IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Text Categorization using Apriori Algorithm

D.Datta[1*], A. Mitra[2], D. Nag[3], N. Roy Choudhury[4]

[1,2,3,4]Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, India

[*]*Corresponding Author: debabrata.datta@sxccal.edu*

Available online at: www.ijcseonline.org

Accepted: 13/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Knowledge exploration from the large set of data, generated as a result of the various data processing activities is an effective application of data mining. Text mining applications have also become important areas of application in the field of document processing. Document clustering has also become an important process for helping the information retrieval systems to organize vast amount of data. This can be tried with categorical data and for image categorization. At the same, time, frequent pattern mining has also become a very important undertaking in data mining. In the research work described in this paper, Apriori algorithm has been applied to generate frequent itemset and this method contains mainly two steps, viz. candidate generation and pruning techniques for the satisfaction of the desired objective. Aim of this paper is to focus on frequent itemset generation from dataset of variable length. Several steps have been executed to achieve the desired result. The primary goal has been to build a method which can be used to find significant items from a text database in an easy and efficient way.

***Keywords*—**Itemsets, Tokenization, Stemming, Apriori algorithm

I. INTRODUCTION

In the decision making process related to commercial applications, a selective approach is very essential where selective means to find the most demanding or frequent item among all. With the advent of e-commerce sites Internet has brought the entire world to the user within a single mouse click. It has also increased the necessity of preference generation. For better user experience, it is required to know the choice of each individual and here comes the main purpose of frequent itemset generation. One of the other important purposes is to use frequent pattern discovery methods in Web log data. Discovering hidden information from Web log data is called Web usage mining [3]. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behaviour of the users. This can be used for advertising purposes, for creating dynamic user profiles etc. Frequent pattern mining was first proposed by Agarwal, Imielinski and Swami for market basket analysis in the form of association rule mining. Frequent Itemset mining came into existence where it is needed to discover useful patterns in customer transaction database. Frequent patterns are patterns (itemsets, subsequences or substructures) that appear frequently in a dataset. A substructure can refer to different structural forms such as subgraphs, subtrees and sub lattices which may be combined with itemsets and subsequences. Finding frequent patterns plays an essential role in mining associations and correlations, and many other interesting relationships among data [1]. Moreover it helps in data classification clustering and many other data mining task. . It is believed that grouping similar documents together into clusters will help the users find relevant information quicker, and will allow them to focus their search in the appropriate direction. Various clustering algorithms found in literature are analyzed and the requirements for an efficient document categorization approach are identified [2]. The frequent pattern mining has become an important data mining task and a focus theme in data mining research. Frequent pattern mining is the discovery of relationships or correlations between items in a dataset. In the case of data streams, one may wish to find the frequent item sets either over a sliding window or the entire data stream [10]. Thus frequent pattern will evaluate a single user's choice after certain modification. This research work concentrates on various processes which will divide frequent pattern mining into several easier steps. It is applicable to all transaction problems and is very much applicable for websites, e-commerce sites or any transactional problem where classification of data will lead to user's choice identification. . It is helpful for user as well as marketers to improve various marketing policies. The next section discusses about the related research work. After that, discussion has been done on the proposed method and the results obtained accordingly. Finally, a concluding section has been added to discuss about the limitations and possible future scope of improvement.

II. RELATED WORK

Several algorithms have been proposed for working with sparse as well as dense data having many rows and columns. Among these most useful methods had been filtered out which can be categorized as scalable methods for mining frequent patterns.

In this regard, one of the major approaches is a method called Direct Hashing and Pruning (DHP) which is an effective hash-based algorithm for mining association rules [5]. This algorithm generates candidate (k+1)-itemsets from large k-itemsets, and large (k+1)-itemsets are found by counting the occurrences of candidate (k+1)-itemsets in the database. DHP algorithm uses a hashing technique to filter out unnecessary itemsets for the generation of the next set of candidate itemsets.

