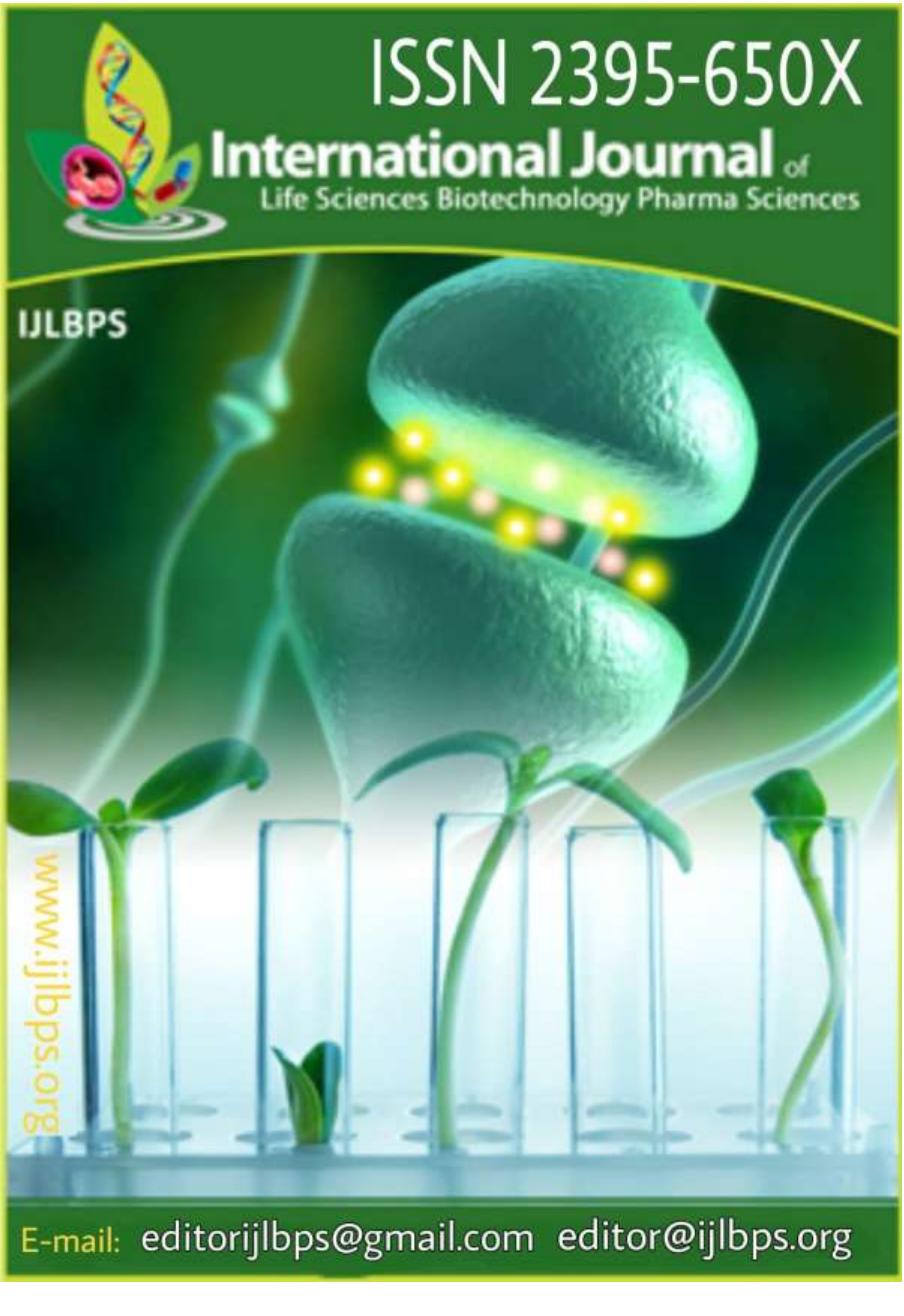

ISSN 2395-650X
International Journal of
Life Sciences Biotechnology Pharma Sciences
IJLBPS
www.ijlbps.org
E-mail: editorijlbps@gmail.com editor@ijlbps.org

