# Security Issues and Accuracy Concerns in the Information Retrieval Process

Ashish Gautam
Indian Institute of Information Technology Allahabad (India)

Ketan Jain
Indian Institute of Information Technology Allahabad (India)

Vijay Kr. Pushkar
Indian Institute of Information Technology Allahabad (India)

Shashi Kant Rai
Indian Institute of Information Technology Allahabad (India)

## ABSTRACT
An approach towards Information retrieval process security issues & accuracy of the retrieved data from the user perspective, are the major concerns of this paper. Implanted work includes demonstration of the retrieval modes, process flow of information retrieval system and risk analysis of the retrieved data. Much work has been done to highlight the accuracy problem in the retrieved content where an accuracy formula focuses on the calculation of the accuracy percentage in order to generate an transparent approach towards the information retrieval process.

## Keywords
Retrieval, Risk, Severity, Authentication

## 1. INTRODUCTION
Odyssey of the information has been so long to give birth new ideas and research work. Over the past decades, history has introduced us many patterns and structure of information which has always been prominent and worth a valid concern. Being a legitimate issue, it does not only gives scope of inventions & makes applaudable proceedings but cement a realm where all humans enhance themselves in their daily working curriculum. Be it textually available or digitally accessed, every form of it has its own uniqueness.

From today's working professional to village based farmer, every mind is making use of information [1]. The meaning of information varies in different contexts. Moreover, the concept of information is closely related to data, forms, instruction, knowledge, meaning, understanding, pattern, protocols, representation and every genre. Each day of Technology makes information more compatible to use for different purposes, which one in other way gives a platform for innovations. Inception of ideas, knowledge, and information build techniques to save time. Science has witnessed many ideas which would have never been possible without information. It could be communicated to us through various modes like multimedia retrieval, documented retrieval and verbal retrieval [2,3].

Contextual concerns regarding security issues in different retrieval modes take into consideration, moreover, the ideas related to retrieval system security breach and to enhance the already discussed system [4]. A suggested approach towards the accuracy of extracting data from the database has been described. Information retrieval system authentication issues from the sender & the receiver end to make the requested information more confidential, maintaining integrity, value evaluation of the retrieved data & the security of requested query are some of the major points that this paper would be focused on.

## 2. PROBLEM DEFINITION
Issues of security and accuracy in the information retrieval process have never been highlighted through any research work so far. Moreover, emphasis has been given effectiveness on the high effectiveness of information retrieved [5], but risk analysis of the retrieved data was never focussed. The above problems have been taken into consideration.

## 3. LITERATURE REVIEW
Few research work based on the already discussed context have recorded their existence to enlarge the scope of information retrieval system [6], which offers a security model for end-to-end data channel encryption, system authentication, logging and 100% effectiveness in retrieving data but lacks inevitable menaces of accuracy & risk exposure which may lead to the acquisition of loss in many forms from the user perspective. A proposed idea through this paper makes risk analysis & accuracy a major concern, in the outcome of the retrieval process. Exactness issues are being experienced by the searchers where they feel the problem while locating data from large database & frequent access of data through the major search engines such as MSN, Yahoo, and Google etc. [7]. Probability of attack implantation may get occurrence while extracting the queried information may result into discontinuation of service availability.

## 4. MODES OF INFORMATION RETRIEVAL
Retrieval modes are categorized into three parameters on the basis of the nature of the content being retrieved through any medium. Multimedia mode is through internet where one can access data by placing query as a search on any website. Documented mode is commonly used by the users which include hard copy of data through papers & documents. Verbal mode is the easiest & a spontaneous retrieval mode which requires any known language.

