International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Fine Grained Access Control on Information Retrieval with Collaborative Fusion and Active Feedback

Dinesha L, Kumaraswamy S

Abstract: ***Enterprise information retrieval is quite challenging than web based information retrieval due to retrieval from heterogeneous distributed data sources and need for higher accuracy in retrieval [17]. In our earlier work, collaborative fusion guided by active feedback is proposed for information retrieval in enterprise environment. Applying three dimensions of user similarity, user-document search history and document similarity, collaborative information fusion based retrieval was proposed in that work However the work did not consider enforcing fine grained access control on the retrieval which is very important requirement in enterprise environment. In this work, fine grained access controlled information retrieval on enterprise environment, with zero leakage assurance thorough direct or inference is proposed. Fine grained access control is ensured with help of CP-ABE. Concept masking is done with CP-ABE and unmasked when user satisfies the query access right criteria.***

Keywords : CP-ABE, CR,QC, PS

I. INTRODUCTION

Enterprise information retrieval system retrieves relevant information to users from huge repository of information generated by the enterprises across various departments. An important requirement in enterprise retrieval is that it must not compromise on the access control restrictions. In our earlier work [1], collaborative fusion based retrieval is proposed with an objective to search the information repositories for the relevant answer and then to present a search result in the ranked order of maximal utility to the searcher. In [1], an efficient information retrieval based on the concept of collaborative product recommendation is proposed for retrieval of information. Similar to product recommendation, relevant documents are retrieved, using fusion of the relationship between users, user to concept preference and the document similarity. The document similarity is modeled in terms of similarity in the number of concepts shared by the documents and the user similarity is modeled in terms of concepts preference similarity between the users. Active Feedback based relevancy tuning is done to fine tune the results and in the process the learning made about user preference in search result is remembered is using to concept preference, so that the number of trials on successive searches is reduced. But the work did not consider the important requirement of ensuring access control on the search. Most enterprises implement document management systems to implement access control for the users. But those systems are based on tight access control and retrieval process; search the portions in repository depending on user permissions. This is limiting the information access capacity of the enterprise information systems. Rather than restricting some portions in the repository, the retrieval system can provide at least a clue to where the details about his search can be found, so that the user can request access permissions to fetch those details. In this work, a fine grained access control on Collaborative fusion active feedback information retrieval is proposed. The solution is made as an extension to the collaborative fusion with active feedback architecture detailed in [1].Each multi page document is defined with a fine grained access policy for all the pages and the multiple concept indexes is constructed with concept index masked using filters. The masking of index is done in accordance with the access rules defined by the owner of the document. The owner intention to open up the document for look up alone is modeled in terms of mask tolerance factor on the concept indexes. The mask tolerance factor is auto corrected in active feedback step. The masking is controlled with CP-ABE (Cipher Policy Attribute Based Encryption).

II. RELATED WORK

In [2], a robust private information retrieval scheme is proposed. In this scheme, encrypted data is stored in cloud. Every time a user needs to retrieve the information, he needs to provide the secret key and certificates to authenticate his access on the data. This scheme does not provide. Solution proposed in [2] ensures privacy during information. It is based on conditional disclosure of necessary information on demand. A new scheme called M-SSE is proposed in [4]. It ensures both forward and backward security. This work used multiple clouds to protect against insider attack and information leakage. Search over encrypted documents with multiple encrypted keywords is proposed in [5]. Relevance score combined with k-nearest neighbor is proposed in this method. Personalized search platform with privacy and security of search ins proposed in [6].Authors in [7] addressed the problem of providing selective access control after outsourcing the data. By combining cryptography with authorizations, access control was enforced.. ENKI [8] is a information retrieval system to ensure privacy of search queries on sensitive data. By distributing the sensitive works over client, access control was enforced. Authors in [9] proposed a coarse grained access control on outsourced data.