Another significant algorithm in this field is Frequent Pattern Growth (FP – Growth) algorithm which does mining operations on frequent patterns without candidate generation [13]. It is basically a two-step approach. Initially, it builds a compact data structure called the FP-tree using two passes over the dataset. Then it extracts frequent itemsets directly from the traversal through FP-Tree.

Vertical data format approach is a relatively new algorithm for fast discovery of association rules [6]. It combines featured itemsets, depending on the format of the database, decomposition techniques, and procedures used search. The algorithm not only minimizes the cost of I/O by simply making a small number of database scans, but also minimizes the cost calculations with efficient search schemes.

Rapid Association Rule Mining (RARM) is another algorithm for identifying the important itemsets from a set of transactions [7]. RARM has been proposed to further push the speed barrier so that association rule mining can be performed more efficiently in electronic commerce. To achieve large speed-ups even at low support, thresholds, RARM constructs a new data structure called Support-OrderedTrieItemset i.e. SOTrieIT.

A significant approach was proposed through Associated Sensor Pattern Mining of Data Stream (ASPMS) [8]. It was mainly applied on Wireless Sensors Network (WSN) dataset. To find the frequent patterns with single scan of database, ASPMS algorithm with ASPS-tree technique has been proposed which was used to generate associated patterns. ASPMS algorithm can extract useful information for the current window of the sensor from the stream contents in a batch-by-batch manner.

Another significant algorithm in this domain is COFI algorithm that was demonstrated to outperform state-of-the-art algorithms on synthetic data [8]. A COFI-tree is a projection of each frequent item in the FP-tree. Each COFI-tree for a given frequent item presents the co-occurrence of this item with other frequent items that have more support than it.

After going through all previously mentioned methods in the domain of frequent itemset mining, Apriori algorithm has been chosen due to its efficient approach towards identifying itemsets. Further, it is always advantageous to choose an approach which can be implemented easily using less complex programming tools. Apriori is easy to implement using any programming language which in return makes the complexity of the whole process of frequent itemset generation little less. Moreover, like other algorithms so far discussed Apriori algorithm does not require to generate any structure, rather it only requires joining and pruning operations. Naturally, no extra memory space is required for storing structures.

III. PROPOSED WORK

The proposed algorithm consists of few major steps like tokenization, removal of stop-words, stemming, frequency generation and generation of significant terms. Before implementing tokenization, all training documents are parsed to remove markup tags and special formatting using a parser [15]. The output of parser is just the content inside the body tag.

In the tokenization phase, a sequence of strings has been broken into pieces containing words, keywords, phrases and symbols. These pieces are called tokens. In the process of tokenization, some characters like punctuation marks were discarded. The tokens became the input for another process like parsing and text mining.

After tokenization was over, stop-words have been removed. Common words such as 'are', 'the', 'with', 'from' etc. that occurred in almost all documents, did not help in deciding whether a document belongs to a category or not. Such words were referred to as stop words. So, these words can be removed by identifying a list of stop-words.

After stop-word removal was over, stemming was done. This is a process which reduces any term to their stems or root variant. For example, words like "computer" or "computing" or "computation" is reduced to "compute" and similarly, "engineering" or "engineered" is reduced to "engineer". The main advantage of using stemming is to reduce computing time and space as different forms of words are stemmed to a single word. The most popular stemmer in English is the Martin Porter's stemming algorithm shown to be empirically effective in many cases. This research work has used Porter's stemming algorithm.

The next step after stemming was to build an inverted index data structure to store a mapping from contents, such as terms to its locations in a set of documents. There are two types of inverted index data structures. In the present work, record level inverted index structure has been used. A record level inverted index contains a list of references to documents for each term.

The present work has concentrated only on those terms whose document frequency was greater than two and less

than 50 and has excluded the remaining terms. Basically, rare terms have been considered to be non-informative for category prediction in global performance and hence could be removed. If terms have document frequency greater than 50, it means that these terms occur in almost half of the document. So, by these terms, one cannot distinguish between two documents and hence can be removed. In this process of significant term generation, Apriori algorithm has been applied. In the process, significant terms have been chosen to be those terms which were frequent.