International Journal of Life Sciences Biotechnology Pharma Sciences

Volume 10 Issue 2 2022

SOURCE RETRIEVAL AND TEXT ALIGNMENT CORPUSCONSTRUCTION FOR PLAGIARISM DETECTION

T.KEERTHANA, M.Tech
Department of IT
Gokaraju Rangaraju Institute of Engineering and Technology

Abstract-
For the task of source retrieval, we focus on the process of Download Filtering. For the procedure from lumping to seek control, we go for high review, and for the procedure of download separating, we dedicate to enhance accuracy. A vote- based methodology and an order based methodology are fused to channel the seeking results to get the written falsification sources. For the undertaking of content arrangement corpus development, we depict the strategies we use to build the Chinese written plagiarism cases. At last, we report the statistics of text alignment dataset submissions.

I.Introduction

Plagiarism detection in PAN is divided into two separated subtasks: source recovery and content arrangement [1]. As of late, the last subtask is changed to corpus development where members are needed to furnish examples of literary theft cases with their source reports.

These cases can happen in two structures: true examples and produced (fake or reproduced) tests. In this paper, as a member of corpus development subtask in PAN 2015, we assess other submitted corpora from the perspective of value and authenticity of unoriginality cases physically, and furthermore examine measurable data of corpora. The corpora are in various dialects, and even there possibly cross-lingual corpora in which source reports are in various dialect from suspicious ones.Worldwide Metadata

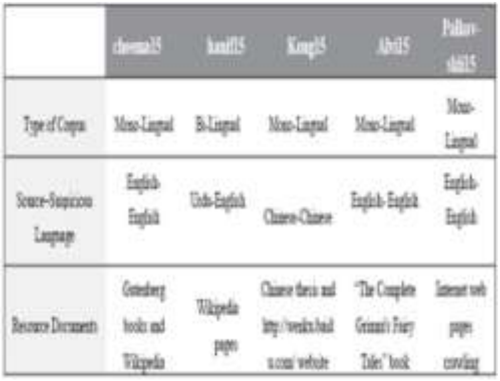
In this area, we portray the worldwide metadata of corpora under assessment. Table 1 demonstrates the metadata of five corpora. As can be appeared in the Table, there is one bilingual and four mono-lingual corpora in English and Chinese. The last line indicates information assets of reports. It very well may be seen that much of the time, the assets for source and suspicious archives are the equivalent.

Table 1: Global information of prepared corpora

Corpora Statistical Information

For evaluation of corpora based on statistical information, we categorized the statistical information

in three different aspects: The first view describes the numerical information about corpora such as number and length of documents and suspicious cases which has been shown in Table 2.



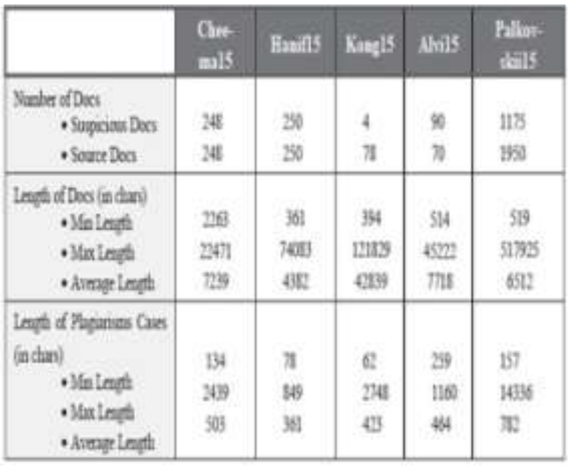
	cheema15	hanif15	Kong15	Alvi15	Palkov-skii15
Type of Corpus	Mono-Lingual	Bi-Lingual	Mono-Lingual	Mono-Lingual	Mono-Lingual
Source-Suspicious Language	English-English	Urdu-English	Chinese-Chinese	English-English	English-English
Resource Documents	Gutenberg books and Wikipedia	Wikipedia pages	Chinese thesis and http://wenku.baidu.com/ website	"The Complete Grimm's Fairy Tales" book	Internet web pages crawling

Table 2 shows the statistical information of the submitted five corpora in text alignment subtask. We categorized statistical information of corpora in three rows:The principal push exhibits the quantity of suspicious and source records. In second line,

the length of reports has been controlled by Min, Average and Max classifications. In the third line, we have demonstrated the data separated from XML documents that give it is possible that coordinated or one-to-numerous connections among source and suspicious sections. This data demonstrates the length of counterfeiting parts.The greater part of corpora have around equivalent number of source and suspicious reports. In spite of
the fact that the corpus of Kong15 has only four suspicious reports, yet it ought to be noticed that it contains genuine copyright infringement cases in suspicious parts in the corpus.

