit the expected data. This has been studied in some detail [9].

2.2 The VAST 2007 Data Set. As part of the IEEE VAST 2007 Symposium on Visual Analytics Science and Technology, a contest was designed to test the ability of teams to use data mining techniques to identify suspicious activities in a large data set. The data set itself consists of about 1500 news stories from a fictitious newspaper, plus a few other items collected by the previous investigators, and a collection of blog entries and other formats. Specifically, the goal of the contest was to find the plots and subplots, and to identify the detailed relations about entities in the story.

Teams from around the world were invited to submit a paper describing their analysis techniques, and what they believed the major subplots of the story to be. After the judges had a chance to review the results, it turned out that the best debriefing of the data came from a team that did a great deal of human analysis, using a collection of non-specific tools like online spreadsheets and physical post-it notes, in lieu of more

advanced data-mining tools that could cover the entire process. In addition, some very interesting visualization work and techniques came out of other papers from the conference.

3 Independent Component Analysis

A mathematical technique first introduced in 1994 [3], and later elaborated on in other texts [6], Independent Component Analysis attempts to separate a set of input signals into additive components in such a way as to maximize the statistical independence of the output components. ICA accepts a set of inputs that have had the goal signals mixed together in some way, and attempts to extract the original components identifying the signals that maximize independence from each other. This attempt to extract these independent signals from the mixed input set is a specialization of a technique called *blind signal separation*, where a set of signals are to be separated from a source set of mixed signals with little information about the nature of the target signals themselves.

The original example of ICA as a technique is the idea of a set of microphones hung over a party, where a number of people are engaged in conversations. From the set of data obtained from the microphones, is it possible to extract the original voices? It turns out that it is possible, and if there is enough statistical independence between the signals, the original voice data for each of the attendees can be isolated.

Independent Component Analysis involves the factorization of a matrix A into two matrices. One of the matrices describes a number of independent components, representing the individual extracted voices from the party described above. The other matrix is a mixing matrix, and holds information about how the independent components themselves were combined to produce the original matrix A [11].

3.1 ICA in Text Mining.

There has been some previous work in using ICA with text mining, including the classification of medical documents [8], and the analysis of dynamically evolving textual data such as that found in Internet chat rooms [1]. In particular, the latter paper uses a technique to help identify keywords related to the topics being discussed at a given point in time. This approach is similar to the one applied in this paper, where topics are identified and viewed using temporal information in addition to semantic content.

In this particular application, the two matrices listed above can be used to identify both the topics and the relevance of documents to those topics. The independent components extracted from the word frequency matrix show which tokens best describe a given signal. The highest and lowest points represent the relevance of tokens at the given position in the signal. For example, if the ith position in a single independent component is the highest value overall in that component, it can be considered the strongest contributor to the signal. If it is the lowest value, it can be considered the strongest token that identifies a document as not being relevant to the particular topic represented by that signal. This signal view is different than just looking at the presence or absence of tokens, as it isn't single tokens that indicate whether a document is or is not relevant, but rather the presence of the entire set of tokens.

The mixing matrix offers a glimpse into how relevant a given signal is to each document. If there are n signals extracted using ICA, each document's row will have n columns corresponding to the significance of each signal. The highest value can be interpreted as indicating the strongest presence of that signal, or the most relevant topic to that document. It is using this approach that leads to a tool developed to extract topics and relevant documents from the VAST data set, described below.

4 Implementation

The process is currently broken down into a number of iterative stages. As input, we accept a set of individual documents. In this specific case, the input is the 1,455 XML news files contained in the VAST 2007 data set. Next, the set of documents is tokenized and stemmed, in order to collect a full list of tokens used through the documents, and to isolate the word stems. Common words, punctuation, and other symbols like the XML markup are stripped from the input.

From these new modified documents, a matrix of token frequencies is generated, where rows correspond to documents and columns correspond to tokens. The value in a given position in the matrix is the frequency of the token's occurrence in the document.