IV. RESULTS AND DISCUSSION

The proposed algorithm as described in the previous section has been implemented with a system having Intel® Core™ i5-6200U processor in Windows 10 operating system. The software tool used to implement the algorithm was JDK SE8. The following dataset has been used to check the outputs of the proposed algorithm. The dataset contains 180 book names of four different subjects, viz., mathematics, physics, chemistry and computer science. The following table mentions all the data used in the present research work.

Table 1. Data Set

Book Number	Book Name
1	the epic quest to solve the world's greatest mathematical problem
2	a romance of many dimensions
3	the science of secrecy from ancient egypt to quantum cryptography
4	the man who loved only numbers
5	zero: the biography of a dangerous idea
6	a beautiful mind
7	the drunkard's walk: how randomness rules our lives
8	journey through genius: the great theorems of mathematics
9	what is mathematics?: an elementary approach to ideas and methods
10	how to lie with statistics
11	euclid's elements
12	the man who knew infinity: a life of the genius ramanujan
13	the music of the primes: searching to solve the greatest mystery in mathematics
14	the number devil: a mathematical adventure
15	uncle petros and goldbach's conjecture: a novel of mathematical obsession
16	a mathematician's apology
17	surely you're joking, mr. feynman!: adventures of a curious character
18	the joy of x: a guided tour of math, from one to infinity
19	innumeracy: mathematical illiteracy and its consequences
20	prime obsession: bernhard riemann and the greatest unsolved problem in mathematics
21	Cryptonomicon
22	men of mathematics
23	monster's battle book 1
24	a history of p
25	here's looking at euclid: a surprising excursion through the astonishing world of math
26	unknown quantity: a real and imaginary history of algebra
27	Relativity
28	the special and the general theory
29	the princeton companion to mathematics
30	a brief history of time
31	a mathematician's lament: how school cheats us out of our most fascinating and imaginative art form
32	the signal and the noise: why so many predictions fail - but some don't
33	fooled by randomness: the hidden role of chance in life and in the markets
34	proofs from the book
35	the golden ratio: the story of phi, the world's most astonishing number
36	the colossal book of mathematics
37	i am a strange loop
38	letters to a young mathematician
39	an imaginary tale: the story of the square root of minus one
40	the mathematical universe: an alphabetical journey through the great proofs, problems, and personalities
41	geometry and the imagination
42	the math book: from pythagoras to the 57th dimension, 250 milestones in the history of mathematics
43	the man who counted: a collection of mathematical adventures
44	the equation that couldn't be solved: how mathematical genius discovered the language of symmetry
45	mathematics: from the birth of numbers
46	love and math: the heart of hidden reality
47	the emperor's new mind concerning computers, minds and the laws of physics
48	the mathematical experience
49	the parrot's theorem
50	number: the language of science
51	an introduction to secondary mathematics
52	foundation of physics for scientists and engineers

