Annotated Bibliography

Stuart Miller
Master's Student in Computer Engineering
Missouri University of Science & Technology

CpE 6330 - Clustering Algorithms

May 2, 2018

Contents

Forward

This paper referenced a number of papers purely as examples of data validation. As the papers were the subject of the data themselves, the contents of which were not analyzed for background knowledge. Their role shall be confined to that of illustration and they not reproduced in annotated form below.

1 A statistical interpretation of term specificity and its application in retrieval

The article by Jones (1972) [1] is the renowned work which arguably originated the field of natural language processing. Working at the time for the University of Cambridge library, Jones presents a method of retrieving documents based on search terms with the goal of maximizing the chance of a request returning appropriate documents while avoiding as many false positives as possible. Her analysis begins with proper choice of search terms and the need to choose keyword with adequate specificity. This is something that scholars in the Internet era are all too familiar with: making heavy use of keyword searches in online databases. The novelty in her proposal, however, comes from the notion of the broader implications of statistical term specificity. Drawing on the statistical properties of Zipf's law (that word frequency nearly doubles for each more common word in a corpus), Jones introduces an inverse weight based on term commonality. That is, terms that appear high up in the Zipf distribution have their index ranking reduces proportionally. By doing so, she is able to experimentally prove that such a weighting scheme reduces false positives and provides a more accurate versions of performing multi-term keyword search rankings.

2 Application of Webpage Optimization for Clustering System on Search Engine V Google Study

Presented here as an example of current research into the term frequency - inverse document frequency (TF-IDF) algorithm as well as evidence of its applicablility to many internet engines, Lin and Chi's 2014 paper [2] presented an investigation into usefulness of TF-IDF for search engine optimization. Taking current results from Google searches, results are parsed and analyzed to select the most relevant words being searched for. This gives the benefit of providing insights to potential webmasters who wish to choose better vocabulary on their site in order to rank higher on search results.

The paper starts by choosing a number of popular search terms in both English and Chinese. Retrieving the language from the results of the top 100 sites for each, results are preprocessed and filtered so that only important terms remain. Words are stemmed and fed through a TF-IDF classifier. A combination of the TF-IDF ranking, site traffic from Alexa (`http://www.alexa.com/`), and BackLink ranking (`http://www.backlinkwatch.com/`) determines each term's overall ranking. The results are then clustered via k-means and grouped into areas of similarly ranked and themed words.

By utilizing the most valuable words in an appropriate cluster, a website can provide a significant boost to their search engine ranking.

3 Research and Improvement of feature words weight based on TFIDF Algorithm

The article by Guo and Yang (2016) [3] presents a novel weighting of the TF-IDF algorithm for big data applications. Their goal is to optimize the TF-IDF algorithm for extremely large repositories of large unclassified data that lacks keywords. Many systems enforce keyword tagging upon document submission, but many legacy systems do not support it, so the need for after-the-fact analysis is present. Principal among the authors' perceived shortcoming with the traditional TF-IDF is the idea that the algorithm does not consider the value of a term's rank relative to similar terms. If a words ranks highly, but cannot be well classified within the scope of other word rankings, it's ranking is not useful and therefore should be reduced when scoring documents relative to it.

In order to compensate for this, Guo and Yang have added a training set and additional parameter K. For any particular word, when nearby words score well K is increased and when nearby words are lacking K is reduced. This addresses one of the major shortcoming of TF-IDF and allows it to look more broadly at the body of documents in as a whole. Moving forward with this notion, Guo and Yang are able to experimentally prove that their method provides slightly higher accuracy over traditional TF-IDF.

Averaged over a massive "big data" set, the slight gains are multiplied and become extremely valuable for their target application.

4 LDA based classification of video surveillance sequences using motion information

The article by Diop, Meza, Gordan, and Vlaicu (2018) is an application of the traditional LDA algorithm to the computer vision field. Included here as a proof of the extensibility and relevance of the LDA, the article presents a novel approach that is quite successful.

In order to classify events of a feed from a collection of video cameras, the authors translated the traditional units of NLP terminology into units relevant for their application. Instead of topics, they use defined actions (i.e. walking, crowd forming, etc.) and instead of documents, they use camera views. Applying the same LDA methods outlined in the other papers mentioned here, they were able to classify blocks of video into probability models of actions and flag certain blocks when concerning actions reached a certain probability. Mathematically, fitness between blocks was calculated by pixel displacement between frames. The result is a novel technique that uses a novel implementation of LDA to rival the leading methods of computer vision.

5 Latent Dirichlet Allocation

The article by Blei, Ng, and Jordan (2003) [4] provides the basis for the latent Dirichlet allocation (LDA) algorithm. Having currently received over 20,000 citations, this highly influential article forms the basis for much of modern natural language processing. Blei, et. al. describe LDA as a generative probabilistic model for analyzing bodies of text. Central to their proposal is the idea that each document can be modeled as a distribution of topics, or a probability simplex. Over this, the body as a whole contains words which belong to their own distribution across the topics. In such a way, words, as well as documents as a whole, can be attributed to clusters of key words. With human interpretation, each cluster can be tied to a real world theme.