With the frequency matrix available, we can now proceed with the Independent Component Analysis step. The matrix is imported into Matlab, and as output, we obtain the estimated independent signal matrix. For the purposes of topic detection, the signal can be considered a way of identifying which tokens help identify good candidates for a given topic, and which ones actually indicate that documents probably do not match.

5 VAST 2007 Results

5.1 Frequency Matrix.

The frequency matrix generated from the VAST 2007 data set is very sparse, and contains 1455 rows and 7402 columns. The values in the matrix itself are 0 if the corresponding token is not

Signal	Tokens
Signal 1	pet fad parmenti said among animal care own chinchilla common diet exotic however know
Signal 2	dog akc pet shop bre friend puppy registration breed eligible obedience register buy compete
Signal 4	abuse elder animal dog neglect perpetrator control cruelty dead percent report victim accord adult
Signal 8	animal hong kong asia china wild chu cat dog foundation government resolution
Signal 10	bill committee said magee ask assembly kuhl lett smith talk 518 albany assemblyman
Signal 12	drug big fix market pharmaceutical book doctor industry new year american become company create
Signal 13	cow disease beef cattle mad alberta canada case ban canadian chicken dog food found four herd
Signal 17	fish have sea that already anyone archive cod fisheries decade fleet newfoundland reveal species trawl
Signal 19	cow disease beef cattle mad alberta canada case ban canadian chicken dog food found four herd
Signal 20	diet animal vegan way pcrm physician two consume enough food more normal people research
Signal 21	food book nation about bestsell fast increasing new system 2001 america antibiotic attention
Signal 22	bechtel iraq public report contract environment from right citizen contractor destruction
Signal 25	research drug favour industry more new study back found fund pharmaceutical positive product

Figure 1: Some Independent Signals with Tokens

found in the document, and less than or equal to 1 if it is found. Non-zero values represent normalized frequency values for the tokens across a document. As mentioned previously, the tokens themselves were cleaned up and stemmed during preprocessing, with punctuation and other common words removed.

5.2 VAST Topic Signals. After applying ICA to the frequency matrix, 25 signals were extracted. This number is fairly arbitrary, and some duplication can be seen in the final list of independent components that are derived, where signals like 13 and 19 share token lists. Some runs were performed using 10, 100, and 150 signals, but the best fit for this data set seemed to fall around 25 signals.

Several interesting trends emerge when looking at the tokens that classify each topic. Some of the more interesting signals are listed in Figure 1, and the full signal list is given at the end of this paper.

The VAST 2007 data set is known to contain several threads of information on topics like veganism, animal rights, and pet care [2]. In addition, other subtopics, such as government laws, fishing rights, and the activities of a number of activist groups are referenced many times.

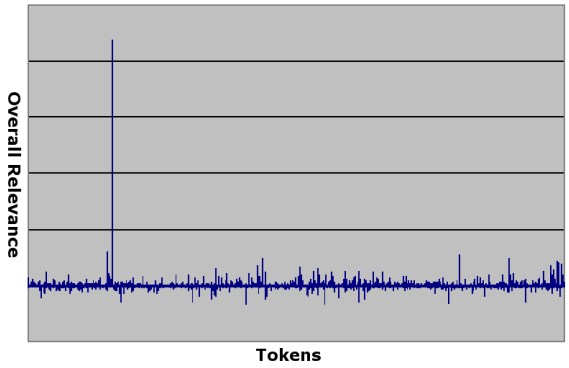

Figure 2: VAST 2007: Signal 19

It can be observed from Figure 2 and Figure 3 that the independent components extracted contain low values for the majority of tokens, with a small number of standouts. These peaks are reflected in Figure 1 as the tokens that identify a signal's topic. Conceptually, higher values for tokens in a given signal indicate that the presence of the token is a good sign that it belongs in the set of related documents.

What is important to note about this method is that all the signals returned share these characteristics. The sample signals provided in the Figures are representative of the type of data returned when applying ICA to the VAST 2007 data. A few select tokens representing a topic that describes the document set well are found having very high relevance values in the signal, while the majority of tokens have values near 0.

