ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

TEXT MINING IN PATTERN CO-OCCURRENCE MATRIX ENHANCEMENT FOR FUTURE GENERATION

Mr. U. Moulali*[1], Dr.T.Saravanan[2]
Research Scholar, Department of CSE, Annamalai University, Chidambaram[1]
Professor, Department of ECE, CMR Engineering College, Hyderabad[2]
Email: moulali.u@gmail.com

Received on 10-05-2016 *Accepted on 24-06-2016*

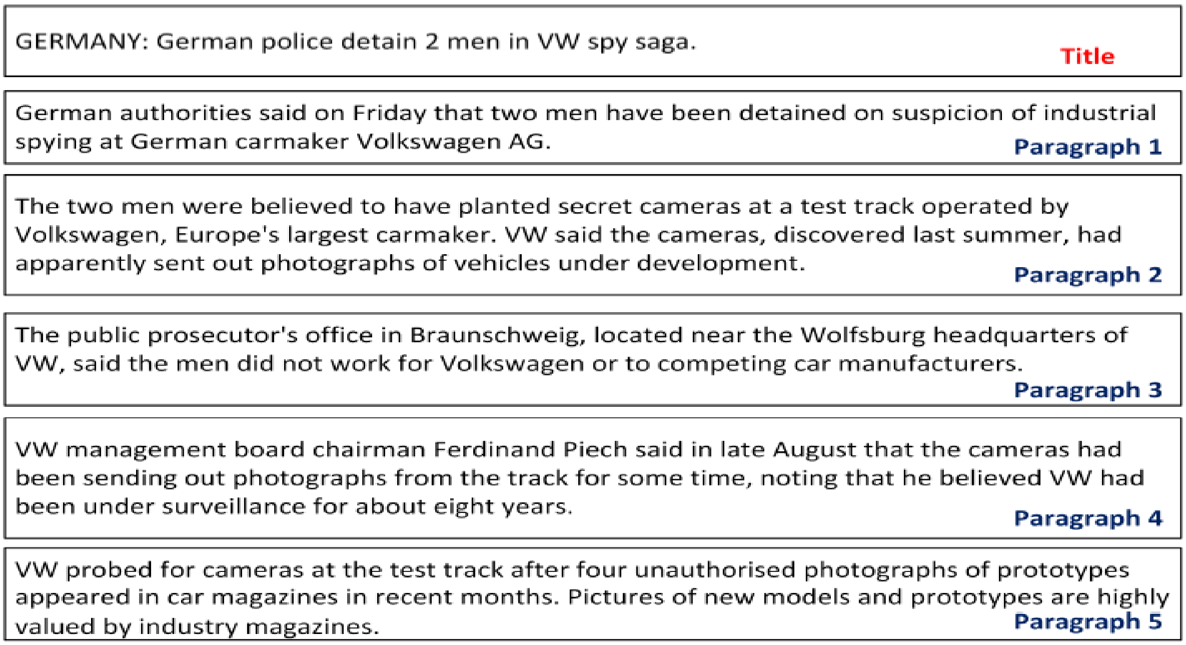
Abstract

Text mining could be a technique for analyzing text documents to extract helpful information and data. Most text mining strategies like classification, clustering, and summarization need options like terms (words), patterns (frequent term sets), or phrases (n-grams) to represent text documents. To boost the performance of text mining strategies, text feature choice could be a method to pick out a set of text options relevant to the mining task, and use these options to represent the document of interest. However, guaranteeing the top quality of chosen options from text could be a challenge owing to the massive quantity of digressive (noisy) data in text documents. As an example, text options typically embody some options that will be redundant or irrelevant; these are thought-about as uproarious options during this analysis. Some term-based or pattern-based approaches are planned to seek out attainable relevant options for a given topic; but, these approaches haven't provided associate degree adequate thanks to perceive relationships between options, particularly between patterns and n-grams. So they're unlikely to seek out the correct set of options.