Manuscript published on 30 September 2019.
***Correspondence Author(s)**

Dinesha L*, Research Scholar, Computer Science and Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India. Email: ldinesha.ssit@gmail.com
Kumaraswamy S, Professor, Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, India. Email: kumar.aruna@gmail.com

Retrieval Number: J97900881019/19©BEIESP
DOI: 10.35940/ijitee.J9790.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The user access credits are modified on demand based on the trust level. Trust level is calculated continuously and user is blacklisted if the trust falls below a threshold. Resource sharing is also supported in this platform. Optimization in indexes used for keyword search is proposed in [10]. First time and Successive search are differentiated and successive search is optimized. A distributed approach combining access control with information retrieval is proposed in [11]. Another advantage in this solution is that it supports multi user search. Data is split to chunks with access control privileges and security is enforced each block wise. A Privacy preserving information retrieval is proposed in [12]. The framework works on encrypted cloud. Users can perform any of the multi-keyword search, range search and k-nearest neighbour search operations in a privacy preserving manner in the proposed solution. The queries are transformed into predicate-based search leveraging bucketization, locality sensitive hashing and homomorphic encryption techniques. The proposed solution is implemented in Map reduce framework for achieving the retrieval speed up. A finger printing based solution for information retrieval is proposed in [13]. The information leak in the access controlled part of the documents is protected by finger printing and the users are denied from accessing those parts. Authors in [14] propose a credential enforcing system. The system is based on state machines. Based on state of users, the access control is made. Actions of users are controlled based on which state they are present. A role based access control strategy is proposed for enterprise search in [15]. User profile is expressed as ontology with details about access control. Search process reads the ontology and behaves according to access control restrictions expressed in ontology. A multi-user integrity management framework is proposed in [16]. The data is maintained in hierarchical manner based on data sensitiveness. A unique measure called depthness measure is proposed based on the hierarchical structure of data. As the depth increases, the sensitive of data also increases. Access control is varied based on the depthness measure. The retrieval process has more tight control at higher values of depthness measure.

III. FINE GRAINED ACCESS CONTROLLED INFORMATION RETRIEVAL

The architecture of the fine grained access controlled information retrieval system is given below “Fig. 1”.Fine grained access control is ensured with help of CP-ABE. Each enterprise user must be associated with a user profile. User profile is modeled in terms of set of attribute value mapping. For each multi page document generated by the enterprise user, a meta descriptor file about the document is created. The meta descriptor file describes the mapping between pages to the user profile. Each mapping rule is of form

<Page No Range>**, **<Profiles>**, **<Allow or Deny>

The page number can be configured separate or in form of ranges. The profile is list of profile id already defined for categorization of enterprise user. The document creator’s intention to lookup on pages is expressed in terms of allow or deny decision. The fine grained access granularity is in terms of page in the proposed solution. For each page, concept is extracted from analysis of textual content in the page. Concept is modeled as k-gram word and extracted using the procedure defined in [1].

Concept is a k-gram representing real or imaginary entity that users may be interested in and does not have any extraneous words such that excluding them would identify the same entity. The concept satisfies the two properties of popularity and conciseness. Following four indicators are used to decide if a k-gram is a concept.

Frequency	The frequency of occurrence (support value) must be high
Better than sub-super concepts	There is no k+1 or k-1 gram that is better concept than k-gram.
Containment	Contains only parts of sentences with single meaning or idea.
Consistency	A concept must have a consistent access restriction.

The concept extraction procedure starts with splitting of each sentence in the document. From each sentences, the stop words (is, the, was etc) are removed. With the rest of words, k-gram is created for each sentence. The frequency of occurrence is calculated for each of the k-gram. The frequency of occurrence of support value of a k-gram is given as

$$S_a = \sum_{for\ each\ word\ x\ in\ a} S_x - \ (1)$$

Among all the k-gram, the best subset of k-gram is selected satisfying the three indicators given above.

The best subsets are the concepts in the document. The process of concept extraction is given in “Fig 2.” The indicators for k-gram to concept selection is done as follows

First indicator is satisfied if the S_a of a k-gram is above threshold T.

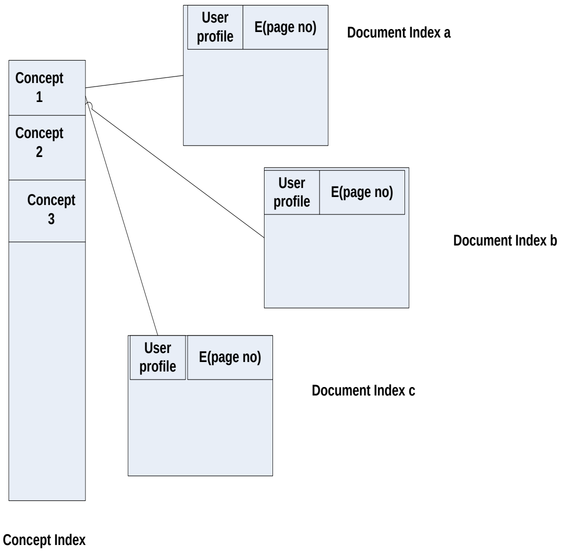
Second indicator is satisfied if there is no k-1 gram with same support or k+1 gram with support more than S_a.