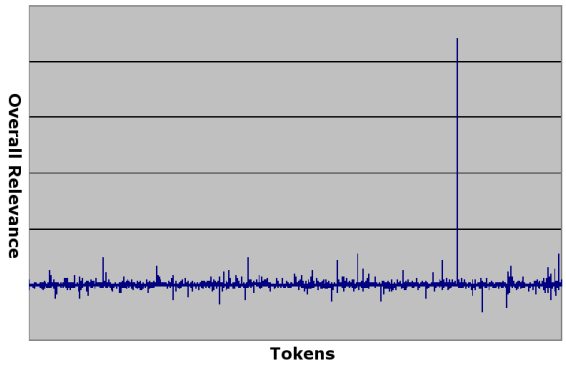

Figure 3: VAST 2007: Signal 4

5.3 VAST Documents. Identifying documents that best match the signals extracted from ICA allows us to actually view documents that follow the topic thread. This classification can be viewed by example, analysing

the documents that are most highly scored when looking at the resultant weight matrix value corresponding to the document in question, and the column for the signal that has the largest value.

Figure 4 shows the sample plot for the document strength of a topic that references Mad Cow outbreaks. Some very distinct high values are easily identifiable, as are a number of low negative values. The majority of the documents fall within a reasonable range from neutral, indicating they have little relevance to the topic. It is the documents that score very high or very low overall that show relevance or irrelevance to a given topic.

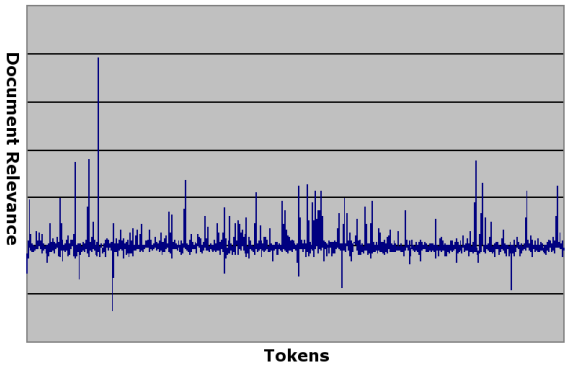

Figure 4: VAST 2007: Signal 19 Document Scores

In order to identify the documents that best match the topics extracted as independent components, the weight matrix generated earlier can be investigated. Strong entries in the matrix corresponding to a given document and signal can be viewed as indications that the topic is heavily used within the document, and low entries indicate the absence of that topic.

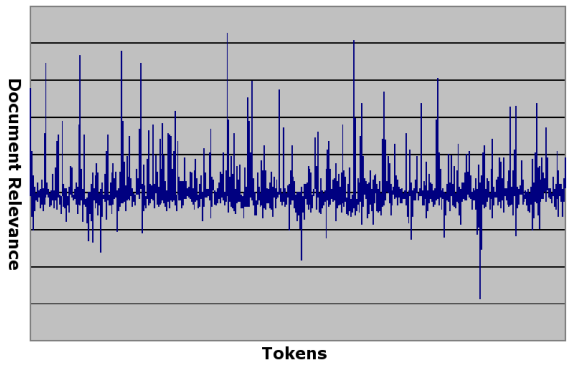

Figure 5: VAST 2007: Signal 4 Document Scores

Using the strongest weighted values from this matrix, it is possible to examine the documents that fall on the high and low points of the signal, and to make a subjective analysis on their relevance. Signal 4 is shown in Figure 5, which seems to relate to issues surrounding animal abuse (tokens like abuse, animal, dog, and neglect, as seen in Figure 1). The best documents that match this signal are shown in Figure 6, and seem to be a good fit.

Partial Filename	Sample Text
20040119.51	Like agribusinesses everywhere, milk producers have tried to increase output while cutting costs. The victims are the cows.
20040308.56	A nutrition student says her rights were trampled on after being forced to work with animal products in one of her classroom experiments or fail her assignment, despite her personal beliefs.
20031124-5.125	Everyone tried to calm me, she said. You just want to do it right. Otherwise the animal would suffer. She did it right. That's what the head slaughterer told her after the first shot.
20031006-4.32	Yesterday, the national animal rights organization People for the Ethical Treatment of Animals (PETA) wrote a letter urging the Fairhaven police to charge Mrs. Hawkins with animal cruelty.

