Finding similar documents using different clustering techniques

[1]R.MADHU MOHAN PRASAD
M.Sc. (Computer Science)
Besant Theosophical College, Madanapalle.

[2]D.VENKATA SIVA REDDY
Head of Department
Besant Theosophical College, Madanapalle

ABSTRACT:

Text clustering is an important application of data mining. It is concerned with grouping similar text documents together. In this paper, several models are built to cluster capstone project documents using three clustering techniques: k-means, k-means fast, and k-medoids. Our datatset is obtained from the library of the College of Computer and Information Sciences, King Saud University, Riyadh. Three similarity measure are tested: cosine similarity, Jaccard similarity, and Correlation Coefficient. The quality of the obtained models is evaluated and compared. The results indicate that the best performance is achieved using k- means and k-medoids combined with cosine similarity. We observe variation in the quality of clustering based on the evaluation measure used. In addition, as the value of k increases, the quality of the resulting cluster improves. Finally, we reveal the categories of graduation projects offered in the Information Technology department for female students.

INTRODUCTION:

Today, with the rapid advancements in technology we are able to accumulate huge amounts of data of different kinds. Data mining emerged as a field concerned with the extraction of useful knowledge from data 1. Data mining techniques have been applied to solve a wide range of real-world problems. Clustering is an unsupervised data mining technique where the labels of data objects are unknown. It is the job of the clustering technique to identify the categorisation of data objects under examination. Clustering can be applied to different kinds of data including text. When dealing with textual data, objects can be documents, paragraphs, or words 2. Text clustering refers to the process of grouping similar text documents together. The problem can be formulated as follows: given a set of documents it is required to divide them into multiple groups, such that documents in the same group are more similar to each other than to documents in other groups. There are many applications of text clustering including: document organisation and browsing, corpus summarisation, and document classification3.

Traditional clustering techniques can be extended to deal with textual data. However, there are many challenges in clustering textual data. The text is usually represented in high dimensional space even when it is actually small. Moreover, correlation between words appearing in the text needs to be considered in the clustering task. The variations in document

sizes is another challenge that affects the representation. Thus, the normalisation of text representation is required2. In this paper, we use data mining techniques in order to cluster capstone projects in information technology. In particular, we study graduation projects offered in the Information Technology department (IT) for female students at the College of Computer and Information Sciences, King Saud University, Riyadh. The goal is to to reveal the areas that the department encourages students to work on. The results of the study will be beneficial to both students and decision makers. For students, clustering graduation projects will help them find previous projects related to their own project idea. The study will also help the administration make right decisions when approving project proposals. We apply and compare three clustering techniques: k-means 4, k-means fast 5, and k-mediods 6. In addition, three similarity measures are used to form clusters: cosine similarity 7, Jaccard similarity, and Correlation Coefficient 1. The goal of the comparison is to find the best combination of clustering technique and similarity measure and to study the effect of increasing the number of clusters, k. The rest of the paper is organised as follows: In Section 2, we review some of the literature in the field of text clustering. Section 3, describes our dataset, the steps taken to prepare it for data mining, and the data mining techniques and the similarity measures used in our experiment. The cluster evaluation measures and our main findings are discuss.

2. Literature Review

Text clustering is one of the important applications of data mining. In this section, we review some of the related work in this field. Luo et al. 3 used the concepts of document neighbors and links in order to enhance the performance of k-means and bisecting k-means clustering. Using a pairwise similarity function and a given similarity threshold, the neighbors of a document are the documents that are considered similar to it. A link between two documents is the number of common neighbors. The concepts were used in the selection of initial cluster centroids and in document similarity measuring. Experimental results using 13 datasets showed better performances as compared to the standard algorithms. Bide and Shedge 8 proposed a clustering pipeline to improve the performance of k-means clustering. The authors adopted a divide-and-conquer approach to cluster documents in the 20 Newsgroup dataset 9. Documents were divided into groups where preprocessing, feature extraction, and k-means clustering were applied on each group. Document similarity was calculated using the cosine similarity measure. The proposed approach achieved better results as compared to standard k-means in terms of both cluster quality and execution time. Mishra et al. 10 used k-means technique to cluster documents based on themes present in each one. The main assumption was that a document may deal with multiple topics. The proposed approach, called inter-passage based clustering, was applied to cluster document segments based on similarity. After segments were preprocessed, keywords were identified for each segment using term frequency-inverse document frequency 11 and sentiment polarity scores 12. Each segment was then represented using keywords and a segment score was calculated. Finally, k-means was applied to all segments. The resulting clusters showed high intra-cluster similarity and low inter-cluster similarity. In general, algorithms used

