A Lookup-Free Approach to Knowledge Extraction from News Feeds

James Little[1] and Chris Painter[2]

[1] Department of Mathematics, Çankaya University, Ankara, Turkey
james@cankaya.edu.tr
[2] Meme-Machines.com, Gloucester, England, U.K
Chris.Painter@zigzag.co.uk

Abstract. Identifying topics without also introducing external assumptions is a major challenge for supervised learning techniques, which by definition classify texts according to precepts. Such approaches identify the presence of pre-classified ideas, but cannot identify new ideas. In this paper, we present the results of applying a well understood unsupervised learning technique, in an innovative way, to news feeds analysis. We identify frequent sets of words using the A-Priori algorithm, and grade those sets according to the significance of the Association Rules that they imply. Such sets of words identify common themes in news feeds autonomously, with stopwords as the only added input. We present in detail this methodology and validate it by examining the identified ideas for their ability to identify actual news.

1 Introduction and Related Work

While the Internet has accelerated the means of production and distribution of Information, it has not in equal measure packaged that Information into formats that humans find easy to use. This disparity between our capacity to produce and our capacity to process is not new; Denis Diderot and d'Alembert in the "Encyclopédie" [1] identified it, and Alvin Toffler [2] introduced it to us as "information overload" in his 1970s best seller "Future Shock". The idea had surfaced in management speak in the 1960s but the Internet has brought the size and perpetuity of the problem into ever sharper relief [3].

The problem has further been exacerbated by the continuing philosophical debate as to what constitutes Meaning. Much of the debate in the 20th century had assumed that meaning was carried in grammatically correct sentences. By the late 1970's, IBM started building considerable models of the relationship between words and parts of speech [4] and by the early 2000s Bayesian statistics had been applied in the context of text modeling, whereby the topic probabilities provide an explicit representation of a document [5].

Behind these approaches, lay a belief that meaning could be extracted by the use of prior knowledge, either about the grammar and syntax of a language, or about topics in existence. As a consequence, much of the effort into mining text for meaning has

exploited the growing number of algorithms for supervised learning, such as Bayesian and Neural Networks and Support Vector Machines [6]. Primarily such techniques are used to categorize new texts against the categories of the training set, which in turn are in the gift of the supervisor. Moreover, cross-validation can be used to improve the performance of the model, albeit at the risk of under-generalisation and over-fitting.

There are however structural limitations in the supervised approach. Quite apart from the dangers of over-fitting by parameter tuning, there is a fundamental question about the codification of bias. What is the correct syntax of this sentence? What is the grammar and what is the dictionary? What decides the categories? For empiricists, the answer should be evident only from the texts under examination. That would however require machine learning with minimal supervision, as close as possible to unsupervised learning.

The experiments we describe in this paper expand on a notion first proposed by John R. Firth, Professor of General Linguistics at SOAS until 1956. Firth stressed the context dependent nature of meaning and is best known by the proposition that "You shall know a word by the company it keeps".

In brief, we apply the A-Priori algorithm to identify Frequent Item Sets (FISs) of words that occur together, and grade those sets by the singleton Association Rules (ARs) that each set implies. Singleton rules in this context contain just one term in the conclusion, as in this example from a Sports News source of a FIS {Hungarian, Grand, Prix}.

AR1: {Hungarian, Prix} $\Rightarrow$ Grand

the presence of the words "Hungarian" and "Prix" strongly imply the present of the word "Grand", within a segment of text

AR2: {Grand, Prix} $\Rightarrow$ Hungarian
AR3: {Grand, Hungarian} $\Rightarrow$ Prix

The literature on AR mining on text corpora does not constitute an overload, yet. There however have been some notable papers. An early approach by Maedche and Staab [7] mined ARs from frequent pairs, and built an ontology from the connection matrix. Our work extends it to FIS of arbitrary length, enabling ontologies with greater granularity to be discovered. Our approach does though focus on the single item conclusion in order to avoid the "powerset" problem described in section 2. At about the same time, Hristovski et al. [8] produced an interactive AR builder which generated ARs from a given start point. In particular, ARs were formed by associating the output of a search engine to the input query start point; the authors then went on to use logical inference to generate new propositions from those ARs, and human experts to prune the results. Given the user supplied start point and pruning, this approach is heavily supervised and contrary therefore to our minimal supervision requirement. It does however point out the additive benefits of rules in general, and Association Rules in particular. We also use this principle to accumulate a rule base which could also be open to logical inference.