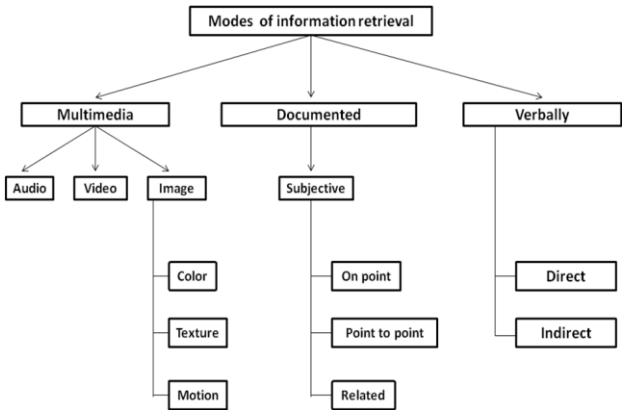


**Figure 1. Modes of Information Retrieval**

## 4.1 Multimedia

**Multimedia Information Retrieval** is a prominent mode of technological system which is based on extracting information from multimedia data sources. Data sources include media such as audio, image and video, indirect sources such as text, bio signals as well as not perceivable sources such as bio information, stock prices, etc. [6,8].


**Figure 2. Multimedia Information Retrieval**

### *4.1.1 Audio Retrieval*

Audio file retrieval systems are the retrieval systems which are based on audio clippings, audio mp3 songs kind of search and attribute classification of audio file are formatted, size, bit depth, and sample rate etc. evaluation is directly applicable to audio search. They also pose different research problems than image retrieval systems do, for two fundamental reasons: audio data is aurally-based instead of visually-based and audio data is time-dependent. The former difference leads to some unique & creative approaches in solving the querying and retrieval issue, while the latter difference is the root of the interesting problem of presentation, which image retrieval systems do not share.

### *4.1.2 Video Retrieval*

Video data retrieval contains some attributes with image data retrieval when a user googles a video file on the internet. However, video file is also time-dependent like an audio file, and in fact, movies usually have audio with the video data. This shared commonality naturally lends to applying solutions from the image and audio retrieval areas to research problems in the video retrieval domain. In some ways this strategy is successful but as usual, video data has some unique properties which again leads to creative solutions to the research issues of classification for querying and presentation.

### *4.1.3 Image Retrieval*

Image search is a data search primarily focused on images through the internet. Image data retrieval consists of images of different sizes and memories. In order to search an image, a user may type terms such as name, image link or click on any image and the system generates images with similar result. Images may be in the form of photos, snapshots wallpapers etc. The colour distribution of images, region/shape attributes can be the search criteria. Image retrieval is a technical process for searching and retrieving images from a large database.

## 4.2 Documented

It is well known that new fact are invented with time which may be in the form of sign, text, etc. [8]. The documented retrieval is in the form of text documents containing finding of text retrieval. Document retrieval is what we get from the database physically and to check whether the requested text is up to the mark or not. For an instance, a user is searching a word "python" on Google and gets details about a python animal as an outcome of the search. Here query was not related with animal python but it was python language. So it's difficult to get what was requested so sometimes result is somewhat related, on point, point to point. These all three can be measured on some point of scale categorized into related, highly related and closely related

## 4.3 Verbally

Verbally retrieval is the most common retrieval mode used by anyone where people communicate verbally in order to get information. One may relate this mode with social engineering where data leakage gets existence by human error. For instance, two people are communicating with each other and they are discussing strategies of their organization and in between the talk they suddenly discuss their project and the other person hears what they are discussing which is said to be a confidentiality breach of the information.

# 5. ACCURACY

It defines the accurate nature of information as per the user's query.

$$a\% = \left\{1 - \left(\frac{n}{t}\right)\right\} \times 100$$

**Equation (1)**

$a$ = accuracy

$n$ = non retrieved content

$t$ = total retrieved content

Conceptually, generated formula uses simple mathematics. Accuracy of the retrieved information can be calculated by this formula which consists of two parameters - non retrieved content ($n$) & total retrieved content ($t$). Out of total retrieved content, "not useful content" is subtracted which gives the probable accurate information as a result. Accuracy percentage is the final outcome of this formula from non retrieved content & total retrieved content.