for clustering texts can be divided into: agglomerative, partitioning-based, and probabilisticbased algorithms 13. Agglomerative algorithms iteratively merge documents into clusters based on pairwise similarity. The resulting clusters are organised into a cluster hierarchy (also dendogram). In partitioning algorithms, documents are split into mutually exclusive (non-hierarchical) clusters. The splitting process optimises the distance between documents within a cluster. Probabilistic clustering is based on building generative models for the documents. Partitioning algorithms for text clustering have been extensively studied in the literature. This is mainly due to the low computational requirements as compared to other clustering algorithms. In this paper, we choose to utilize three partitioning-based algorithms: k-means 4, k-means fast 5, and k-mediods 6 in order to cluster capstone projects.

3. Methodology

3.1. Data collection and preprocessing

The dataset was collected manually from the library of the College of Computer and Information Sciences, King Saud University, Riyadh. We selected capstone projects with dates between 2010 to 2014. A total of 63 projects were collected. For each project, the following attributes were considered: project title, abstract, and supervisor name. Pre-processing was conducted as follows: 1. Tokenization: the first step was to split text into element called tokens 14. A token can be a symbol, a word, a phrase, or a sentence. We split our datatset into tokens using whitespace as a delimiter. 2. Filtering: The result of the tokenisation step was filtered to remove meaningless words. Filtering was done based on minimum length. All tokens with lengths less than three characters were removed. 3. Stemming: this is an important step in text mining where words are reduced to their root forms. 4. Cases Transformation: finally, all the words were converted to lowercase. 3.2. Document Representation The vector space model 7 is a common representation of text documents. Let D be a collection of documents and let T = {t1, t2, ..., tn} be the set of terms appearing in D. A document x ∈ D can be represented as an n−dimensional vector in the term space T. Let wx,ti be the number of times a term ti∈ T appears in x, then the vector of x is defined as: x = {wx,t1 , wx,t2 , ..., wx,tn } (1) 3.3. Data Mining Clustering Algorithms: k-means and k-medoids are well-known and widely applicable clustering algorithms. Here, we provide a brief description of these algorithms. • k-means 4 is an iterative clustering algorithm. It is based on partitioning data points into k clusters using the concept of centroid. The cluster centroid is the mean value of the data points within a cluster. The produced partitions feature high intra-cluster similarity and inter-cluster variation. The number of clusters, k, is a predetermined parameter of the algorithm. k-means works as follows: 1) k data points are arbitrarily selected as cluster centroids. 2) the similarity of each data point to each cluster centroid is calculated. Then data point are re-assigned to the cluster of the closest centroid. 3) the k centroids are updated based on the newly assigned data points. 4) steps 2 and 3 are repeated until convergence is reached. • k-medoid 6 is a partitioning-based clustering algorithm similar to

k-means. However, the k-medoid algorithm uses actual data points to represent clusters. The algorithm is less sensitive to outliers than k-means and works as follows: 1) k data points are arbitrarily selected to form the set of current cluster representatives (medoids). 2) the remaining data points are assigned to the cluster of the closest representative. 3) a data point that is not in the current set of cluster representatives is randomly selected. 4) the total cost of replacing one of the cluster representative points with the randomly selected one is calculated. 4) the replacement takes place only if the quality of the resulting clusters is improved. 5) steps 2-4 are repeated until no improvement can be achieved. • k-means fast 5 is an accelerated version of k-means where many un-necessary distance calculations are avoided using triangle inequality. The k-means fast algorithm is suitable for larger values of k and for datasets with large number of attributes. However, it requires more memory. Similarity Measures: There are many metrics for measuring document similarity. We focus on three common measures in this domain which are: cosine similarity 7, Jaccard similarity coefficient, and Correlation Coeffiecent.