Figure 6: VAST 2007: Signal 4 Documents

The trend continues through other topic threads. Signal 20, with tokens relating to diet research and alternate eating, provides a set of documents after investigating the weight matrix that reference debate and discussion about the Atkins diet, and about a vegetarian lifestyle.

Partial Filename	Sample Text
20040223.89	The Physicians Committee for Responsible Medicine, which promotes a vegetarian diet, issued a warning about Atkins and other high-protein, low-carb diets
20040209-3.11	The PCRM is a fiercely anti-meat, pro-vegetarian brains trust that has opposed the Atkins diet for years, apparently with the best of intentions.
20040209-2.35	The Atkins company deplored the leaking of Dr Atkins's medical records to "a known group of vegan and animal rights extremists".
20040223.90	The militant vegan at work discovered this morning that I am on the Atkins diet and wigged out. She's always trying to get everyone to quit eating meat.

Figure 7: VAST 2007: Signal 20 Documents

5.4 Quality of Results. In order to gauge the quality of the results, we can look at the online solution set provided by the VAST 2007 judges [2]. One solution document is a table that outlines the key Bioterror subplots and events. By using this as a reference, it is possible to look at the output signals obtained from applying Independent Component Analysis to the document set, and see how they line up.

The initial set of events are described "Chinchillas become fad exotic pet." Looking at the ICA results, the primary tokens that stand out in Signal 1 seem to describe this subplot quite well. The strongest tokens in Signal 1 include *pet*, *fad*, *animal*, *care*, *chinchilla*, *common*, and *exotic*.

The next set of events listed in the Bioterror subplot table discuss the activities of a radical animal rights group called the Animal Justice League (AJL). One summary of the events include "AJL hits pet shops in LA," where "AJL is shown to have violent tendencies." From the ICA results, Signal 18 seems to describe activities that may line up with this, using strong tokens such as *animal*, *ajl*, *group*, *peta*, *petsmart*, *store*, *activities*, *angeles*, *authorities*, and *investigation*.

It isn't clear how well ICA performed on extracting information about individuals, although some signals appear to reference political figures who may have implications to the subplots. In addition, many related topics appear to be extracted successfully, such as animal disease outbreaks on farms, vegetarian lifestyles, and industrial and pharmaceutical research. Overall, the technique appears to have very promising results.

6 Future Work

One primary avenue for expansion of this work seems to be improvement of the preprocessing step. There has been a large set of research in semantic analysis and text markup that may improve the ability to identify good tokens which can then be used in topic extraction. Cerno, a tool implemented in the structural transformation system TXL [4], is a framework for generating annotations semi-automatically using a lightweight text-analysis approach [7]. It has already been used to perform Information Extraction from document sets, and shows potential for an application like this. In addition, general Information Extraction tools would help improve the source data used to generate the frequency matrix, and would arguably allow for more accurate topic detection. Since we are losing multi-word tokens, and are not taking implicit advantage of things like acronyms, any help in this area would improve the ability to identify threads of related information.

Further analysis is needed to ensure that the best documents for a given signal are extracted after the independent components are identified. It is often the case that single documents contain references to several topics, and choosing a best topic for each document may not identify its actual relevance accurately. By working with smaller data sets, or data that has been tagged with appropriate topic information, the accuracy of each technique can be better gauged.

Finally, the current toolset to visualize the results is a custom solution that blends Python and Matlab code together. Using standard techniques for user-interface design would help make the application much more usable, and would allow a user to more accurately extract the information threads identified through ICA.

7 Conclusion

Independent Component Analysis offers an interesting and exciting way to automate the extraction of topics from a source data set with no existing semantic knowledge. By tokenizing and stemming the documents, a frequency matrix can be generated, and ICA can be used to determine a set of signals that represent topics used within the documents. From this, documents that most closely relate to the topics in question can be identified.