Recreation of real instances of unoriginality cases necessitates that the suspicious archives have enough length to install some copied sections inside their content. As appeared in the table, the archives in Kong15's corpus have most prominent normal length, while the reports in Cheema15's corpus have most noteworthy least length. In this way, in them two, we can possibly embed increasingly and bigger appropriated sections with the end goal to build suspicious reports. Three of corpora have around same normal length of written falsification cases; Due to the short length of counterfeiting cases in Hanif15's corpus, even with a medium rate of muddling, the literary theft location will turn out to be more troublesome. Then again, Palkovskii15 corpus has long written falsification cases and needs to perform more changes with the end goal to assemble a testing corpus. These will be examined later in this paper.

Table 2: The statistical information of the five corpora



	Chee-ma15	Hanif15	Kong15	Alvi15	Palkov-skii15
Number of Docs					
• Suspicious Docs	248	250	4	90	1175
• Source Docs	248	250	78	70	1950
Length of Docs (in chars)					
• Min Length	2263	361	394	514	519
• Max Length	23471	74083	121829	45222	517925
• Average Length	7239	4382	42839	7718	6512
Length of Plagiarisms Cases (in chars)					
• Min Length	134	78	62	259	157
• Max Length	2439	849	2748	1160	14336
• Average Length	503	361	423	464	782

Additional data likewise can be separated from referenced XML records, for example, obscurity procedure. A few members have utilized one sort of jumbling, for example, Cheema15 and Hanif15 which connected reenacted obscurity in their corpora. Kong15 corpus incorporates simply genuine confusion methodology of counterfeiting with no additional pieces to suspicious records, where everyone of suspicious archives have sections either are the literary theft cases or can possibly be written falsification.

Then again, two members have numerous jumbling techniques in their corpora: Alvi15 corpus has utilized three sorts of confusion: "retelling-human" is like reproduced obscurity; "character-substitution" and "programmed" is like fake muddling. "Character- Substitution" muddling trades vowel sounds with same character glyphs however with various Unicode. Likewise Palkovskii15 corpus covers four sorts of confusion: "None" muddling which is a precise of parts, "cyclic Translation", "outline obscurity" and "irregular jumbling".

II. Literature Work

Y. Zheng, “Trajectory data mining: an overview” The advances in area procurement and versatile registering systems have produced enormous spatial direction information, which speak to the portability of an assorted variety of moving items, for example, individuals, vehicles, and creatures. Numerous strategies have been proposed for handling, overseeing, and mining direction information in the previous decade, cultivating an expansive scope of uses. In this article, we direct a deliberate overview on the real examination into direction information mining, giving a scene of the field and also the extent of its exploration themes. Following a guide from the determination of direction information, to direction information preprocessing, to direction information the executives, and to an assortment of mining undertakings, (for example, direction design mining, anomaly location, and direction arrangement), the review investigates the associations, connections, and contrasts among these current procedures. This review additionally presents the strategies that change directions into other information positions, for example, diagrams, networks, and tensors, to which more information mining and machine learning procedures can be connected. At long last, some open direction datasets are exhibited. This overview can help shape the field of direction information mining, giving a fast comprehension of this field to the network.

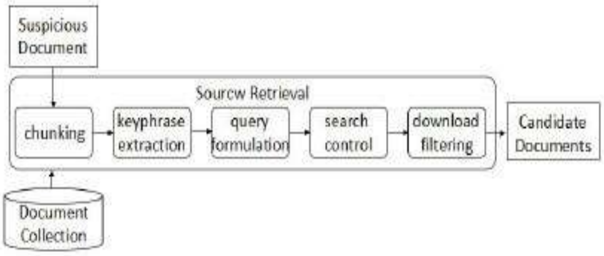
Y. Li, J. Luo, C.Y. Chow, K.- L. Chan, Y. Ding, and