Cosine similarity measures the cosine of the angle between the vectors of two documents. Given two vectors x and y, each of length n, the cosine similarity can be calculated as follows: cos(x, y) = x.y ||x||||y|| (2) where ||x|| = x2 1 + x2 2 + ... + x2 n. The Jaccard similarity coefficient, also known as Jaccard index, is a popular measure of similarity and is calculated as follows: Jaccard(x, y) = q q + r + s (3) where, q is the total number of terms that are present in both documents, r is total number of terms that are present in x but not in y, and s is the total number of terms that are present in y but not in x. The value of both cosine and Jaccard range between 0 (no similarity) and 1 (identical matches). The correlation coefficient can be used to measure the degree of relatedness for two vectors. The value of correlation coefficient ranges from −1 (negative correlation) and 1 (positive correlation). The correlation coefficient can be calculated as follows: r(x, y) = nn t=1 w(x, t).w(y, t) − T Fx.TFy

---

[nn t=1 w(x, t)2 − T F2 x].[nn t=1 w(y, t)2 − T F2 y] (4) where, T Fx = n t=1 w(x, t) 4. Experimental Results Here, we cluster our dataset using k-means 4, k-means fast 5, and k-mediods 6. With each clustering technique, we build models using different values of k and the three similarity measures described above. The RapidMiner 15 platform was used in our experiment. This open source platform provides a friendly GUI and supports all the steps of Knowledge Discovery from Data, including: data pre-processing, data mining, model validation, and result visualisation. Figure 1 shows the main steps of this study. 4.1. Evaluation Measures We evaluated and compared the quality of the obtained clustering models using two cluster evaluation measures: the average within cluster distance 15 and Davies-Bouldin Index (DB) 16. The average within cluster distance is defined as the average of the distance between a cluster centroid and all elements in a cluster. As for DB, given as set of clusters, this metric measures the average similarity between each cluster and its most similar one. This metric is calculated as follows: DB = 1 n ni=1,i=j max σi + σj d(ci, cj)) (5) where n is the number of clusters, σi is the average distance of all elements in cluster i to the cluster centroid ci, and σj is the average

distance of all elements in cluster j to the cluster centroid cj , and d(ci, cj) is the distance between the clusters centroids ci and andcj . Since the optimal clusters should be compact and have the least similarity to each other, the value of DB should be minimised. 4.2. Result Discussion In this section, we discuss the quality of the obtained clustering models based on the values of the clustering evaluation measures. We compare all the obtained models to find the best combination of clustering technique and similarity measure. In addition, we look into individual clustering algorithms to find for each,

1. The accuracy of the clustering techniques using cosine similarity

Average within cluster distance Davis Bouldin Index

k	KM	KMF	KMD	KM	KMF	KMD
5	0.872	0.872	1.696	4.623	4.623	
1.788	6	0.852	0.852	1.636	4.082	
4.082	1.759	7	0.834	0.834	1.602	
3.962	3.962	1.683	8	0.816	0.816	
1.571	3.685	3.685	1.711	9	0.792	

Based on the average within cluster distance, the results indicate that k-means and k-means fast perform similarly when the cosine similarity is used. This could be partially due to the ability of the cosine similarity measure to ignore document length. However, k-means outperforms both k-means fast and k-medoids for the Jaccard similarity and correlation coefficient. In terms of DB index, k-medoids shows better performance than k-means and k-means fast for all similarity measures. The worst performance is obtained with k-means fast and Jaccard similarity. For all clustering techniques, the best average within cluster distance is achieved when the cosine similarity is used. We observed variation in the quality of clustering of k-means and k-means fast. The two clustering techniques show better quality when the average within cluster distance is used. As for k-medoids, the quality of clustering is similar regardless of the evaluation measure used. We found that the quality of clustering models improves as the value of k increases. Overall, the best performance is obtained using k-means and k-medoids combined with cosine similarity. As shown in Figure 2, we found that capstone project ideas can be generally divided into the following categories: E-health applications, Arabic and Islamic applications, location-based applications, voice, image, and signal recognition, games, and e-learning applications.

2. The accuracy of the clustering techniques using Jaccard similarity

Average within cluster distance			Davis Bouldin Index		
k	KM	KMF	KMD	KM	KMF	K	MD
5	0.882	0.897	1.697	4.719	∞	1.810
6	0.862	0.890	1.665	4.365	∞	1.790
7	0.835	0.876	1.631	3.815	∞	1.745
8	0.819	0.958	1.597	3.650	∞	1.718
9	0.802	0.896	1.563	3.392	∞	1.675
10	0.786	0.895	1.527	3.212	∞	41.666

The accuracy of the clustering techniques using correlation coefficient

Average within cluster distance				Davis Bouldin Index		
k	KM	KMF	KMD	KM	KMF	KMD
5	0.882	0.884	1.720	4.691	4.414	1.817
6	0.864	0.868	1.667	4.367	4.161	1.783
7	0.840	0.855	1.639	3.905	4.113	1.747
8	0.836	0.857	1.608	∞	∞	∞
9	0.804	0.848	1.569	3.457	∞	1.680
10	0.789	0.846	1.538	3.330	∞	1.609

.