There are some very exciting avenues of research that follow from this work. First, more efficient preprocessing of the token set would give better data for ICA to work with, and would help identify appropriate threads more effectively. In addition, more work can be done isolating documents that best fit the topics that are extracted. The technique is very effective at isolating topics, but more work is needed to ensure that the documents that match the signals are in fact the best candidates for the topic in question. Using ICA in different sets of data should show similar successes in classification, and it is our intent to experiment further.

Overall, the VAST 2007 data set has provided an interesting testbed for research, and has been an excellent candidate for identifying and evaluating good text mining techniques.

References

[1] E. Bingham, A. Kab, and M. Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters 17: 1-15*, pages 69–83, 2003.

[2] VAST Contest Committee. IEEE VAST 2007 Contest.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[4] J.R. Cordy. The TXL Source Transformation Language. *Science of Computer Programming*, 61(3):190–210, 2006.

[5] Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[6] Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. J. Wiley, New York, 2001.

[7] N Kiyavitskaya, N Zeni, T.D. Breaux, A.I. Antn, J.R. Cordy, L. Mich, and J. Mylopoulos. Extracting rights and obligations from regulations: Towards a tool-supported process. *ASE 2007, 22nd IEEE/ACM In-*

ternational Conference on Automated Software Engineering, pages 429–432, 2007.

[8] T. Kolenda and L. Hansen. Independent components in text. *Submitted to NIPS'99.*, 1999.

[9] S. McConnell and D. Skillicorn. Outlier detection using semi-discrete decomposition. Technical Report 2001-452, Dept. of Computing and Information Science, Queen's University, 2001.

[10] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[11] David Skillicorn. *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. CRC Press, 2007.

Appendix A: Full ICA Signal Results

Signal	Tokens
Signal 1	pet fad parmenti said among animal care own chinchilla common diet exotic however know
Signal 2	dog akc pet shop bre friend puppy registration breed eligible obedience register buy compete
Signal 3	were gorilla kano nigeria pandrillu center environment airport amp bright douala friday
Signal 4	abuse elder animal dog neglect perpetrator control cruelty dead percent report victim accord adult
Signal 5	dolphin bruge dolphinarium taverni without aft close year able animal breed day help hour just long
Signal 6	vogue you animal let office fur know monday peta 2nd against day demonstration build
Signal 7	animal will appear cruelty game government greece greek new athens concern hotel humane
Signal 8	animal hong kong asia china wild chu cat dog foundation government resolution
Signal 9	animal they that one community make money contract county kill made two vega actual mary
Signal 10	bill committee said magee ask assembly kuhl lett smith talk 518 albany assemblyman
Signal 11	bill committee said magee ask assembly kuhl lett smith talk 518 albany assemblyman
Signal 12	drug big fix market pharmaceutical book doctor industry new year american become company create
Signal 13	cow disease beef cattle mad alberta canada case ban canadian chicken dog food found four herd
Signal 14	experiment with scientist capital one second every vivisection animal make other problem science
Signal 15	fluffy iraq dog army joyce force handle home now once out special war 1st american
Signal 16	oregon bill ferrioli research said senate timber
Signal 17	fish have sea that already anyone archive cod fisheries decade fleet newfoundland reveal species trawl
Signal 18	animal that ajl care group last peta petsmart store activities angeles authorities investigation
Signal 19	cow disease beef cattle mad alberta canada case ban canadian chicken dog food found four herd
Signal 20	diet animal vegan way pcrm physician two consume enough food more normal people research
Signal 21	food book nation about bestsell fast increasing new system 2001 america antibiotic attention
Signal 22	bechtel iraq public report contract environment from right citizen contractor destruction
Signal 23	bird teflon death account chicago coat cook cookware fume kill stick toxic veterinarian
Signal 24	animal that people can cruel defense not other real story they
Signal 25	research drug favour industry more new study back found fund pharmaceutical positive product