F. Zhang, "Becoming the charging station arrange for electric vehicles with direction information analytics"With ongoing advances in battery innovation and the subsequent diminishing in the charging times, open charging stations are turning into a practical alternative for Electric Vehicle (EV) drivers. Simultaneously, across the board utilizationof area GPS beacons in cell phones and wearable gadgets makes it conceivable to follow singular dimension human developments to an extraordinary spatial and fleeting grain. Spurred by these improvements, we propose a novel approach to perform information driven advancement of EV charging stations area. We define the issue as a discrete streamlining issue on a topographical matrix, with the target of covering the whole interest district while limiting a proportion of drivers' distress. Since ideally taking care of the issue is computationally infeasible, we present computationally proficient, close ideal arrangements dependent on covetous and hereditary calculations. We at that point apply the proposed technique to streamline EV charging stations area in the city of Boston, beginning from monstrous wireless informational collections covering 1 million clients more than 4 months. Results demonstrate that hereditary calculation based streamlining gives the best arrangements regarding drivers' uneasiness and the quantity of charging stations required, which are both diminished about 10% when contrasted with a randomized arrangement.D. Yan, Z. Zhao, and W. Ng, "Productive preparing of ideal gathering point inquiries in Euclidean space and street networks",Finding an ideal gathering point (OMP) for a gathering of individuals (or an arrangement of items) at various areas is a critical issue in spatial inquiry handling. There are some genuine applications identified with this issue, for example, deciding the area of a meeting setting, choosing the get area of a vacationer transport, and arranging strategies of man-made reasoning continuously methodology diversions. In this paper, we consider OMP questions in the accompanying two spatial settings which are basic, all things considered, applications: Euclidean space and street systems. In the setting of Euclidean space, we propose a general system for noting all OMP inquiry variations and furthermore distinguish the best calculations for specific sorts of OMP inquiries in the writing. In the setting of street systems, we think about how to get to just piece of the street organize and look at part of the applicants.Y. Wang, Y. Zheng, and Y. Xue, "Travel time estimation of a way utilizing inadequate directions," Here propose a citywide and constant model for evaluating the movement time of any way (spoke to as a grouping of associated street fragments) continuously in a city, in light of the GPS directions of vehicles got in current availabilities and over a time of history and also delineate sources. In spite of the fact that this is a deliberately imperativeundertaking in many rush hour gridlock observing and steering frameworks, the issue has not been all around settled yet given the accompanying three difficulties. The first is the information sparsity issue, i.e., numerous street fragments may not be gone by any GPS-prepared vehicles in present vacancy. Much of the time, we can't discover a direction precisely crossing a question way either. Second, for the piece of a way with directions, they are numerous methods for utilizing (or joining) the directions to assess the comparing travel time. Finding an ideal mix is a testing issue, subject to a tradeoff between the length of a way and the quantity of directions crossing the way (i.e., bolster). Third, we have to immediately answer clients' questions which may happen in any piece of a given city. This requires a proficient, adaptable and compelling arrangement that can empower citywide and ongoing travel time estimation.Z. Chen, Y. Liu, R. C. Wong, J. Xiong, G. Mai, and

C. Long, "Productive calculations for ideal area inquiries in street systems," Here examination the ideal area inquiry issue dependent on street systems. In particular, we have a street organize on which a few customers and servers are found. Every customer finds the server that is nearest to her for administration and her expense of getting served is equivalent to the (organize) separate between the customer and the server serving her increased by her weight or significance. The ideal area question issue is to discover an area for setting up another server to such an extent that the most extreme expense of customers being served by the servers (counting the new server) is limited. This issue has been considered previously, yet the best in class is as yet not sufficiently productive. In this paper, we propose a productive calculation for the ideal area question issue, which depends on an original thought of\emph{nearest area component}. We likewise examine three augmentations of the ideal area question issue, to be specific the ideal numerous area inquiry issue, the ideal area inquiry issue on 3D street systems, and the ideal area question issue with another target.

S. Li and O. Svensson, "Approximating k-middle by means of pseudo approximation,Here present a novel estimation calculation for k-middle that accomplishes a guess certification of 1+√3+ε, enhancing the decade-old proportion of 3+ε. Our methodology depends on two segments, every one of which, we accept, is of autonomous intrigue. In the first place, we demonstrate that with the end goal to give a α- guess calculation for k-middle, it is adequate to give

a pseudo-estimate calculation that finds a α-inexact arrangement by opening k+O(1) offices. This is a fairly astounding outcome as there exist examples for which opening k+1 offices may prompt a noteworthy littler expense than if just k offices were opened.