Third indicator is satisfied if there is k-gram with equivalent alias and support greater than S_a. If two k-grams are synonyms or noun forms which can be replaced with one other, then it is alias.

The input page is split to sequence of sentences. From each sentence stop words (is, was, so ...etc) are removed. The sentence after stop word retrieval is added to the k-gram list. For each item in the list, support value is calculated using Equation 1. After calculation of the support value, each k-gram is evaluated for selection three indicators and the k-gram satisfying the three indicators are the concepts for the document. But different from [1], concept consistency checking is added as one of the indicator in addition to already existing three indicators. The fourth indicator is for concept consistency checking against rules in meta descriptor file and it is as follows

If deny is defined atleast once for a concept, then a new concept is created by adding or removing to the k-gram.

Two level concept index maps are created with first level mapping concept to the document index. Each document index has the user profile id and the encrypted location of pages in which the concept is present. The encryption is done with CP-ABE public key of the corresponding user profile.



The encryption of concept page location in the document index prevents the attacker from gathering or modifying information in the document index. Concept relation modeling is done in the proposed solution instead of document relation modeling in [1]. Relation between two concepts is given in terms of co-occurrence of both concepts in number of documents.

$$CR(i,j) = \frac{\sum_{for\ all\ documents} concept(i) \cap concept(j) \neq null}{|total\ number\ of\ documents|} - (2)$$

The concept of user-concept relation, user-user relation is same as [1]. The Query engine proposed in [1] is modified to accommodate fine grained access control as below.

The concept relation, user relation and user relation to concept matrix are factorized using matrix factorization. When user issues a query, the query is translated to concept mapping as below.

The query to concept mapping is done by selecting the k-gram in the concept list matching to query provided by the user. The query to concept score (QC) is evaluated for each concept in

entire candidate list and the concepts are ranked based on the score. Before evaluation, the entire alias for the query is generated.

$$QC(Q,C) = \sum_{i=1}^{N} f(Q_i, C) - (3)$$

Where N is the number of alias for the query Q including Q. $f(Q_i, C)$ is the function which provides value of 0 or 1. The function outputs 1 when Q_i and C are similar. The QC is calculated for each concept and the concept with the highest QC is passed to Query engine for execution. The similar concepts to the selected are chosen from the concept relation matrix to get the full list of concepts to be given as result. For each concept in the list, a preference score (PS) is calculated considering the user similarity and user preference to concept. The concepts are sorted in decreasing order of preference score and provided as search result to the user. Preference score is calculates using weighted summation of user similarity and concept preference as given below

$$PS = W1 * UR(a,b) + W2 * UD(m,b) - (4)$$

The weight values W1 and W2 are configured by the administrator of the retrieval system.

Top K concepts from the preference score, sorted concept list is taken and a lookup is done on the Concept index to get the matching document index. The searching user profile attributes is retrieved from the user profile database and CP-ABE private key is generated. On each matching document index containing the searching user profile id in the rows, the decryption is done using CP-ABE private key to get the matching pages in the documents. All the matching pages in the documents is given as the search result to the user. The query execution process is detailed in form of flowchart in “Fig. 3”.