Conclusion :

We built several clustering models for graduation project documents at King Saud University. Three cluster similarity measures were tested and the quality of the resulting clusters was evaluated and compared. We found that the best performance can be obtained using k-

means and k-medoids combined with cosine similarity. The documents in our dataset were of various lengths and fell into different topics. Since the cosine similarity measure is independent of document length, it was able to better deal with our dataset. There was a variation in the quality of clustering based on the cluster evaluation measure used. We also found that as the value of k increased, the quality of the resulting clusters improved. Finally, we concluded that project ideas usually fall into the following categories: E-health applications, Arabic and Islamic applications, location-based applications, voice, image, and signal recognition, games, and e-learning applications. As a future work, we plan to build a system using these clustering techniques to help students find similar projects. The system should also serve as a repository of capstone project documents, since no similar system exists.

References

1. Han, J., Kamber, M.. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 3rd ed.; 2011. ISBN 978-0-12-381479-1.

2. Aggarwal, C.C., Zhai, C.. A Survey of Text Clustering Algorithms. In: Aggarwal, C.C., Zhai, C., editors. Mining Text Data. Springer US; 2012, p. 77–128.

3. Luo, C., Li, Y., Chung, S.M.. Text document clustering based on neighbors. Data & Knowledge Engineering 2009;68(11):1271–1288.

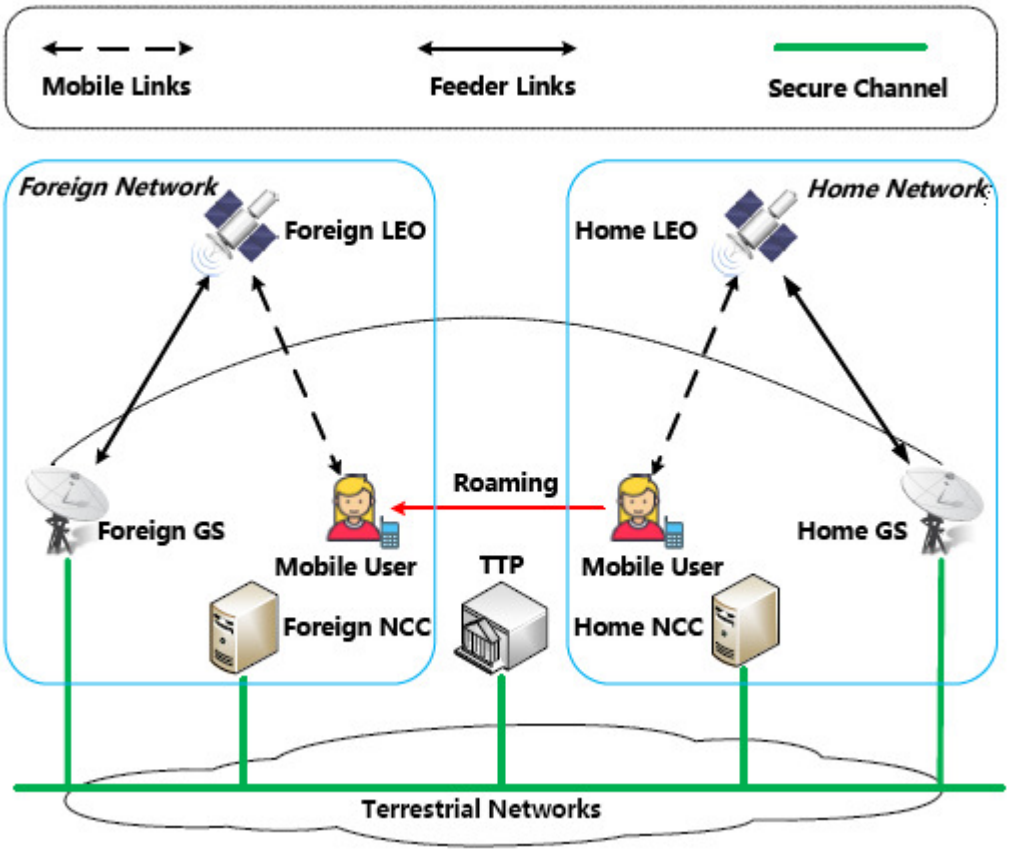
4. Hartigan, J.A.. Clustering Algorithms. New York, NY, USA: John Wiley & Sons, Inc.; 99th ed.; 1975. ISBN 978-0-471-35645-5.