III. **Problem Statement**

Source retrieval is a core task of plagiarism detection. The source recovery errand can be portrayed as: given a suspicious record and a web internet searcher, the undertaking is to recover the source archives from which content has been reused. The examination of written falsification source recovery calculation is a significant work which is something other than for the advancement of copyright infringement programming. Discovering copyright infringement sources from a huge number of site pages is a testing work for all of specialists. Skillet sorted out Source Retrieval Evaluation from 2012. Potthast et al. condensed the general procedure by investigating the algorithms committed by contestants.

IV. Source Retrieval in Plagiarism Detection Source retrieval is a center errand of copyright infringement recognition. The source recovery errand canbe portrayed as: given a suspicious report and aweb internet searcher, the assignment is toretrieve the source archives from which content has been reused [1]. The examination ofplagiarism source recovery calculation is a significant work which is something other than forthe advancement of written falsification programming. Discovering unoriginality sources from tens ofmillions of site pages is a testing work for all of scientists. Container composed Source Retrieval Evaluation from 2012. Potthast et al. summarizedthe general process by dissecting the calculations.Pursued the above procedure, we center around download sifting process in this year'sevaluation. For the procedure from piecing to look control, we go for high recall,and for the procedure of the download separating, we dedicate to enhance precision.Given a settled suspicious content lumping strategy and a settled downloading number ofretrieval results, we find there is no remarkable contrast on assessment measurerecall on the off chance that we hold enough recovery results (for instance, 100 recovery results for aquery) without thinking about exactness. In this way, we choose to accomplish a high review by submitting however many questions as could beexpected under the circumstances to the web search tool and holding as manyretrieval results as would be prudent.

Fig1: A general process of plagiarism source retrieval

Chunking

Firstly, the suspicious writings are divided into sections that are madeup of just a single sentence. Particularly, it is discovered that the suspicious documentsgenerally contain a few headings. On the off chance that there are unfilled lines in front and one behind andthe word number of the line is under 10, the current line are saw asheadings. We endeavor to utilize just headings as inquiries to recover the literary theft sourceswhen we didn't recover any sources on some suspicious archives, however the sourcesare still not found by utilizing these headings. So the headings are converged into thesentence which was adjoining them.

Key phrase Extracting

After getting all sentences, each word in each paragraph istagged using the Stanford POS Tagger[2] and only nouns and verbs are considered asquery keyphrase.

Query Formulation

Inquiries are developed by removing each sentence of kkeywords, where k = 10. On the off chance that the quantity of things and action words in a single sentence is morethan 10, we hold just best 10 with high term frequencies. Also, if the number is lessthan 10, all things and action words are viewed as the inquiry. At that point these questions aresubmitted to ChatNoir look engine[3] to recover literary theft sources.

Search Control

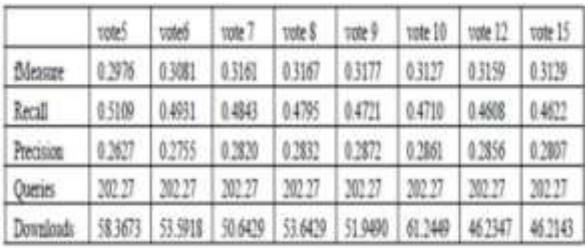
Since each query is produced by just a single sentence, it representsthe point which the sentence endeavors to express, and possibly strayed from the subjectwhich the written falsification section which the sentence originate from. The outcome is that manypositive unoriginality sources are positioned beneath. Along these lines, for each inquiry, we keep thetop 100 outcomes. This strategy makes us claim a higher review before download sifting.Download Filtering

There can be no contention that the quantity of recovery resultshas an extensive impact on the execution, and expanding the number will prompt anincrease in review and a reduction in exactness. In the means of watchwords extraction,except for the substance of suspicious archive and its content lump, we have extremely littleinformation. Submitting more inquiries might be the best decision without considering theretrieval cost. However, in the wake of recovering, we can get copious data including varioussimilarity scores among question and record, the length of report, the length ofwords, sentences and characters of archive, the snippet(the length of scrap werequested is 500 characters, etc. By abusing the recovery results and themeta-information returned by ChatNoir API, we structure a two-advance download filteringalgorithm.