More recently, Word2Vec [9] proposed an approach which recognises meaning through the "company the words keep". In Word2Vec, each word is associated with a

context derived from other words in its vicinity and encoded in a vector. This is achieved through a neural network learning approach. Similar words found in the same space will have a high similarity in their vectors as measured by the cosine distance. Such vectors are immediately usable by computers, and have proved extremely useful in machine translation, where similarity is a key criterion. That being said, the approach is not immune to traditional "Black Box" objections to neural networks; the output vectors have little meaning for humans, and so may not ultimately reduce information overload.

The area of topic discovery from multiple text streams (hence differently expressed) is highly relevant to our research. Wang et al. [10] claim that correlating multiple streams "provides more meaningful and comprehensive clues for topic mining". Their angle is to identify and bring together the same topics over long periods of time. Wang et al. [11] also agree on convergence of sources, but in contrast, focus on short periods of time. Here, they identify 'bursty" types of news, which is the same item intensively covered across multiple streams. Our approach shares the same objctives of identifying useful patterns in news. However, the novel feature is that we do not require a vocabulary of words as both the above papers do. This means that our approach is much more versatile to changing news agendas.

Finally, work in the detection of word collocations is fundamental to parsing, translation and summarization [12]. Collocations are frequent combinations of words that co-occur more often than expected by chance [13]. Examples of collocations include, compound proper nouns, compound nouns, noun-adjective combinations, proverbs and quotations. In our work, the words are those which identify a news story and which occur in a proximity to each other. These can also include compound proper nouns, such as peoples' names or countries. In collocation algorithms, once the text has been broken into n-grams, lookup and/or statistical frequency are deployed to isolate collocations from the text. There is similarity here in that our Association Rule approach also uses the frequency or "support" to propose a set of words. Out of choice, we do not use lookup in our approach. Therefore we may identify frequent item sets such as, {Romney, Trump, Mitt, Donald}, but identifying the collocated words within it, would impose an order of Mitt Romney and Donald Trump. This would come for us in a post-processing stage.

In the sections that immediately follow we describe the properties and challenges of Association Rules, and then move in section 3 to explain how our system "Sherlock", adapts this approach. Section 4 describes the results of an experiment to extract key topics from RSS newsfeeds and is followed by our conclusions from the work to date, and suggestions for ongoing work.

2 Association Rules and Frequent Item Sets

In formal terms, ARs are statements in the Propositional Calculus, which is a branch of Mathematical Logic concerned with deriving the truth or falsity of propositions derived from other propositions. It is the oldest form of logical manipulation, being attributed to Chrysippus in the third century BC. Correct use of its rules of inference remains a

test for arguments to this day, where it is known as Boolean algebra. A less formal point is that we are very accustomed to digesting complex information presented in this way. ARs extend the calculus by adding an empirical measure of interestingness to such statements, in order that validly inferred rules may be compared. They are used to discover correlations between variables, and drive the Market Basket Analysis we see at Amazon, which shows purchase patterns similar to a proposed purchase [14].

Behind every AR there must be a set of terms forming the premise, "if" part of the rule, and the terms that form the conclusion, "then", part of the rule. Such sets are called Frequent Item Sets (FIS). Identifying FIS is a combinational problem that involves exploring and counting the Lattice of variable combinations. Such a Lattice for just three variables a, b, and c is shown in figure 1. In order to show a property of all such lattices, namely that of downward closure, a set of length N has a maximum frequency of X, where X is minimum frequency found in its constituent sets of length N-1. Exploiting this property, efficient algorithms such as A-Priori can find all frequent item-sets.

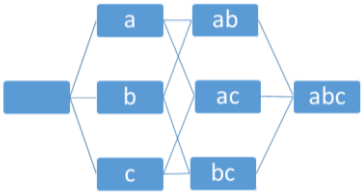

Figure 1: Lattice for Generating FIS

The exploration of the lattice of possible combinations can be conducted depth or breadth first, building either trees with counts for each node, or arrays expressing the same empirical evidence about the frequency of each set's constituents. Interested readers should visit Christian Borghelt's [15] repository of algorithms to see the range of approaches that have been adopted. Whatever the approach taken to the identification of FISs and their counts, two issues need resolution before statistically sound ARs can be extracted from the FISs identified. The first is the "powerset" problem and the second concerns the selection of the measure of "Interestingness".