5. Elkan, C.. Using the Triangle Inequality to Accelerate k-Means. In: Fawcett, T., Mishra, N., editors. Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA. AAAI Press; 2003, p. 147–153.

6. Kaufman, L. and Rousseeuw, P.J., . Clustering by means of Medoids. In: Y. Dodge and North-Holland, , editor. Statistical Data Analysis Based on the L1-Norm and Related Methods. Springer US; 1987, p. 405–416

. 7. Blair, D.C.. Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50. Journal of the American Society for Information Science 1979;30(6):374–375.

Architecture:



The sample of giving worldwide meandering in forms of systems makes it crucial for the SIN to present wandering help of its wandering clients. The meandering state of affairs in SIN is without loss of all inclusive announcement, we simply do not forget the framework show that patron wanders between the homogeneous SINs, and the scenario of meandering to SIN from specific heterogeneous structures (e.g., cell structures) is equal to this. The framework display in our plan contains of a worldwide disconnected confided in outsider (TTP) and some areas, and every area incorporates a machine manage recognition (NCC), door stations (GSs), low earth circle satellites (LEOs) and transportable clients. Following delineates the capacities and duties of every element:

• TTP is chargeable for overseeing and appropriating open/private key sets for NCCs in numerous spaces. These keys are utilized for verifying amongst those NCCs, with the goal that they are able to alternate records effectively.

• NCC is the administration of its device area. It offers enlistment and accreditation to clients to get to the house/outdoor system.

• GS is a middle element among the NCC and LEOs. It buddies with the NCC via the earthly structures, and gives a floor interface to LEOs. • LEO is the passage for customers to get to the system. With the satellite assembling innovation progression, nowadays LEO satellites could have positive figuring skills to execute some thoughts boggling capacities.

• Users get to the machine to gather its club administrations. In this paper, we bear in mind the situation wherein a meandering client is out of its home machine and traveling an out of doors system.

Conclusion:

Space records arrange (SIN) can wreck provincial confinements and grant extra sizable inclusion contrasting and commonplace Internet. The pattern of wandering to SIN might be another aspect of things to come back organize, which requires planning every other meandering validation conspire for SIN. While challenges exist for structuring a meandering affirmation framework for SIN because of its splendid condition (e.g., the dynamic and insecure topology, the relatively uncovered connections, the lengthy inertness). Propelled via the significance of customer confirmation deferral and obscurity for wandering in SIN, we shape an unknown and brief meandering verification convention (named AnFRA). In AnFRA, we use the collection mark and underline the affirmation of remote LEO (FLEO), that implies the FLEO can legitimately approve wandering clients to get to the outdoor machine without the regular inclusion of home device manage consciousness (HNCC) and without security divulgence. Besides, a denial aspect dependent explicitly for the framework is joined into the meandering verification plan to help clients' repudiation. Despite the reality that a touch degree of overhead is gotten inferable from the repudiation system.

References:

[1] M. Perry, K. O'hara, A. Sellen, B. Brown, and R. Harper, "Dealing with mobility: understanding access anytime, anywhere," ACM Transactions on Computer-Human Interaction, vol. 8, no. 4, pp. 323–347, 2001.

[2] J. Mukherjee and B. Ramamurthy, "Communication technologies and architectures for space network and interplanetary internet," IEEE Communications Surveys & Tutorials, vol. 15, no. 2, pp. 881–897, 2013.

[3] G. Miao, J. Zander, K. W. Sung, and S. B. Slimane, Fundamentals of Mobile Data Networks. Cambridge University Press, 2016.

[4] Q. A. Arain, D. Zhongliang, I. Memon, S. Arain, F. K. Shaikh, A. Zubedi, M. A. Unar, A. Ashraf, and R. Shaikh, "Privacy preserving dynamic pseudonymbased multiple mix-zones authentication protocol over road networks," Wireless Personal Communications, vol. 95, no. 2, pp. 505–521, 2017.