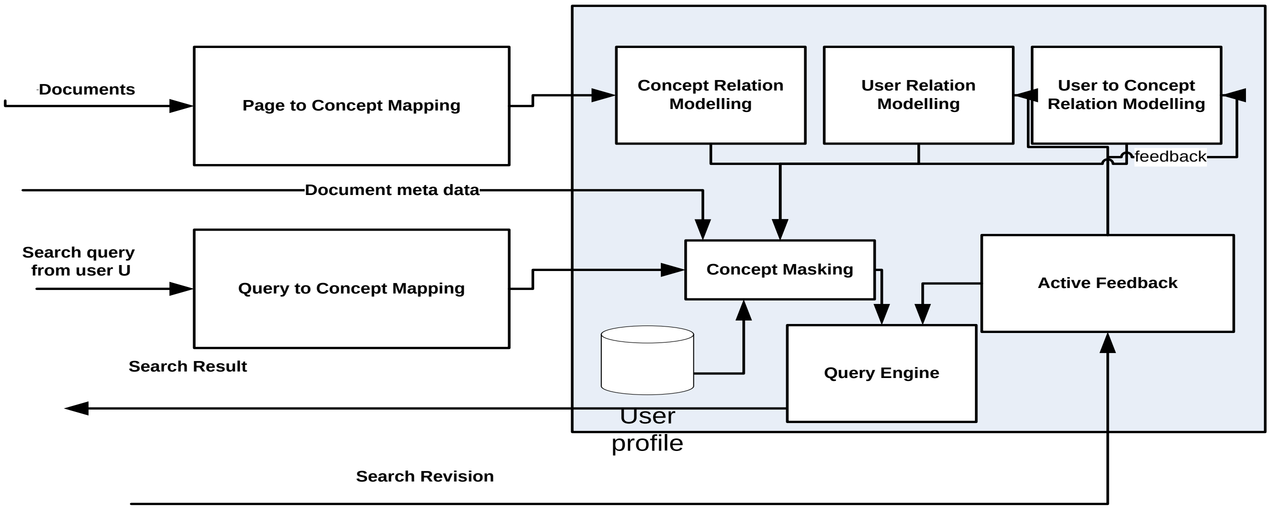


Fig.1: Fine grained access controlled information retrieval

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

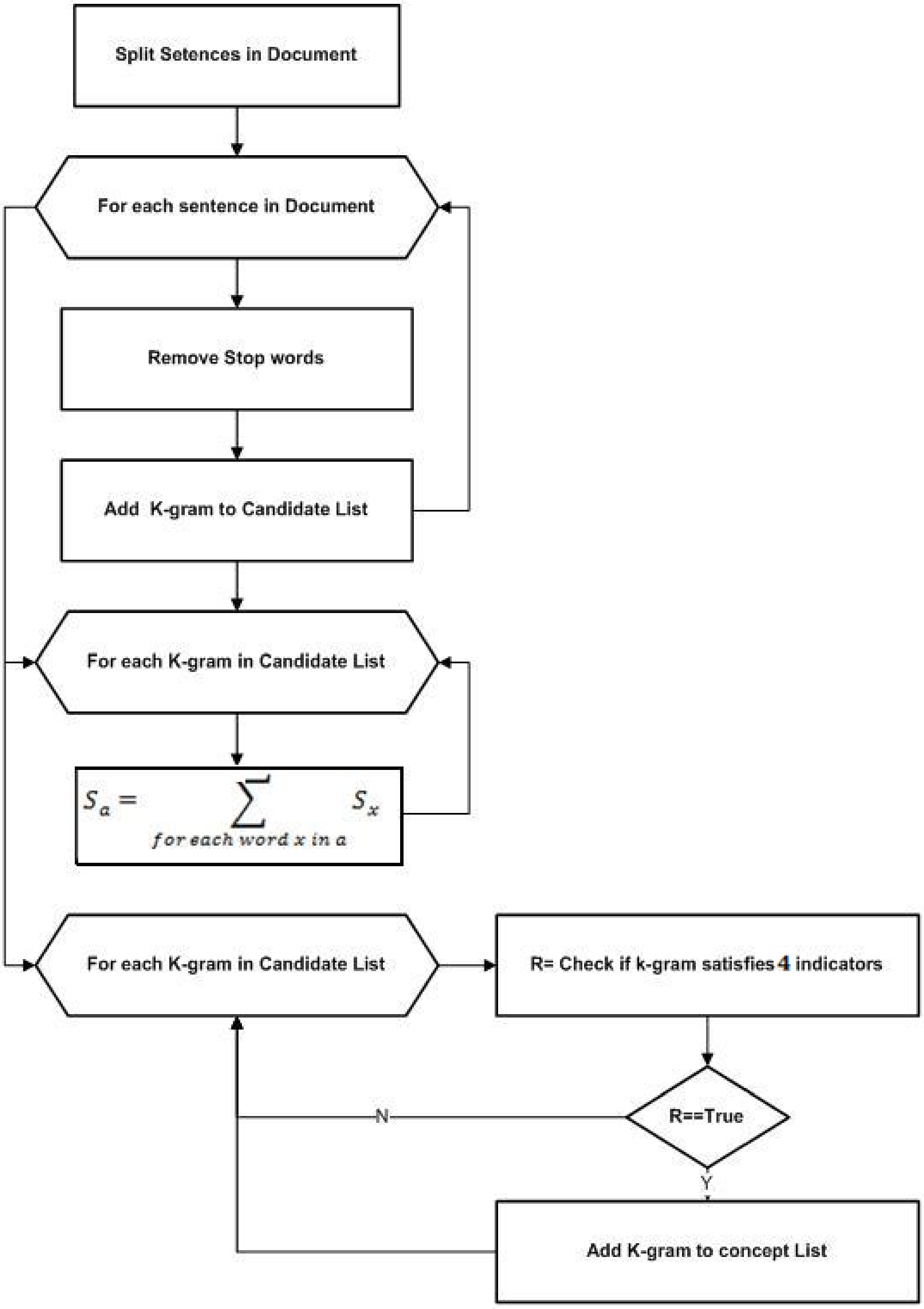

Split Setences in Document
For each sentence in Document
Remove Stop words
Add K-gram to Candidate List
For each K-gram in Candidate List
$S_a = \sum_{for\ each\ word\ x\ in\ a} S_x$
For each K-gram in Candidate List
R= Check if k-gram satisfies 4 indicators
R==True
N
Y
Add K-gram to concept List