53	elementary physics
54	introduction to vectors
55	chemical thermodynamics
56	electromagnetism for electronic engineers
57	electricity and magnetism
58	introduction to particle physics
59	introduction to quantum mechanics
60	introduction to lagrangian & hamiltonian mechanics
61	mechanics and oscillations
62	sound and electromagnetic waves and optics
63	fluid bed particle processing
64	introduction to solid state physics
65	quantum theory of solids
66	essential relativity
67	particle physics and introduction to field theory
68	introduction to high energy physics
69	quantum field theory
70	space, time, and gravity: the theory of the big bang and black holes
71	theoretical nuclear physics
72	fundamental astronomy.
73	hydrodynamics and hydromagnetic stability
74	principles of optics: electromagnetic theory of propagation
75	the principles of nonlinear optics
76	the theory of atomic spectra
77	theory of superconductivity
78	introduction to quantum mechanics
79	space time and black holes
80	solid state physics
81	quantum optics & quantum gases
82	basic principle of biophysics
83	a textbook of oscillations, wavesand acoustics
84	plasma physics and engineering
85	fundamentals of magnetism and electricity
86	solid state devices and electronics
87	seven brief lessons on physics
88	the physics of time
89	the cosmic web
90	the fabric of the cosmos
91	solar electricity handbook
92	mathematical methods in the physical sciences
93	classical mechanics and general properties of matter
94	fourier analysis
95	semiconductor physics and devices
96	stochastic dynamics
97	jam physics
98	university physics with modern physics
99	molecular spectroscopy
100	optoelectronic devices: advanced simulation & analysis
101	applied physics
102	royal society of chemistry
103	archaeological chemistry
104	atmospheric chemistry
105	a first course in electrode processes 2nd edition
106	basic water treatment 5th edition
107	biocatalysis in organic synthesis
108	brewing 2nd edition
109	carbohydrate chemistry and biochemistry 2nd edition
110	chemical information for chemists
111	chemical processes for a sustainable future
112	chemistry, nutrition and therapy
113	colour chemistry 2nd edition
114	compound-specific stable isotope analysis
115	comprehensive organic chemistry experiments for the laboratory classroom
116	concepts of chemical engineering for chemists 2nd edition
117	contemporary catalysis
118	detection of drug misuse
119	edible nanostructures
120	the chemistry of its components
121	forensic toxicology
122	general chemistry 4th edition
123	student solutions manual to accompany general chemistry
124	green chemistry 3rd edition
125	industrial polymer applications
126	intermolecular interactions in crystals
127	introduction to photocatalysis
128	isotope dilution mass spectrometry
129	mimicking the extracellular matrix
130	nanochemistry 2nd edition
131	physical chemistry for the chemical sciences
132	polymer structure characterization 2nd edition
133	pulse chemistry and technology
134	synthetic methods in organic electronic and photonic materials
135	tanning chemistry
136	the chemistry of cosmic dust
137	the chemistry of explosives 3rd edition
138	the chemistry of plants
139	the chemistry of textile fibres 2nd edition
140	the handbook of medicinal chemistry
141	worldwide trends in green chemistry education
142	d- and f-block chemistry
143	aromatic chemistry
144	atomic structure and periodicity
145	basic atomic and molecular spectroscopy

146	biophysical chemistry 2nd edition
147	functional group chemistry
148	heterocyclic chemistry
149	inorganic chemistry in aqueous solution
150	main group chemistry
151	become an x coder
152	data structures and algorithms with object-oriented design patterns in java
153	dive into accessibility
154	getting started with awk
155	how to design programs
156	introduction to design patterns in c++ with qt 4
157	linux device drivers
158	linux network administrator's guide
159	logic, programming and prolog
160	the opengl programming guide
161	practical common lisp
162	programming in lua
163	programming ruby: the pragmatic programmer's guide
164	the scheme programming language
165	software engineering for internet applications
166	sql for web nerds
167	structure and interpretation of computer programs
168	a tutorial on pointers and arrays in c
169	introduction to algorithms
170	structure and interpretation of computer programs
171	the c programming language
172	the pragmatic programmer: from journeyman to master
173	the art of computer programming, volumes 1-3 boxed set
174	design patterns: elements of reusable object-oriented software
175	code: the hidden language of computer hardware and software
176	the mythical man-month: essays on software engineering
177	artificial intelligence: a modern approach
178	code complete
179	the protocols (tcp/ip illustrated, volume 1)
180	advanced programming in the unix environment

With the above mentioned dataset, the following results were obtained.