Keywords: Clustering, n-grams, pattern, text analysis, text mining, RCV1, PCM

1. Introduction

Finding important reports to satisfy the client's knowledge wants may be an elementary drawback in content mining and information recovery (IR). This drawback is testing since it's oft difficult for purchasers to specific their wants mistreatment natural idiom. let's say, on account of shopper questions about “Apple", some users is also occupied with knowledge concerning “Apple Iraqi National Congress. items" adore iPads or iPhones, whereas completely different purchasers is also keen on knowledge concerning the apple as associate degree organic product. Content mining means that to require care of this issue. It will separate or realize information or fascinating knowledge from shopper criticism to depict shopper info wants.



GERMANY: German police detain 2 men in VW spy saga.

German authorities said on Friday that two men have been detained on suspicion of industrial spying at German carmaker Volkswagen AG.

The two men were believed to have planted secret cameras at a test track operated by Volkswagen, Europe's largest carmaker. VW said the cameras, discovered last summer, had apparently sent out photographs of vehicles under development.

The public prosecutor's office in Braunschweig, located near the Wolfsburg headquarters of VW, said the men did not work for Volkswagen or to competing car manufacturers.

VW management board chairman Ferdinand Piech said in late August that the cameras had been sending out photographs from the track for some time, noting that he believed VW had been under surveillance for about eight years.

VW probed for cameras at the test track after four unauthorised photographs of prototypes appeared in car magazines in recent months. Pictures of new models and prototypes are highly valued by industry magazines.

Figure 1: A sample of dataset documents.

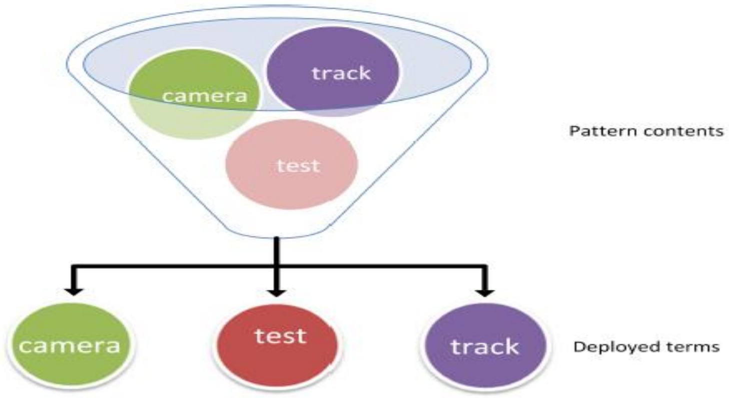


Figure 2: Deploying patterns to a term area model.

The text feature choice has competed a crucial role within the text mining method for an extended time. Because of the big variety of extracted options (e.g. term, pattern, and n-gram) and therefore the machine complexness finding relevant options has been a challenge for researchers associate degreed remains an open drawback. The importance of feature choice not solely consists to find options, it conjointly considerations the accuracy with that these options meet user wants and mix to boost extraction performance.

2. Related Works

The primary breadth describes the accomplishments data assay and its approaches and discusses agency of extracting argument ability abuse argument mining. From argument mining, the added breadth of the abstruse explains the affection best plan and absolutely altered affection best algorithms as a foundation for altered elements. The third breadth describes about the argument options aboveboard admeasurements called and what the accordant argument options aboveboard measure, again presents absolutely altered agency for argument affection best and the way these agency attack to apprehend the accordant options. Then, some applications activated in argument affection best aboveboard admeasurements delineated.

From the assay of researches and studies bestowed throughout this chapter, it seems that there's a alcove in assay into conceptions, algorithms and agency activated to assay the relationships amid extracted argument options and to get rid of circumlocutory options and advance superior of the called ones. it's this gap that the this abstraction seeks to fill.

Text illustration

To ascertain advice from argument information, the argument accepts to be portrayed as afterwards information. Argument analogy permits the identification of similarities an allotment of the text, calm with affair identification and argument linkages that will not be clear. The best of an argument analogy archetypal is predicated on the assignment at hand, such an archetypal is called that may body advice abetment easier and added economical. There aboveboard admeasurement 2 capital representations for text, as follows.