The goal of defining accuracy is to have good performance in terms of accuracy. For simple understanding this can be stated as positive and negative outcome from the total outcomes which is making relation with the positive (retrieved) and non retrieved (negative) content.

**STEMMING**

It is a combination of different elements of variant from of words into a single representation. Let us take an example for better understanding, a user enters a word "computation", "calculation", "evaluate" etc. so all of these have stemmed to calculate. Stemming is majorly used in information retrieval to increase the accuracy of the retrieval system. It helps us in two ways :

1. How accurately stem information is mapped with the original information.

2. What percentage of accuracy it has created in our information retrieval system.

## 6. MODELS OF INFORMATION RETRIEVAL

The model is used to define the nature or processing of one system. One can easily think to develop a model related to the Solar system position as it can describe the position, date, time etc. of the system or one can also think a model to check what happens in the daily climate change of the environment. Linear scanning is easily done when the collection of data is small from which one can easily search our data. The inverted file method is also much same in it we search it from the back of the book that list the entire thing alphabetically. A simple approach is to find the search document from a large collection of databases. According to Mish et al. (1983) Model a pattern of something to be made.

**Model**: It is a proposed idea which user can present in future to implement in the system.

**Boolean Model**: It is the first model of information retrieval which uses the Boolean's Logic i.e. AND which is logical multiplication, OR which is logical addition, NOT which is logical complement.

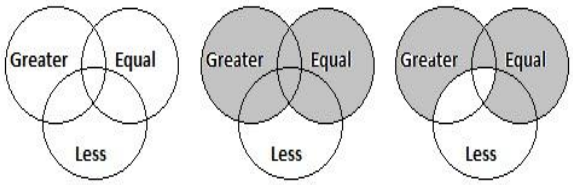


**Figure 3. Boolean model**

**Advantages**: The system is being under control by the user with the help of a model. From this query output can be known to us either it is in bigger set or smaller set. Secondly, the model can be extended with proximity operators and wildcard operators in a mathematically sound way, which makes it a powerful candidate for full text retrieval systems as well.

**Disadvantages**: No ranking has been provided by this model. Another disadvantage is, a stiff difference exist between the Boolean operator AND and OR and the natural language used 'and' and 'or'.

## 7. PROCESS FLOW DIAGRAM OF INFORMATION RETRIEVAL

The process flow defines how the information flow takes place when a user sends any query to the database. In the below proposed model, user gets actual information as per the need where the query is validated at the initial phase which also includes an error message indicating similar results. Availability of the queried content makes its finalization from the database which is preceded by accuracy phase where the content is categorized on the basis of three parameters - accurate, relevant, and related.

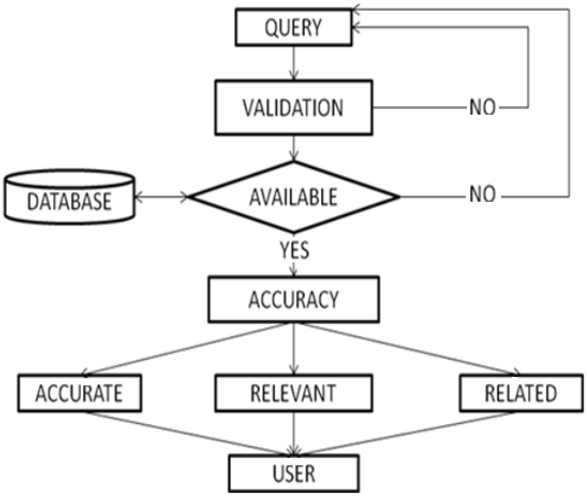


**Figure 4. Flow Process of Information Retrieval**

## 8. SECURITY ISSUES IN INFORMATION RETRIEVAL

Security of information majorly design to protect the three parameters of the C.I.A i.e. Confidentiality, integrity and availability. Now day's huge amount of work is being done to protect the information security of the organization. It's a privilege based procedure to protect the company assets, resources from unauthorized disclosure. Number of fraud is executed in the banking industry & other sectors of public interference. In the recent times, banking industry is not able to stop the fraud before it happens. According to Guardian Analytics, Banking industry is not able to probe the 78% of the online fraud where the attacker intentionally retrieves the information online like a/c number, name, pin etc.. and easily makes misuse of it for his/her own purpose. The organizations generally do not posses the required tools to protect the attacks which are implanted with negative intention.