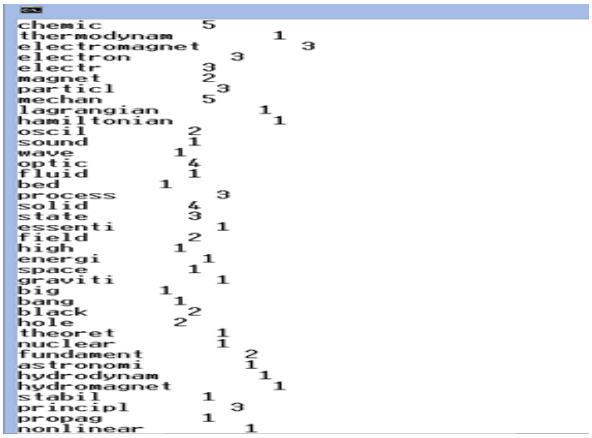

Figure 1. A snapshot of the output of the frequency generation process of the keywords

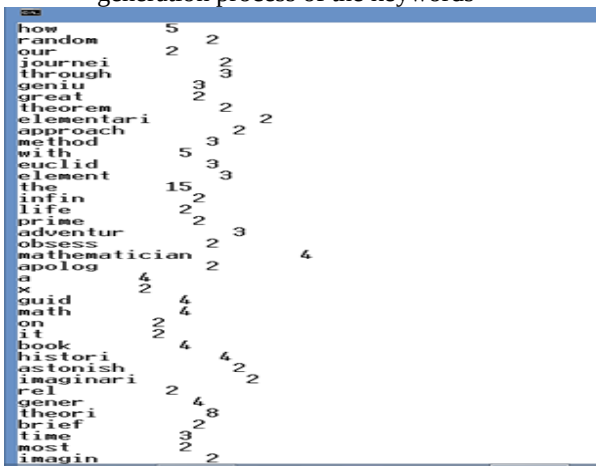

Figure 2. A snapshot of the output after the pruning step of Apriori algorithm

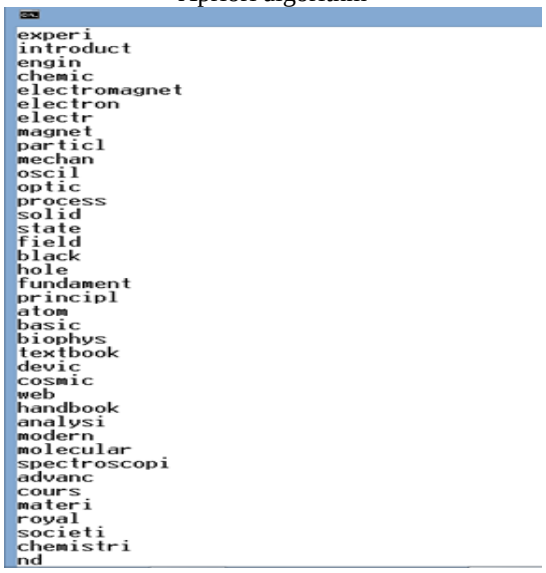

Figure 3. A snapshot of the output after the word generation

As shown in figure 1, the frequency of each significant word has been obtained and based on this result; the pruning step

as mentioned in Apriori algorithm, has been applied. Figure 2 demonstrates that. Finally, figure 3 mentions the word generation process. Accordingly, the significant terms have been obtained.

V. CONCLUSION AND FUTURE SCOPE

After executing the proposed method based on Apriori algorithm, some drawbacks were found which are depicted below:
The method needs several iterations on the specified data set. Moreover, it was difficult to find rarely occuring events.
As have been mentioned before, frequent itemset generation is the initial and the most important approach in data mining. So, with the obtained result, the next step forward may be in the field of text categorization which implies that objects are grouped into categories, usually for some specific purposes. Due to the increased availability of ever larger numbers of text documents in digital form and the ensuing need to organize them for easier use, text categorization has become one of the key techniques for handling and organizing text data. The next improvement on the proposed method as described in the present research work may be on the field of text categorization.