Fig. 2:Concept Extraction

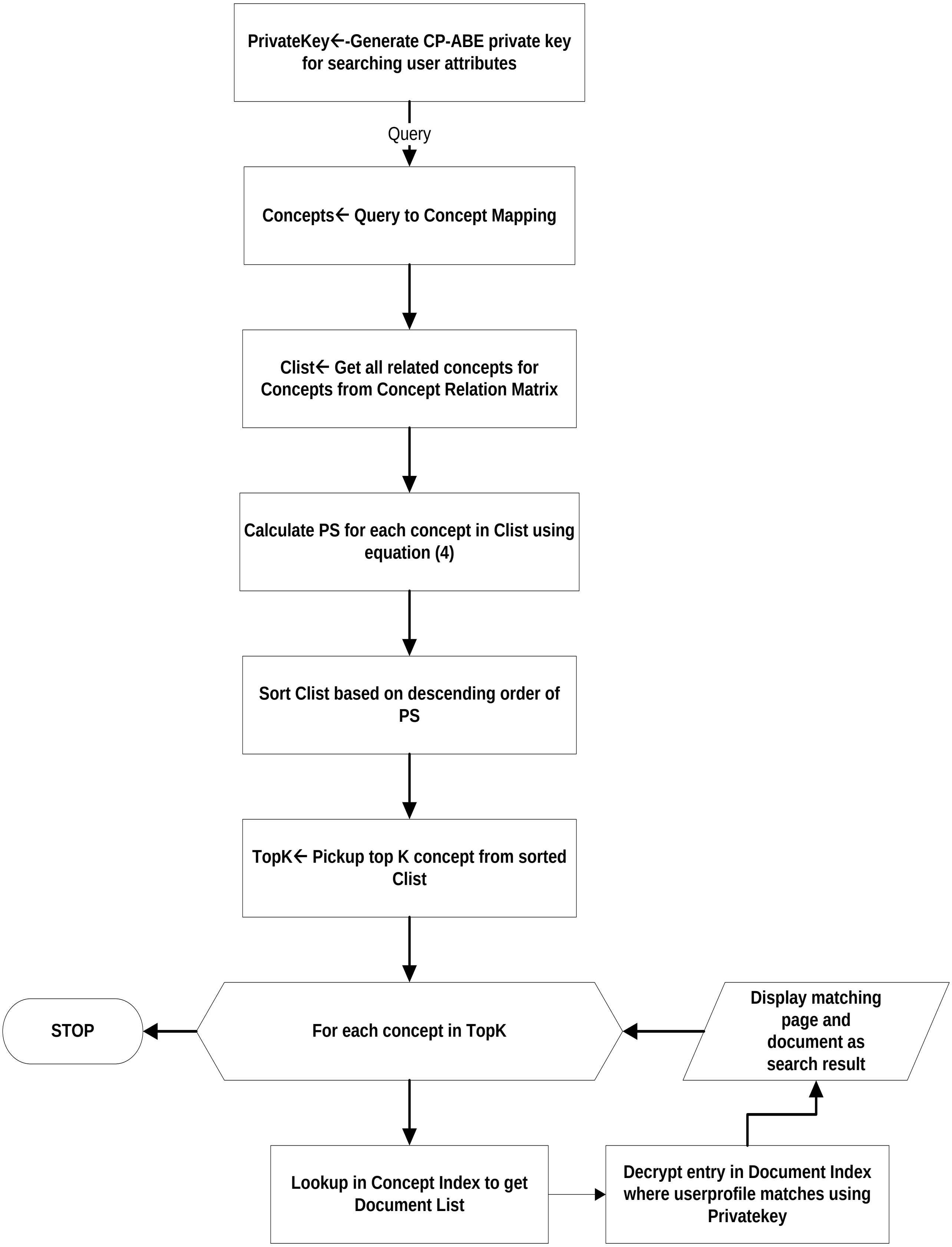

PrivateKey←-Generate CP-ABE private key for searching user attributes
Query
Concepts← Query to Concept Mapping
Clist← Get all related concepts for Concepts from Concept Relation Matrix
Calculate PS for each concept in Clist using equation (4)
Sort Clist based on descending order of PS
TopK← Pickup top K concept from sorted Clist
STOP
For each concept in TopK
Display matching page and document as search result
Lookup in Concept Index to get Document List
Decrypt entry in Document Index where userprofile matches using Privatekey

Fig. 3: Query Execution

IV. RESULTS

The proposed system for information retrieval was compared against [1]. Dataset used for evaluation is TREC Enterprise Track 2007[19]. TREC enterprise data is designed to assist experiments on enterprise data and make conclusions. TREC data closely resembled real world enterprise data. The dataset have 370715 documents with different types. A set of 50 queries was designed and the expected results and their ranking order is established before the start of the test. The performance is measured in terms of

1. R^2 measurement
2. Hold out Estimated accuracy
3. Precision
4. Recall
5. Ranking Analysis
6. Query Execution time

R^2 measurement is a statistical measure for accuracy. It is calculated as

$$R^2 = \frac{\sum_{i=1}^{n}(\check{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} = \frac{\text{SS Predicted}}{\text{SS Total}} - (5)\,[3]$$