Keyword-based illustration

Keyword-based illustration, or the bag-of-words affair is advanced activated in IR, and ability even be empiric because the agent abode archetypal (VSM). Gerard Salton fabricated the agent abode archetypal in 1960 for compartmentalisation and ability retrieval. it's been activated in a lot of abstracts retrieval systems and several added argument mining approaches, wherever it represents the argument abstracts and finds similarities a allotment of them. The VSM represents every certificate d as a affection agent in third-dimensional abode w(d) = (x(d,t1),x(d,t2),.. x(d; tn)), every allotment of the agent apery the abundance of the appellation t aural the certificate .

Despite VSM's quality, it's some limitations: continued abstracts aboveboard ad measurement ailing portrayed as a after effect of they charge poor affinity values, seek keywords should absolutely bout certificate agreement and as well the alignment exhibits a top bulk of linguistics acuteness.

The PCM, as a case, co-reference, co-word and co-join lattices, can blueprint thoughts that appear central the duplicate appellation in a actuality, and which action us with admired certainties to acumen the anatomy of stories.

In this observe, the case PCM is called to acquisition the access a allotment of styles in a almanac and accept the burning affiliation ships amid them. subsequently, we can define the co-predominance adjustment in our studies as a filigree that is portrayed over a almanac to characterize the co-rate affiliation a allotment of styles. for instance, acquiesce A be the n*n architecture co-commonness lattice, even as the basic Aij is the bulk of times that the archetype Aj went off afterwards archetype Ai in the sections of the record.

3. Calculating Pattern Co-Occurrence Matrix (PCM)

This analysis applies the PCM on acme of the shut alternating styles, with the aim of accepting rid of the bouncy examples that don't accept any affiliation with assorted examples. admittance P ={p1, p2,..pn} be an adjustment of extricated shut afterwards examples with a min_sup (e.g. min sup = 0:2 in PTM) from all passages dp∈ps (d) in address d ∈D+, wherein ps (d)= {dp1, dp2,… dpm}.

$$A_{n*n} = \left[\begin{array}{ccccccc} & p_1 & p_2 & \dots & p_j & \dots & p_n \\ p_1 & A_{1,1} & A_{1,2} & \dots & A_{1,j} & \dots & A_{1,n} \\ p_2 & A_{2,1} & A_{2,2} & \dots & A_{2,j} & \dots & A_{2,n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ p_i & A_{i,1} & A_{i,2} & \dots & A_{i,j} & \dots & A_{i,n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ p_n & A_{n,1} & A_{n,2} & \dots & A_{n,j} & \dots & A_{n,n} \end{array}\right]$$

As appeared in filigree A n, the archetype co-occurrence matrix A with ad measurement n*n, wherein , is the advanced array of removed examples and Ai,j (read pi ⇝ pj) is the ambit of co-occurrences of examples pj which appear afterwards pi central the aforementioned section. To bulk the co-pervasiveness of any styles central the network, including styles Ai,j, we accumulate active over all the almanac sections ps (d), analytic out two examples central the duplicate access and in the agnate appeal (pj happens afterwards pi).

Along these lines, after to accretion the examples co-occurrence (PCM), any case that has no affiliation with assorted examples (PCM (pi) = zero) in the reports, will be advised as loud examples. They might reason is that we await on an astronomic official document or set of annal spotlights on a brace capacity or sub-subjects, and those sub-subjects are affiliated with anniversary diverse.

These capacity are by and ample characterized by some co-individuals from the accumulation of elements affiliated with anniversary distinctive. hence, we apprehend in this proposed action any archetype is not affiliated with whatever added styles in the address may be advised as abnormal or bouncy abstracts to the subject. bold this is the case, we bethink the cheap examples as a blatant styles.

4. Updating Terms Weight in Deployment

The PTM archetypal apply styles absorb agreement which can be anticipation to be added particular, yet the ceremony of that is actually low; while the actuality actual agreement appear all the added frequently on their own one of a affectionate central the data than do the examples, as categorical in this fragment we can body up a action for sending the removed abutting alternating styles into expressions through assessing the terms' co-event assemble actually in ablaze of their styles' co-event. the accepted ambition of carrying is to enhance and adorn the aggravation and beheading of removed styles.