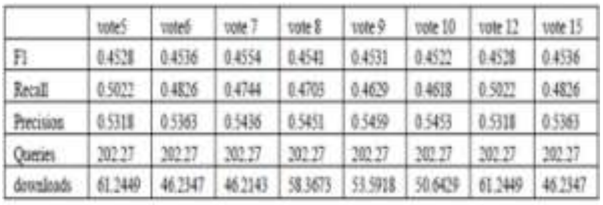
As we known, the assessment calculation of source recovery figures recall,precision and fMeasure by utilizing the downloading records, so before implementingour download sifting calculation, we choose to channel some recovery results right off the bat. Wesuppose that the inquiries can recover a similar written falsification sources in the event that they come fromthe same copyright infringement portion of suspicious archive. At that point, for one suspiciousdocument, a similar recovery results will happen ordinarily. The underlyingassumption is that more conceivable literary theft sources are probably going to get more searchresults casting a ballot from various questions of suspicious record. Thus, we utilize a straightforward votealgorithm to allot a weight to each archive of the recovery results set. On the off chance that adocument is recovered by an inquiry, the heaviness of the report will include 1. We havealso attempted the weighted vote approach by giving the record which positioning at thefront more higher weight, yet it don't perform superior to the straightforward vote approach.After actualizing vote calculation, the aftereffects of vote are viewed as the candidateplagiarism sources. In the event that the span of result list is under 20, we pick the best

50results as indicated by the best casting a ballot results as the hopefuls.

Table 1 demonstrates the execution of source recovery just utilizing vote approach tofilter the recovery results, which is called Han15 by PAN in [4]. Analyses wereperformed on the train dataset pan14-source-recovery preparing corpus-2014-12-01 ofsource recovery which contains 98 suspicious records. The numbers in the columnheaders implies the check of vote, and the line headers are the assessment measures ofsource recovery. We pick vote 8 when we present our source recovery programming toPAN.



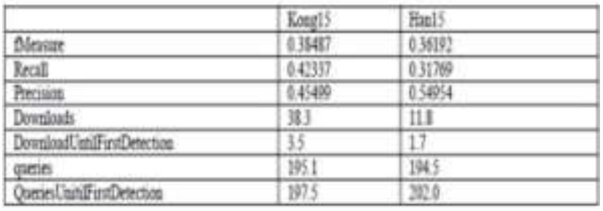
	vote5	vote6	vote 7	vote 8	vote 9	vote 10	vote 12	vote 15
fMeasure	0.2976	0.3081	0.3161	0.3167	0.3177	0.3127	0.3159	0.3129
Recall	0.5109	0.4931	0.4843	0.4795	0.4721	0.4710	0.4608	0.4622
Precision	0.2627	0.2755	0.2820	0.2832	0.2872	0.2861	0.2856	0.2807
Queries	202.27	202.27	202.27	202.27	202.27	202.27	202.27	202.27
Downloads	58.3673	53.5918	50.6429	53.6429	51.9490	61.2449	46.2347	46.2143



	vote5	vote6	vote 7	vote 8	vote 9	vote 10	vote 12	vote 15
F1	0.4528	0.4536	0.4554	0.4541	0.4531	0.4522	0.4528	0.4536
Recall	0.5022	0.4826	0.4744	0.4703	0.4629	0.4618	0.5022	0.4826
Precision	0.5318	0.5363	0.5436	0.5451	0.5459	0.5453	0.5318	0.5363
Queries	202.27	202.27	202.27	202.27	202.27	202.27	202.27	202.27
downloads	61.2449	46.2347	46.2143	58.3673	53.5918	50.6429	61.2449	46.2347

Table 2. Results of combining vote and classification approach

Our two evaluation results reported by PAN are shown in Table 3.