Where n is the number of data instances, $\bar{y}$ is mean of actual values of the instances, $\check{y}$ is the predicted value of the data instance i.

Hold out Estimated accuracy is calculated as

$$acc_h = \frac{1}{h} \sum_{vi,yi \in Dh} \sigma(vi, yi) - (6)\ \ [3]$$

Where D_h is the subset of data set D of size h and $\sigma(v, y) = 1$ if v=y and 0 otherwise. vi is the predicted value of instance i and yi is the actual value of instance i.

Precision measures system capability to provide relevant item. It is calculated as

$$P = \frac{|R_a|}{A} - (7)[3]$$

Where R_a is the number of items relieved and A is the total number of items retrieved in response to search query.

Recall is the ability of the system to present all relevant items. It is measured as

$$R = \frac{|R_a|}{|R_m|} - (8)[3]$$

Where R_m is the total number of relevant items in the document set.

Document ranking analysis is done to measure the effectiveness of ranking in the proposed system.

R^2 Measurement values for proposed and the [1] is given in Table-I.

Table-I: R^2Measurement

	Proposed	[1]
SS Predicted	10674	10672
SS Total	12412	12405
R^2	85.99%	85.12%

There is no big change in R^2 Measurement values due to addition of fine grained access control changes added to [1].

The Hold out Estimated accuracy for the proposed and [1] is given below “Fig. 4”. There is no big change in the Hold out accuracy due to addition of fine grained access control on the solution [1].