REFERENCES

[1] R. Agarwal, R. Srikant "Fast Algorithms for Mining Association Rules", In Proceedings Of Int. Conf. on Very Lata Bases, pp. 487 – 499, 1994.
[2] B. Babcock, S. Babu, M. Datar,R. Motwani, J. Widom, "Models and Issues in Data Stream Systems". In Proceedings Of ACM Symp. on Principles of Database Systems, pp. 1-16, 2002.
[3] R. Agrawal, T. Imielinski, and A. Swami,"Mining Association Rules between Sets of Items in Large Databases", In Proceedings of ACM-SIGMOD International Conference on Management of Data, pp. 207–216, 1993.
[4] G. Manku, R. Motwani, "Approximate Frequency Counts over Data Streams", In Proceedings of International Conference on Very Large Data Bases, pp. 346-357, 2002.
[5] S. Ozel, H. Atlay, "An Algorithm for Mining Association Rules Using Perfect Hashing and Database Pruning", Güvenir Bilkent University, Department of Computer Engineering, Ankara, Turkey.
[6] J. Reynaldo, D.B. Tonara, "Data Mining Application using Association Rule Mining ECLAT Algorithm Based on SPMF", 3rd International Conference on Electrical Systems, Technology and Information, 2017.
[7] S. Rewatkar, A. Pimpalkar, "Associated Sensor Patterns Mining of Data Stream from WSN Dataset", International Journal on Computer Science and engineering, Vol 8, Issue 10, 2016.
[8] M. El-Hajj, O.R. Zaiane, "COFI Approach for Mining Frequent Itemsets Revisited", In Proceedings of the Ninth ACM SIGMOD Workshop on Resesrach Isssues in Data Mining and Knowledge Discovery, pp. 70-75, 2004.
[9] W. Cheung, O.R. Zaïane, "Incremental Mining of Frequent Patterns Without Candidate Generation or Support Constraint", In Proceedings of the Seventh International Database Engineering and Applications Symposium, 2003.
[10] X.Y. Wang, J. Zhang, H.B. Ma, Y.F. Hu, "A New Self-Adaptive Algorithm For Frequent Pattern Mining" , In Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, pp. 13-16, 2006.
[11] S. Aggarwal, R. Kaur, "Comparative Study of Various Improved Versions of AprioriAlgorithm", International Journal of Engineering Trends and Technogy, Vol 4, Issue 4, pp 687-690, 2013.
[12] M.J. Zaki, "Parallel and Distributed Association Mining: A Survey", In Proceedings of Concurrency IEEE, Vol 7, Issue 4, pp 14-25, 1999.
[13] S. Brin, R. Motwani, J. D. Ullman S. Tsur, "Dynamic Itemset counting and Implication Rules for Market Basket Data", ACM SIGMOD, Vol 26, Issue 2, pp. 255-264, 1997.
[14] Tsay, Y. Jiuan, T. J. Hsu, Y. J. Rung, "FIUT: A New Method for Mining Frequent Itemsets" Information Sciences, Vol 179, Issue 11, 2009.
[15] G.Pyun, U.Yun, K.H.Ryu, "Efficient Frequent Pattern Mining Based on Linear Prefix Tree", Knowledge-Based Systems,Vo. 55, Issue 1, pp 125-139, 2014.
[16] D. Xin, J. Han, X. Yan, H. Cheng, "Mining Compressed Frequent-Pattern Sets", Proceedings of the Thirty First international Conference on Very Large Data Bases, pp709-720, 2005..

Authors Profile

Mr. D Datta pursued Master of Technology from University of Calcutta, India and he is currently pursuing his Ph.D. in Technology from the same university. He is an Assistant Professor in the department of Computer Science, St. Xavier's College (Autonomous), Kolkata, India He is a life member of IETE. He has published more than 20 research papers in reputed international journals and conferences His main research work focuses on Data Analysis. He has more than 10 years of teaching experience and has more than 4 years of Research Experience.

Miss A Mitra pursued her B.Sc. in Computer Science from St. Xavier's College (Autonomous), Kolkata, India and is currently doing her Masters in Computer Science from the same institute.

Mr. D Nag has completed his B.Sc. in Computer Science from St. Xavier's College (Autonomous), Kolkata, India.

Miss N Roy Choudhury pursued her B.Sc. in Computer Science from St. Xavier's College (Autonomous), Kolkata, India and is currently doing her Masters in Computer Science from the same institute.