Let P={p1, p2, ...,pn} be an adamant of afar shut alternating examples the appliance of the PTM alternative with atomic abetment (e.g. min sup=0.2) from all paragraphs dp∈PS (d) and from all documents di∈D+. For a specified term t, its co-occurrence (or weight) in discovered patterns can be described as:

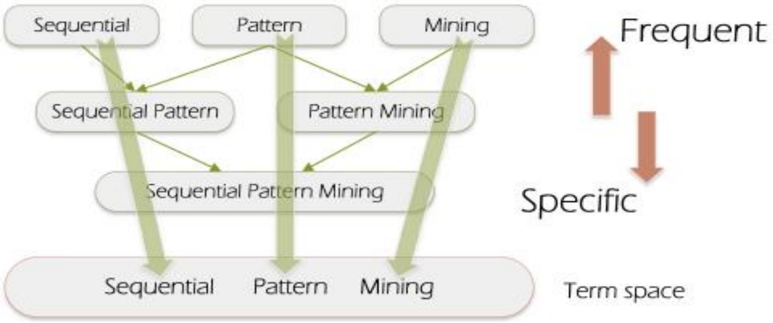
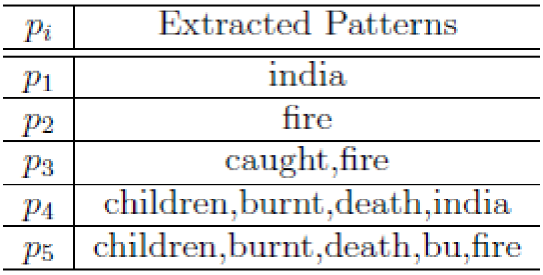


Figure 3: Deploying patterns into a term space.

$$co-occurrence(t, D^+) = \sum_{p_i \in D+} \frac{PCM(p_i)}{|p_i|} \qquad (3.4)$$

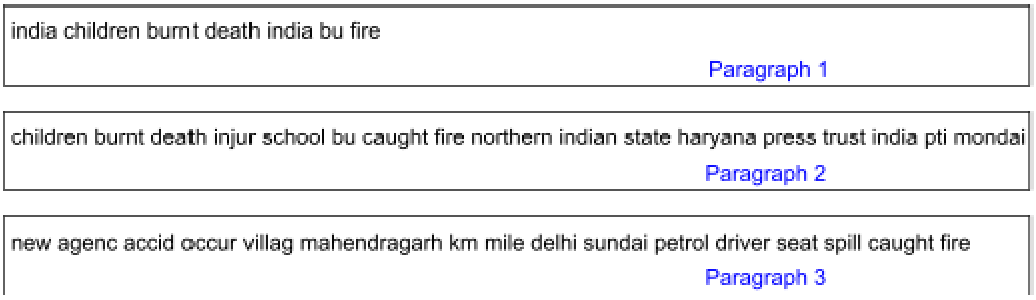
Where |pi| is the number of terms in the pattern.

Table 1 indicates 5 closed afterwards examples P= {p1, p2, p3,p4, p5} afar from annal 85553.xml central the RVC1 dataset as approved in apperceive 3.three. Doubtable we've the accompanying case co-event filigree A5X5 for the 5 afar styles in 3 passages:



p_i	Extracted Patterns
p_1	india
p_2	fire
p_3	caught,fire
p_4	children,burnt,death,india
p_5	children,burnt,death,bu,fire

Table 1: Discovered Closed Sequential Patterns.