	Kong15	Han15
fMeasure	0.38487	0.36192
Recall	0.42337	0.31769
Precision	0.45499	0.54954
Downloads	38.3	11.8
DownloadUntilFirstDetection	3.5	1.7
queries	195.1	194.5
QueriesUntilFirstDetection	197.5	202.0

Table 3. Results of PAN@CLEF2015 Source Retrieval subtask

Table1:Results of only using vote approach

The data in above table 1 is assessed by our own assessment finder which isdesigned as indicated by Ref. [1]. Be that as it may, we just executed the previous two-wayapproach to decide genuine positive identifications since we didn't know whichalgorithm was utilized to extricate literary theft sections' set which were connected to computethe control relationship.

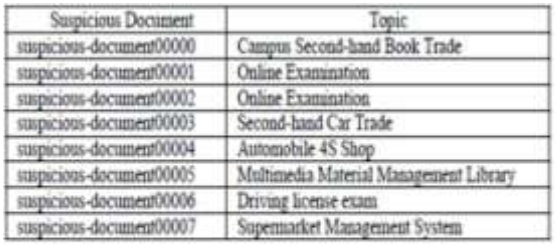
In the previous year's assessment, Williams et al.[5] proposed a sifting approach whichviewed the separating procedure of applicant literary theft sources as a classificationproblem. A directed learning strategy dependent on LDA(Linear Discriminant Analysis)was used to take in an order model to choose which hopeful counterfeiting sourcewas the positive location before downloading them. This year, we pursued theiridea and included four new highlights. They are Document-piece word 2-gram, 3-gram,4-gram and 8gram crossing point. The arrangement of word 2, 3, 4 and 8 grams from thesuspicious archive and piece are removed independently, and the normal n-gramsare figured. We picked SVM as our arrangement show. The open instruments SVMlight(http://www.cs.cornell.edu/People/tj/svm_li ght/) is utilized as our classifier. Weonly prepared theparameter c in preparing set which was built by Ref.[6]. In the wake of casting a ballot, every one of the outcomes which are certain case made a decision by classifier aredownloaded. The vote methodology pursues Han15. This methodology dependent on vote andclassification is called Kong15 by PAN in [4].

Utilizing the Source Oracle, we sifted our outcomes. The last log document announced thefiltered aftereffects of source recovery. Table 2 demonstrates the outcomes by utilizing the classificationtactics.

V. **Text Alignment Corpus Construction**

For the task of text alignment corpus development, we present a corpus whichcontains 7 literary theft cases. The written falsification cases are developed by utilizing realplagiarism.

Initially, we selected 10 volunteers to compose a paper as indicated by a point weproposed. We pick 7 of 10 to present our corpus. Table 4 records the 7 topic.For each exposition, we ask for ten thousand Chinese characters at any rate. The volunteersretrieved the related substance regarding the matter by utilizing the predetermined web index andwrote the paper. Particularly, the Baidu is utilized to internet searcher. The number ofsources has not been not constrained.At that point papers were submitted to a renowned Chinese written falsification recognition softwarewhich are utilized in numerous Chinese schools and colleges. This literary theft detectionsoftware utilizes the unique mark innovationto recognize the written falsification. Next, the volunteersmodified the substance which were distinguished by this product. The alteration tacticsinclude: changing the words' structure, supplanting the words and summarizing adjustment. Be that as it may, regardless of what sorts of adjusting strategies they received, they mustensure that the paper subsequent to changing is coherent and reliable with the first paper'smeaning. Finally, the altered papers were submitted to the counterfeiting detectionsoftware until the point that the product could never again recognize any unoriginality. The modifiedpapers were submitted to PAN as the content arrangement corpus.



Suspicious Document	Topic
suspicious-document00000	Campus Second-hand Book Trade
suspicious-document00001	Online Examination
suspicious-document00002	Online Examination
suspicious-document00003	Second-hand Car Trade
suspicious-document00004	Automobile 4S Shop
suspicious-document00005	Multimedia Material Management Library
suspicious-document00006	Driving license exam
suspicious-document00007	Supermarket Management System

Table 4. Topics of text alignment corpus construction

VI. **Conclusion**

In the past year's evaluation, Williams proposed a sifting approach which saw the separating procedure of competitor copyright infringement sources as an order issue. A directed learning technique dependent on LDA(Linear Discriminant Analysis) was utilized to take in a grouping model to choose which applicant literary theft source was the positive discoveries previously downloading them. This year, we pursued their thought and included four new highlights. They are Document-bit word 2-gram, 3- gram, 4-gram and 8gram crossing point. The arrangement of word 2, 3, 4 and 8 grams from the suspicious report and piece are removed independently, and the normal n-grams are processed.