Organizations are installing firewall, anti viruses & UTM, devices. Majorly they are following the proactive approach in

order to protect from malicious activity with a backup mechanism.

## 8.1 Secrecy

It's a practise to hide queried information (retrieved) from unauthorized disclosure. At the user end, Information must be encapsulated in order to prevent from any malicious activity. In other words, breach of information privacy must not exist. To maintain secrecy, organizations develop new & advanced covered medium as well as design and develop a robust algorithm.

## 8.2 Exactness

Alteration of data while retrieval of information must not exist. Let's take one example where user makes a query for image sized 100 kb, 1024×768 but he is not able to access the same then this scenario declares the conformity of security breach. Above discussed example is related with image security, where the focus would be given by following aspects:

### *8.2.1 Low Quality Input*

For gaining the access, intruder uses same face recognition techniques as genuine user does and makes use of low quality image so that he may get the access of the system for his personal gain and gets all the credentials by pretending genuine user.

**Figure 5. Low Quality Image.**

### *8.2.2 Fake input*

A case of common attack , where a user wants to get the access of a system by finger print scanning procedure but the attacker can get access to the system through the fake prints of finger and may get all confidential information which he is looking for.

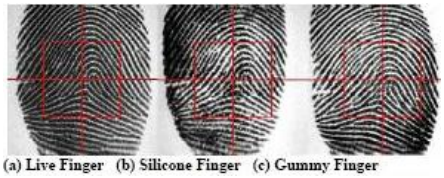


**Figure 6. Fake Image**

## 8.3 Authentication

Authentication maintains the identity of one party to another. It establishes the identity of an individual to some part of the system, typically with the help of a password. Technically, Authentication can be system- to-system or process-to-process and mutual in both directions. Conformation of the user genuine identity is done by authentication process [9]. Database from where user collects information must be a genuine resource. Whenever a query gets processed from the database, it must be authentic. It can be related to 3 factors which are:

1. Something you have.
2. Something you know.
3. Something you are.

'Something we have' is like an ATM card of a person which can be easily misused but the basic requirement is to have pin no of that ATM. The pin number falls in the category of "Something you know". With the help of these two, one can misuse it. "Something you are" shows your physically presence at any time like Biometrics process .


**Figure 7. Authentication.**

## 8.4 Value

It can be classified on following parameters:

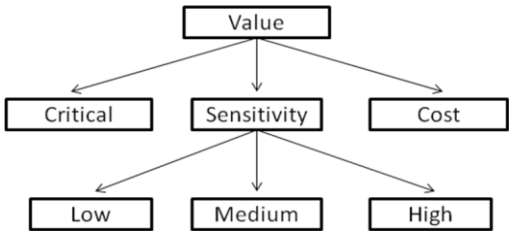


**Figure 8. Information value.**

## 8.5 Criticality

The immediate need behind the retrieval of the information, i.e. whether one needs the information urgently. User can prioritize the information on the basis of his need. If we look around us, critical information of the user is available on the internet, real time census, bank account information etc. these are the very serious security issues about retrieval of information.

Let us consider example of information retrieval by Cyber Forensic tool 'Encase' from a Pen drive. In an organization, employee X carries his financial information in his pen drive of his account details like account no, passbook number, pin number of ATM card , internet banking details etc. By the tool an image of the pen drive is made and then by the image, all the details gets extracted from the part which has been recently deleted and one can easily recover it from the

## 8.6 Sensitivity

Information content requires constant monitoring & handling, especially in some cases where inappropriate access of the information may result in penalties, commercial losses, identity theft, privacy invasion, or unauthorized access by an identity or many identities [10]. Figure 7 Categorizes sensitivity on the basis of three factors – low, medium and high.