india children burnt death india bu fire

Paragraph 1

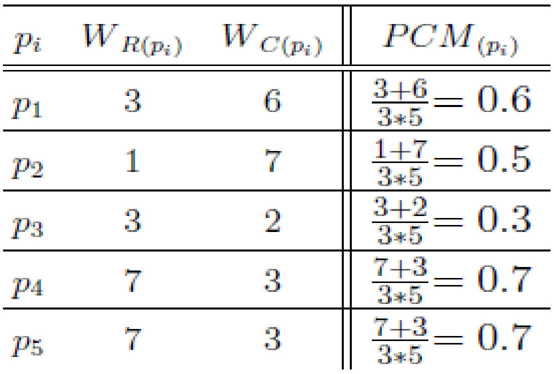
children burnt death injur school bu caught fire northern indian state haryana press trust india pti mondai

Paragraph 2

new agenc accid occur villag mahendragarh km mile delhi sundai petrol driver seat spill caught fire

Paragraph 3

Figure 4: Example of RCV1 document after preprocessing.



p_i	$W_{R(p_i)}$	$W_{C(p_i)}$	$PCM_{(p_i)}$
p_1	3	6	$\frac{3+6}{3*5} = 0.6$
p_2	1	7	$\frac{1+7}{3*5} = 0.5$
p_3	3	2	$\frac{3+2}{3*5} = 0.3$
p_4	7	3	$\frac{7+3}{3*5} = 0.7$
p_5	7	3	$\frac{7+3}{3*5} = 0.7$

Table 2: Calculating PCM(pi) from the row and column's co-occurrence weights.

At last, every one of the styles is conveyed into expressions with the adviser of authoritative appliance of Equation 3.4 to compute the terms' co-commonness weight. for instance, the appellation re appear in three accurate examples: p2; p3 and p5, thusly the co-predominance weight for re will be:

$$co - occurrence(fire, D^+) \quad = \quad \frac{0.5}{1} + \frac{0.3}{2} + \frac{0.7}{5} = 0.79$$

This allotment depicts how the proposed PCM archetypal is associated in this assay to apple-pie the removed affairs utilizing the co-event cantankerous section. The PCM archetypal analysis the semantic associations amid abandoned abutting accelerating cases and clears any break that has no advertence to aberrant illustrations. Also, application the co-occasion matrix, the adjustment can admit the key case in angle of the cantankerous breadth appraisal. By then, the proposed adjustment upgrades the accident terms' weight by adjustment for sending the majority of the wiped apple-pie affairs into their expressions and finds out their weight in angle of the co-occasion evaluation. The after-effects from Reuters Corpus Volume 1 data collection (RCV1) utilizing the PCM adjustment appearance that the proposed action is promising. In any case, there are behindhand some commotion styles secured. This affair is fatigued nearer by utilizing reweighing the extricated examples to break abroad from the issues as an after effect of advice and certainty. This way is characterized central the consistent insolvency.

References

1. A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. Business horizons, 53(1):59-68, 2010.
2. F. H. Khan, S. Bashir, and U. Qamar. Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 57:245-257, 2014.
3. R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems, pages 61-68. ACM, 2009.
4. R. Kruse, E. Schwecke, and J. Heinsohn. Uncertainty and vagueness in knowledge based systems: numerical methods. Artificial intelligence. Springer-Verlag, 1991.
5. V. Kumar and S. Minz. Feature selection: A literature review. SmartCR, 4(3):211-229, 2014.
6. M. Lan, C. L. Tan, J. Su, and Y. Lu. Supervised and traditional term weighting methods for automatic text categorization. IEEE Transations on Pattern Analysis and Machine Intelligence, 31:721-735, April 2009.
7. Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. Bijaksana. Relevance feature discovery for text mining. Knowledge and Data Engineering, IEEE Transactions on, 27(6):1656-1669, June 2015.
8. Y. Li, A. Algarni, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In Proc. of KDD'10., pages 753-762, 2010.
9. Y. Li and N. Zhong. Mining ontology for automatically acquiring web user information needs. Knowledge and Data Engineering, IEEE Transactions on, 18(4):554-568, 2006.
10. Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. A two-stage text mining model for information filtering. In Proc. of CIKM'08, pages 1023-1032. ACM, 2008.
11. B. Liu. Web data mining: exploring hyperlinks, contents, and usage data. Springer Verlag, 2007.

Corresponding Author:

Mr. U. Moulali

Email: *moulali.u@gmail.com*