The "powerset" problem is that any member of the powerset of a FIS, except both the full and empty sets, could be the premise set of an AR, so a FIS of length 3, as shown above implies (2^N) -2 possible premise sets, and therefore (2^N) -2 rules. In this case one FIS implies 6 rules, each of which need counting. Possible premise sets are, {a}, {b}, {c}, {a, b}, {a, c}, {b, c}. The longest FIS found in our Newsfeed experiment has 11 members. This means $2^{11} - 2$, or 2046 different ARs, all of which need to be counted. The issue is therefore one of combinatorial explosion of FIS which need to be examined.

The second issue that needs to be addressed is the selection of an appropriate measure of "Interestingness". In their paper, Pang-Ning et al. [16] demonstrate that all measures of Interestingness have biases such that rules may be graded differently by different measures. We commend that paper to the reader, along with a comprehensive collection of links to measures of Interestingness maintained by Hahsler (2016).

3 Finding the FIS and ARs

Sherlock uses the A-Priori algorithm to generate FISs, restricts rule generation to a single term conclusions in order to avoid the "powerset" problem, and uses Iterative Proportional Fitting (IPF) to produce the most objective measure of Interestingness (Pang-Ning et al, 2004). Sherlock therefore uses IPF to adjust the row and column totals of the observed data towards the balanced totals that would occur if the chances of the premise being true were 50%, and the chances of the conclusion being true were also 50%. By standardising the contingency table of each rule in this way we are able, albeit at further computational cost, to avoid introducing bias associated with any scoring system. The A-Priori algorithm generates FISs in a breadth first search, the set of all pairs is generated from the set of all sufficiently frequent single items, the set of all of all triples is generated from the set of all pairs, and so on until no further generation is possible.

The question that Sherlock seeks to answer is, what are the most important N variables (words) in this data (news) and how are they connected?

In order to avoid overwhelming available memory, the program operates by reducing the frequency threshold until the set of ARs, sorted by IPF adjusted score, contains at least N unique items. Using the rules derived from the data we can construct a network of the most important terms to demonstrate the connections, as in Fig. 1.

The data structure produced is a dynamic array in which each row represents a singleton AR; that is a rule with only one item in the conclusion, along with its score, where 0.5 is the maximum, due to iterative proportional fitting. It grows as data comes in; it starts with frequent items, it then appends pairs that also exceed the frequency threshold, from these are generated candidate triples, triples that get over the frequency threshold get appended to the array, and so on, towards larger sets.

Table 1 shows a typical output, starting with the last frequent items and the first 2 pairs. Therefore, for the 1st row, this singleton AR is derived from the 1st FIS of size one and has been seen 13 times already. Rows 3 and 4 constitute two rules from the same FIS of size two. Furthermore, we can see that the first rule derived from the second pair {rows 5 and 6} had the highest score for a singleton rule (0.37998). Note that frequent item sets of size 1 do not constitute singleton rules, as the premise would have to be an empty set.

Table 1. Snapshot of Singleton Rules

Row	FIS ID	Position	Set Size	Item ID	Set Count	Premise Count	Conclusion Count	Interest Score
1	1	1	1	23424	13	13	13	0.5
2	1	1	1	23457	14	14	14	0.5
3	0	0	2	16	6	58	16	0.36536
4	0	1	2	2272	6	16	58	0.36531
5	1	0	2	16	7	57	16	0.37998
6	1	1	2	10730	7	16	57	0.37982

In Table 2 we see that the table ends with eight quintuples, the last being made from the FIS {5798,8563,19708,7708,17210} which occurred 6 times and that the best rule produced by that set was rule number 245958 namely, {5798,8563,19708,7708} => {17210}, which had the highest score. This states that whenever the premise occurred in the data 17210 occurred nearly 10 times as often as the set.