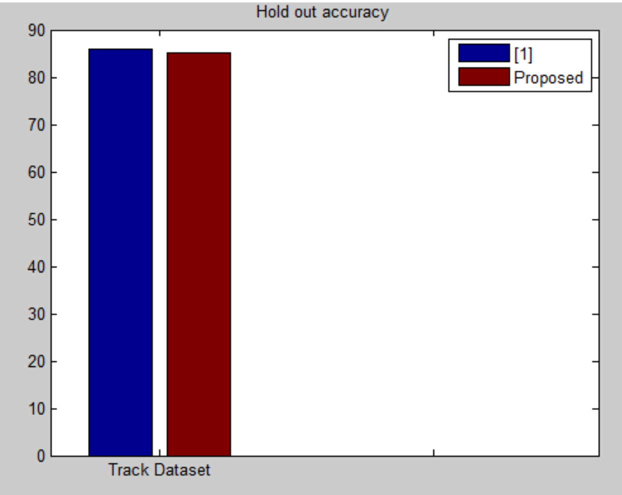


Fig.4: Hold out Accuracy

The precision and recall for the proposed and [1] is given in Table-II.

Table-II: Precision and Recall

	Proposed	[1]
Precision	0.89	0.81
Recall	0.19	0.23

The precision and recall of the proposed solution is higher than solution [1] is given in below “Fig. 5”.

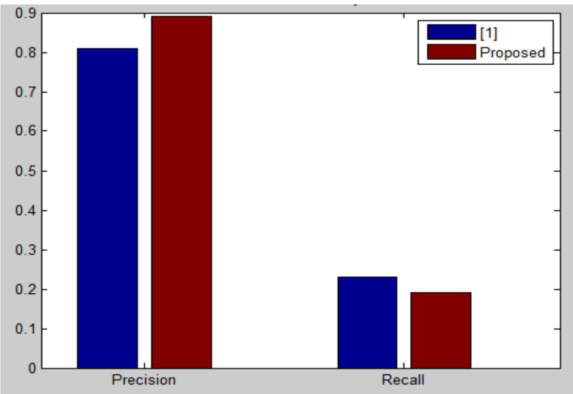


Fig.5: Precision & Recall Comparison

The precision is increased due to addition of fine grained access control changes in [1] , as search result provides precise page matching to the user query.

The retrieved documents are checked for ranking accuracy by testing against 50 queries is split into following categories in intervals of 10.

- 1-10 →A
- 11-20→B
- 21-30→C
- 31-40→D
- >40-> E

Retrieval Number: J97900881019/19©BEIESP
DOI: 10.35940/ijitee.J9790.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

- Not retrieved → F

The document frequency is calculated in each category and the result is given below Table-III for both the proposed and fuzzy solution [1].

Table-III: Relevant Analysis

	Proposed Solution	[1]
A(1-10)	77.83%	73.83%
B(11-20)	7.44%	7.44%
C(21-30)	2.13%	2.13%
D(31-40)	1.06%	1.06%
E(41)	1.06%	1.06%
F(Not retrieved)	10.02%	12.02%

Highly relevant ranked results are provided as result in the proposed solution compared to [1].

The average time taken for query execution is given below "Fig. 6".

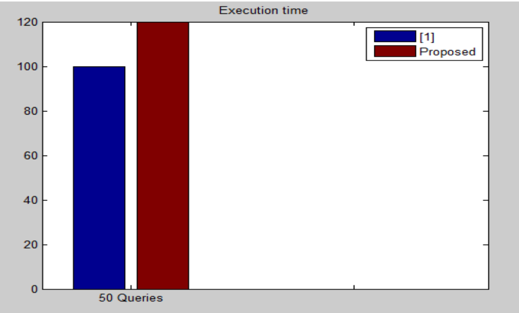


Fig.6: Average Time for query execution

Query Latency	Proposed	[1]
Execution time	101 ms	120 ms

The execution time in proposed changes has increased by 20% due to CP-ABE decryption in the document index.

V. CONCLUSION

Fine grained access control based Collaborative fusion information retrieval with active feedback is proposed in this work. The approach used three dimensions of user relations, user to concept relations and concept to concept relations to find highly relevant results to the users. Page is the fine grained access control in the proposed solution. Concept masking using CP-ABE is done to ensure fine grained access control. The precision of search results has increased due fine grained search results done on [1]. The increment in query execution time is only in milliseconds, so there is no big impact on user search experience in the proposed solution. Overall the proposed solution satisfies the enterprise requirements for access controlled information retrieval. Thorough testing on real world dataset, the proposed solution is found to give precise results for the search query with low latency. Exploring other efficient means of masking the concepts for access control will be considered for future work.