[5] Y. Hu and V. O. Li, "Satellite-based internet: a tutorial," IEEE Communications Magazine, vol. 39, no. 3, pp. 154– 162, 2001.

[6] T. B. Zahariadis, K. G. Vaxevanakis, C. P. Tsantilas, N. A. Zervos, and N. A. Nikolaou, "Global roaming in next-generation networks," IEEE Communications Magazine, vol. 40, no. 2, pp. 145–151, 2002.

[7] F. Li, L. Yang, W. Wu, L. Zhang, and Z. Shi, "Research status and development trends of security assurance for space-ground integration information network," Journal on Communications, vol. 37, no. 11, pp. 156–168, 2016.

[8] Y. Jiang, C. Lin, X. Shen, and M. Shi, "Mutual authentication and key exchange protocols for roaming services in wireless mobile networks," IEEE Transactions on Wireless Communications, vol. 5, no. 9, pp. 2569– 2577, 2006.

[9] P. Gope and T. Hwang, "Lightweight and energy-efficient mutual authentication and key agreement scheme withuser anonymity for secure communication in global mobility networks," IEEE Systems Journal, vol. 10, no. 4, pp. 1370–1379, 2016.

[10] I. F. Akyildiz, H. Uzunalioglu, and M. D. Bender, "Han- ˇ dover management in low earth orbit (LEO) satellite networks," Mobile Networks and Applications, vol. 4, no. 4, pp. 301–310, 1999.

[11] D. He, J. Bu, S. Chan, C. Chen, and M. Yin, "Privacypreserving universal authentication protocol for wireless communications," IEEE Transactions on Wireless Communications, vol. 10, no. 2, pp. 431–436, 2011.

[12] G. Yang, Q. Huang, D. S. Wong, and X. Deng, "Universal authentication protocols for anonymous wireless communications," IEEE Transactions on Wireless Communications, vol. 9, no. 1, 2010.

[13] J. S. Warner and R. G. Johnston, "GPS spoofing countermeasures," Homeland Security Journal, vol. 25, no. 2, pp. 19–27, 2003.

[14] J. Lei, Z. Han, M. A. Vazquez-Castro, and ´ A. Hjorungnes, "Secure satellite communication systems design with individual secrecy rate constraints," IEEE Transactions on Information Forensics and Security, vol. 6, no. 3, pp. 661–671, 2011.

[15] J. A. Larcom and H. Liu, "Modeling and characterization of GPS spoofing," in Proceedings of 2013 IEEE International Conference on Technologies for Homeland Security (HST 2013). IEEE, 2013, pp. 729–734.

[16] G. Zheng, P.-D. Arapoglou, and B. Ottersten, "Physical layer security in multibeam satellite systems," IEEE Transactions on wireless communications, vol. 11, no. 2, pp. 852–863, 2012.

[17] H. Cruickshank, "A security system for satellite networks," in Proceedings of Fifth International Conference on Satellite Systems for Mobile Communications and Navigation. IET, 1996, pp. 187–190. [18] M.-S. Hwang, C.-C. Yang, and C.-Y. Shiu, "An authentication scheme for mobile satellite communication systems," ACM SIGOPS Operating Systems Review, vol. 37, no. 4, pp. 42–47, 2003.

[19] C.-L. Chen, K.-W. Cheng, Y.-L. Chen, C. Chang, and C.- C. Lee, "An improvement on the self-verification authentication mechanism for a mobile satellite communication system," Applied Mathematics & Information Sciences, vol. 8, no. 1L, pp. 97–106, 2014.

[20] W. Zhao, A. Zhang, J. Li, X. Wu, and Y. Liu, "Analysis and design of an authentication protocol for space information network," in Proceedings of 2016 Military Communications Conference (MILCOM 2016). IEEE, 2016, pp. 43–48.

[21] J.-L. Tsai and N.-W. Lo, "Provably secure anonymous authentication with batch verification for mobile roaming services," Ad Hoc Networks, vol. 44, pp. 19–31, 2016.

[22] D. Wang, H. Cheng, D. He, and P. Wang, "On the challenges in designing identity-based privacy-preserving authentication schemes for mobile devices," IEEE Systems Journal, vol. 12, no. 1, pp. 916–925, 201