VII. **References**

[1] L. A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung,
C. Hung, and W. Peng, "On discovery of traveling companions from streaming trajectories," in IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012, 2012, pp. 186–197.

[2] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang, "On discovery of gathering patterns from trajectories," in ICDE, 2013, pp. 242–253.

[3] A. Kharrat, I. S. Popa, K. Zeitouni, and S. Faïz, "Clustering algorithm for network constraint trajectories," in Headway in Spatial Data Handling, 13th International Symposium on Spatial Data Handling, Montpellier, France, 23-25 July 2008, 2008, pp. 631–647.

[4] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. E. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. de Lara, "Mobility detection using everyday GSM traces," in UbiComp 2006: Ubiquitous Computing, 8th International Conference, UbiComp 2006, Orange County, CA, USA, September 17-21, 2006, 2006, pp. 212–224.

[5] J. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-anddetect framework," in Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, México, 2008, pp. 140–149.

[6] Y. Zheng, X. Xie, and W. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," IEEE Data Eng. Bull., vol. 33, no. 2, pp. 32–39, 2010.

[7] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 220–232, 2013.[Online]. Available: http://dx.doi.org/10.1109/TKDE.2011.200

[8] J. Bao, Y. Zheng, D. Wilkie, and M. F. Mokbel, "A survey on recommendations in location-based social networks," ACM Transaction on Intelligent Systems and Technology, 2013.

[9] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu, "SmartAdp: Visual analytics of large- scale taxi trajectories for selecting billboard locations," IEEE Transactions on Visualization and Computer Graphics (Proceedings of IEEE VAST 2016, vol. 20, no. 12, 2014.

[10] F. Chierichetti, R. Kumar, and A. Tomkins, "Max-cover in map-reduce," in WWW. ACM, 2010, pp. 231–240.

[11] D. S. Hochba, "Approximation algorithms for np-hard problems," SIGACT News, vol. 28, no. 2, pp. 40–52, 1997.

[12] N. Alon, B. Awerbuch, Y. Azar, N. Buchbinder, and J. Naor, "The online set cover problem," in STOC, 2003, pp. 100–105.

[13] G. Cormode, H. J. Karloff, and A. Wirth, "Set cover algorithms for very large datasets," in CIKM, 2010, pp. 479–488.

[14] B. Saha and L. Getoor, "On maximum coverage in the streaming model & application to multi-topic blog-watch," in SDM, 2009, pp. 697–708.

[15] "Smartadp," smartadp.chinacloudapp.cn.

[16] D. S. Hochbaum, "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems," in Approximation algorithms for NP-hard problems. PWS Publishing Co., 1996, pp. 94–143.

[17] Y. Li, J. Bao, Y. Li, Y. Wu, Z. Gong, and Y. Zheng, "Mining the most influential k-location set from massive trajectories," in SIGSPATIAL. ACM, 2016.

[18] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Mapmatching for low-sampling-rate gps trajectories," in SIGSPATIAL. ACM, 2009, pp. 352–361.

[19] T. Sellis, N. Roussopoulos, and C. Faloutsos, "The r+-tree: A dynamic index for multi-dimensional objects," 1987.

[20] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Comput. Surv., vol. 31, no. 3, pp. 264–323, 1999.

[21] J. L. Bentley, "Multidimensional binary search trees used for associative searching," Communications of the ACM, vol. 18, no. 9, pp. 509–517, 1975.

[22] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve," TKDE, vol. 13, no. 1, pp. 124–141, 2001.

[23] U. Feige, "A threshold of ln n for approximating set cover," Journal of the ACM (JACM), vol. 45, no. 4, pp. 634–652, 1998.

[24] J. E. Smith and J. R. Goodman, "Instruction cache replacement policies and organizations," Trans. Computers, vol. 34, no. 3, pp. 234–241, 1985.

[25] Y. Li, Y. Zheng, S. Ji, W. Wang, L. H. U, and Z. Gong, "Location selection for ambulance stations: a data-driven approach," in SIGSPATIAL. ACM, 2015, pp. 85:1–85:4.