Table 2. Long Association Rule

Row	FIS ID	Position	Set Size	Item ID	Set Count	Premise Count	Conclusion Count	Interest Score
245954	7	0	5	5798	6	8	61	0.42815
245955	7	1	5	8563	6	6	65	0.49549
245956	7	2	5	19078	6	7	39	0.45877
245957	7	3	5	7708	6	7	46	0.45468
245958	7	4	5	17210	6	6	56	0.49557

4 Data Processing

Using Newsblur.com, we established a pipeline of news which would append new stories from appropriate feeds defined in Newsblur to an input file in the Cloud. The news channels watched by Newsblur include,

- www.econmist.com/rss/the_world_this_week_rss.xml
- www.telegraph.co.uk/news/worldnews/rss
- rss.cnn.com/rss/edition_world.rss
- feeds.bbci.co.uk/news/rss.xml? edition=int
- feeds.skynews.com/feeds/rss/world.xml

Sherlock first converts each news item into a vector of interesting words, after stripping out tags and stopwords. These are words which do not carry significant meaning from the text, such as ‘and’ and ‘or’. If they were left in, then vast amounts of superfluous rules would be generated. Currently Sherlock supports English, Russian and Arabic texts, but there are no reason this cannot be extended to all languages that have separators to distinguish words. Sherlock then converts to binary the occurrence matrix into an integer vector against which counting can be performed in parallel, and repeatedly reduces the frequency threshold in the manner described above until the sorted singleton rule list that is returned contains at least 12 unique terms. In practice, due to the breadth-first nature of the search, the returned rule set can jump to exceed the minimum term threshold. In the screenshot shown in Fig 2 we see that when the

frequency threshold was reduced to 1.4% 20 unique items were in the results, in 6 distinct clusters. The diagram can be interpreted from the point of view of a node having one or more links into it. This node is the Conclusion and the nodes pointing to it are the premise. Therefore, we have from Figure 2 we have both, {Nepal, Prince} ⇒ {Harry}, {Nepal, Harry} ⇒ {Prince} – two rules referring to a story about Prince Harry visiting Nepal, but compressed into one cluster.

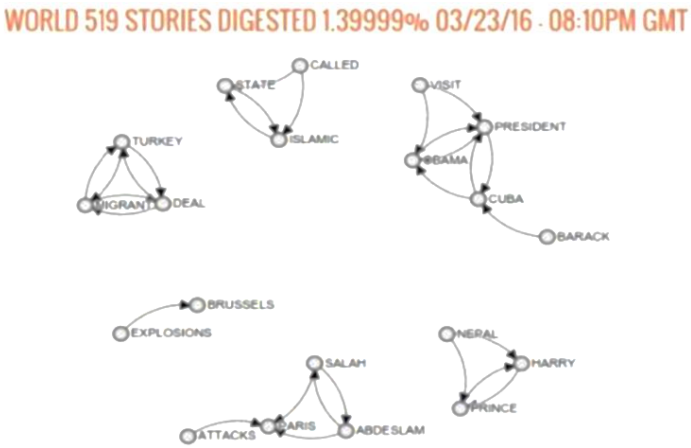

Figure 2: Graphical Illustration of ARs forming Clusters (Rule Sets)

5 Results

Over the period Feb 2015 to March 2016 more than 32,000 updates from news sources have been received. Each causes Sherlock to check for existing ARs and to record findings in a database, along with a summary report. The 227,515 FIS to March 26 are available for download at http://www.meme-machines.com/world4.csv.

963 unique FISs were identified, ranging in length from 2 to 11 according to the following distribution in Table 3.

Table 3. Distribution of FIPs in terms of Number of Items

FIS Length	Frequency
11	1
7	4
6	7
5	24
4	119
3	439
2	369
Total	963

Due to the nature of the A-Priori algorithm, there are many cases of a FIS being a subset of a longer FIS. These subsets are equally plausible, but contain less information about the news item. For example, taking as a starting the longest rule (11), we have found the following FIP decomposition which is also "interesting".

['Airport', 'Atlantic', 'Beam', 'Flight', 'Heathrow', 'Incident', 'Laser', 'London', 'Plane', 'Virgin', 'York']
['Airport', 'Atlantic', 'Beam', 'Flight', 'Heathrow', 'Laser', 'Virgin']

We can also find examples where a FIS is completely constructed from other subset FISs,

['Apple', 'Bernardino', 'FBI', 'San']
['Apple', 'FBI']
['Bernardino', 'San']

Subsets that remain after the subtraction of identified FISs also offer areas of potential interest.