Motivated by prior work in the field, notably research into improving TF-IDF and upcoming probabilistic and predictive models, Blei, et. al. based their LDA model on the (at the time) foremost research into Bayesian parameter estimation (called the Dirichlet parameter herein). Their novel method involves taking a Dirichlet probability density parameter and tuning it according to the obtained probability. If the Dirichlet parameter is equal to one, it represents an even distribution; higher represents a distribution concentrated in the middle of both variables, lower represents concentrated on the edges near to the bounds of the variables. As the goal is to obtain the primary topic(s) of each document and word, uneven distributions are highly desirable. Once enough iterations have been completed to tune the parameters to an acceptable value, the distributions can be calculated out by word and summed across each document to get a final topic distribution.

The article includes extensive proofs, both mathematical and experimental, showing the breakdown of the probability simplexes for various data sets and the applicability of the final model.

6 Inference of population structure using multilocus genotype data

The article by Pritchard, Stephens, and Donnelly (2000) [5] describes a novel model-based clustering method to assign individuals to populations based on genotype data. Developing much the same model design as Blei [4], Pritchard, et. al. present their model in order to classify various thrushs into three distinct geographic

populations. Due to crosssbreeding and mixed genotypes, each thrush represents a proportion of a parent population and thus a three-way probability simplex is required for the model.

By taking a given distribution and calculating it back over the assumed variables, a probability can be obtained. If a bird is indeed belonging to genetic group A, then genetic markers x, y, and z should be present in it and similar birds should also exhibit genetic markers x, y, and z. Iterating in such a way to obtain higher probabilities allows the authors to obtain probability distribution with less and less statistical error.

Their analysis focuses on classifying birds in the manner of "is bird x likely to be an immigrant from b? or native to a". Using the developed probability models help answer these questions and more and allowed biologists to develop a ecological history of the thrushs being studied.

7 Efficient methods for topic model inference on streaming document collections

The article by Yao, Mimno, and McCallum (2009) [6] presents collapsed Gibbs sampling, a sampling-based inference method of Bayesian learning that today is most common in implementations of latent Dirichlet allocation (LDA) for natural language processing. With the goal of drastically improving performance of existing LDA implementations, Yao, et. al. applied Gibbs sampling to current methods.

Gibbs sampling is an existing method. In this method, variables are first randomly initialized. Then, by fixing various parts of the variables that are known to be more steady while iterating on the more volatile ones, a quicker iteration cycle can be achieved. Various versions of this existed previously, but this article presents a new implementation targeted at NLP called SparseLDA. Taking advantage of the natural sparsity that not all words will appear in all documents, collapsed Gibbs sampling focuses on iterating around topic variation and does not concern itself as much with individual word assignments. The probability estimation can be divided into three parts: a summation of the the current Dirichlet parameters assigned, the parameters of the other topics in a particular document, and the parameters assigned other terms that are assigned to this topic across the entire body. Such a method cuts out summing every combination of parameters and focuses on the parameters combinations most influential to determining a valid probability.

Experimental validation of this included in the paper features runtimes that are astronomically improved, boasting improvements that are multiple orders of magnitude.

8 Computer information retrieval using latent semantic structure

The patent by Deerwater, et. al. and asssigned to Telcordia Technologies Inc in 1988 [7] is the initial presentation of the latent semantic analysis (LSA) method of natural language processing. Published in a United States patent, and not an academic journal, the concept was initially protected from public use. As the patent is now expired, this is no longer true. Unfortunately, the patent includes little beyond the technical details of the algorithm outlined, and lacks the details of a traditional academic paper.

LSA relies on singular value decomposition of a term-document matrix over the entire body of work. As the term-document matrix is naturally sparse (due to not all words appearing in all documents), a valuable reduction is to reduce the number of rows (terms) while retaining the number of columns (documents). Singular value decomposition accomplishes this and is a relatively straightforward procedure involving established matrix math. Upon completion, the similarity between columns(documents) is obtained and reduced down to only a few rows. The matrix will be weighted to the top, so lower values at the bottom rows can safely be discarded. The number of rows remaining is the number of topics desired. Plotting the remaining rows of each document-column gives a multi-dimensional topic vector for each document. The vector points most closely towards it's primary topic and it's direction can be analyzed for similarity to other documents.

9 An algorithm for suffix stripping

The article by Porter (1980) [8] outlines his namesake stemming algorithm. An information scientist at the Cambridge Computer Laboratory, Porter required a method of standardizing search terms across the many forms in which they were suffixed, conjugated, and appended in literature. In order to unify the diversity

of current stemming systems being used and to create something both comprehensive and simplistic, Porter presented his own stemming algorithm.

The algorithm consists of five steps. Step one deals with plurals and past participles. It must includes exceptions for some edge cases (i.e. removed "-ed" except in words with "-eed" like "seed"). Step two deals with various cases of strange suffixes. Porter proposes indexing by the penultimate letter as they are most unique and provide for for fast indexing. Step three requires full removal of a list of suffixes such as "-icate", "-itive", "-ness", etc. Step four provides that if the preceding letter count is grater than one, more suffixes such as "-al" and "-ance" can be removed. Finally, step five involves clean up for any hanging "e" and double letter endings. These are removed.

Porter is careful to account for all edge and corner cases and thereby presents a complete and comprehensive algorithm that is still widely used today.

References

[1] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[2] T. F. Lin and Y. P. Chi, "Application of webpage optimization for clustering system on search engine v google study," in *2014 International Symposium on Computer, Consumer and Control*, June 2014, pp. 698–701.

[3] A. Guo and T. Yang, "Research and improvement of feature words weight based on tfidf algorithm," in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, May 2016, pp. 415–419.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[6] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 937–946.

[7] S. C. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter, "Computer information retrieval using latent semantic structure," May 1988, uS Patent 4,839,853.

[8] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.