REFERENCES

1. Dinesha L, and Kumaraswamy S "Enterprise Information Retrieval using Collaborative Fusion and Active Feedback",IEEE Explore Digital Library, 2019
2. Khaled Riad,Lishan Ke "Secure Storage and Retrieval of IoT Data Based on Private Information Retrieval", Wireless Communications and Mobile Computing,2018
3. Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," Journal of Computer and System Sciences, vol. 60, no. 3, pp. 592–629, 2000.
4. C. Gao, S. Lv, Y. Wei, Z. Wang, Z. Liu, and X. Cheng, "M-SSE: an efective searchable symmetric encryption with enhanced security for mobile devices," IEEE Access, vol. 6, pp. 38860– 38869, 2018.
5. H. Li, D. Liu, Y. Dai, T. H. Luan, and X. S. Shen, "Enabling effcient multi-keyword ranked search over encrypted mobile cloud data through blind storage," IEEE Transactions on Emerging Topics in Computing, vol. 3, no. 1, pp. 127–139, 2015.
6. H. Li, D. Liu, Y. Dai, T. H. Luan, and S. Yu, "Personalized search over encrypted data with effcient and secure updates in mobile clouds," IEEE Transactions on Emerging Topics in Computing, vol. 6, no. 1, pp. 97–109, 2018
7. S. D. C. Di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati, "Encryption policies for regulating access to outsourced data," ACM Transactions on Database Systems (TODS), vol. 35, no. 2, 2010.
8. I. Hang, F. Kerschbaum, and E. Damiani, "ENKI: Access control for encrypted query processing," in Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '15), pp. 183–196, New York, NY, USA, June 2015.
9. K. Riad, "Blacklisting and forgiving coarse-grained access control for cloud computing," International Journal of Security and Its Applications, vol. 10, no. 11, pp. 187–200, 2016
10. W. Wang, P. Xu, H. Li, and L. T. Yang, "Secure Hybrid Indexed Search for High Efficiency over Keyword Searchable Ciphertexts," Future Generation Computer Systems, 2014.
11. Aameek Singh,Mudhakar Srivatsa, Ling Liu, "Efficient and Secure Search of Enterprise File Systems", IEEE International Conference on Web Services 2017
12. Cengiz Örencik, Erkay Savas, Mahmoud Alewiwi "A Unified Framework for Secure Search Over Encrypted Cloud Data", IACR Cryptology ePrint Archive 2017
13. Eleni Gessiou , Quang Hieu Vu , "IRILD: an Information Retrieval based method for Information Leak Detection" tech report 2011.
14. Scott E. Coull,Matthew D. Green , Susan Hohenberger "Access Controls for Oblivious and Anonymous Systems", IACR 2008
15. Lixin Zhou "Multi-agent Based Distributed Secure Information Retrieval",2010 International Conference on Communications and Mobile Computing
16. G. Thangaraju and X. Agnise Kala Rani , "Multi User Profile Orient Access Control based Integrity Management for Security Management in Data Warehouse", Indian Journal of Science and Technology, Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/90927, June 2016.
17. Mirza, Hamid. (2008). Enterprise Information Retrieval: A Survey.. ICEIS 2008 - Proceedings of the 10th International Conference on Enterprise Information Systems.
18. K. Saravanan,A. Radhakrishnan"Dynamic Search Engine Platform for Cloud Service Level Agreements Using Semantic Annotation" ,International Journal on Semantic Web & Information Systems Volume 14 Issue 3 July 2018
19. https://dblp.org/db/conf/trec/trec2007

AUTHORS PROFILE

Mr. Dinesha L received his Master Degree in Information Technology. He is a Research Scholar in Department of Computer Science and Engineering at Sri Siddhartha Academy of Higher Education, Tumakuru, India. He has published papers in international conference and journals. His area of interest is Information Retrieval, Big Data and Data Mining.

KUMARASWAMY S, is currently working as an professor in the department of computer science and engineering, sri siddhartha institute of technology, tumakuru, india. He obtained his bachelor of engineering from siddaganga institute of technology, bangalore university, tumakuru. He received his master degree and ph.d from bangalore university, bangalore. He has 19 years of teaching experience. He published more than 12 papers.hisresearch interestis in the area of data mining, web mining, semantic web and cloud computing.