['Airport', 'Atlantic', 'Beam', 'Flight', 'Heathrow', 'Incident', 'Laser', 'London', 'Plane', 'Virgin', 'York']
['Airport', 'Atlantic', 'Beam', 'Flight', 'Heathrow', 'Laser', 'Virgin']

suggests ['Incident', 'London', 'Plane', 'York'], while

['Donald', 'Mitt', 'Presidential', 'Republican', 'Romney', 'Trump']
['Donald', 'Mitt', 'Romney', 'Trump']

suggests ['Presidential', 'Republican'].

Some sets are not decomposable to the smallest basic length of 2. We contrast [Amanda, Knox, Murder], which does decomposes to [Amanda Knox], with [Demi, Drowned, Man, Moore, Pool] which does not decompose further. All this follows from how the news stories are written by the various news sources and their development over time. With filtering in place, only news stories covered frequently will emerge. Further, if they are reported in a consistent way they will more likely appear. With proper compound nouns, such as name or places these are unlikely to be changed from one news item to another. The event surrounding the name may though be described differently.

6 Conclusions and Future Work

With no knowledge of the domain, the grammar, or the language (subject to constraints about stopwords), the derived Association Rules describe meaningful major and minor pieces of news using logic and frequent words. Even with multiple sources of data, news items do use the same set of words to identify topics. Information overload, as we proposed at the beginning, can be reduced automatically to coherent pieces, pointing towards more detail descriptions through Internet search. The results come directly

from the combination of news reporter articles and the A-Priori algorithm. As such, the results reflect how a news item may be written to have a dominant theme (longest item set), while also going on to develop the article through subthemes, evidenced by subsets of FIS. Subtracting item sets from FISs suggest likely candidates for shorter FISs or topics for the reporter. The automatic identification of news themes though gives a way of indexing them for easier retrieval later. The time dimension, although not explored here, can give an insight into the life cycle of a news item; when it is reported and its strength of reporting over time.

Sherlock is available and can be validated at www.meme-machines.com

References

1. Diderot, D., le Rond d'Alembert, J.: Encyclopédie, Vol. V. Paris (1756)
2. Toffler, A.: Future Shock, Bantam, Books, New York (1970)
3. Evaristo, R., Adams, C., Curley, S.: Information Load Revisited: A Theoretical Model. In: Proceedings of the 16th Annual International Conference on Information Systems, Amsterdam, pp. 197-206 (1965)
4. Bahl, L. R.., Baker, J. K., Cohen, P. S., Jelinek, F., Lewis, B. L., Mercer, R. L.: Recognition of Continuously Read Natural Corpus. In Proc. IEEE Int. Conference in Acoustic Speech and Signal processing, Tulsa, OK, US, pp 422-424 (1978)
5. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. The Journal of Machine Learning Research, Vol. 3, pp 993-1022 (2003)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment Classification using Machine Learning Techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics (2002)
7. Maedche, A., Staab. S.: Discovering Conceptual Relations from Text. In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'2000). Vol. 321, no. 325. Ed. Horn, W (2000)
8. Hristovskia, D., Starea, J., Peterlinb, B., Dzeroskic, S.: Supporting Discovery in Medicine by Association Rule Mining, in Medline and UMLS. Medinfo. Vol. 10, pp 1344–1348 (2001)
9. Mikolov. T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR (2013)
10. Wang, X., Zhang, K., Jin, X., & Shen, D.: Mining Common Topics from Multiple Asynchronous Text Streams. In Proceedings of the Second ACM International Conference on Web Search and Data Mining pp 192-201 (2009)
11. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining Correlated Bursty Topic Patterns from Coordinated Text Streams. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining pp 784-793 (2007)
12. Verma , R., Vuppuluri, V., Nguyen, A., Mukherjee, A., Mammar, G., Baki, S., Armstron, R.: Mining the Web for Collocations: IR Models of Term Associations. In Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics (2016)
13. Smadja, F.: Retrieving Collocations from Text: Xtract. Computational Linguistics Vol 19, pp 143–177 (1993)

14. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I.: Fast Discovery of Association Rules. Advances in knowledge discovery and data mining, 12(1), pp 307-328 (1996)
15. Hahsler. M.: A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules, http://michael.hahsler.net/research/association_rules/measures.html (2016)
16. Pang-Ning, T., Vipin, K., Jaideep, S.: Selecting the Right Objective Measure for Association Analysis, Information Systems. Vol. 29, no. 4, pp 293-313 (2004)