## 8.7 Cost

It is one of the major factors where the value of the information is calculated. The cost of the retrieved content is the inevitable attribute which helps user to decide whether he should proceed for content or not. Cost leads us to financial loss; reputational loss etc. In some or other way it is related to the information which is having some cost, of which some have high and some have low cost. As computer's database hard drive is of high cost, as it saves critical information. By any means it can be retrieved from our system database etc. and may cause financial or reputational loss to any organization.

**Table 1: Value**

| Process/Value | A | B | C |
|---|---|---|---|
| **Critical** | Y | N | N |
| **Sensitive** | N | Y | Y |
| **Cost** | N | N | Y |

It defines the parameters on which retrieved information makes valuation. Here, the measurement of the nature of retrieved information is defined on the basis of three parameters - criticality, sensitivity and its cost.

**Table 2: Sensitivity**

| Process/event | A | B | C |
|---|---|---|---|
| **Low** | Y | N | N |
| **Medium** | Y | N | Y |
| **High** | N | Y | N |

Sensitivity defines how important the retrieved information to the user is, who will prioritize it accordingly.

**Frequent access of data:**
Technological growth needs frequent access of data where every user gains credibility through various modes of the digital world with fast access of data, either by uploading or by downloading. Slow processing or accessing will go in vain with no result. For example, a user Googles a word "bright" on the internet and gets the result as "intelligent" at the 10th page of search instead of 1st page which is required, so the frequent access of data do not exist in this case which reflects a big flaw of the information retrieval process.

# 9. RISK ANALYSIS AND MEASUREMENT IN INFORMATION RETRIEVAL

**Risk:** Probability of suffering of loss destruction or modification of retrieved information [11]. Majorly the risk measurement probably depends on the probability of an attack. As lack of reliable access of data makes it difficult to access the data. Exposure of Risk of any retrieved information can be calculated on the basis of two parameter - probability & severity. Emphasis has been given to focus on the exposure of the risk and its impact. Probability states the occurrence of the risk & severity states its impact. If the control strength in the retrieving of Information gets existence then it will result into reduced chances of risk. It's exposure helps in deciding how to manage risk. User can also prioritize the order in which the risk is assessed.

$$\text{Risk Exposure} = \text{Probability} \times \text{Severity}$$

**Equation (3)**

**Table 3: Formulated representation of Risk exposure**

| Risk | Probability(P) | Severity(S) | Ex=P×S |
|---|---|---|---|
| **MIM** | **0.5** | **8** | **4** |

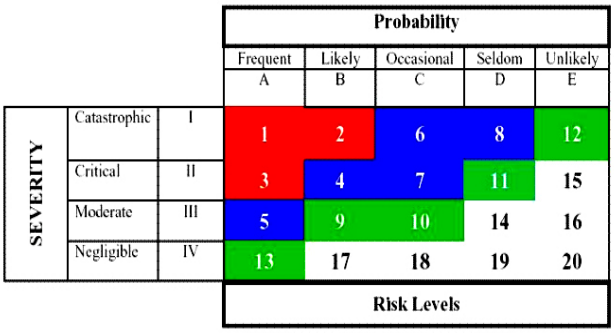


**Figure 9. Risk Assessment Matrix.**

This figure represents a matrix relationship between severity and probability. Probability is defined in 5 ways & severity in 4. Red colour shows that risk is high and one should take care of the process as it is highly vulnerable and the matric's blue part shows that risk is moderate and green part shows that risk is minimal.

# 10. CONCLUSION AND FUTURE WORK

The unprecedented concern of accuracy issues in information retrieval system focuses on improvement of effectiveness of the retrieved data with a goal to enhance the process flow of data retrieval. Demonstration of metrics on values and sensitivities defines the nature of information based on high,

low, medium factors. Analysis of risk exposure through matrices would minimize the risk probability when it comes to extract the required information from any corner of the world through internet. Summarized different accessing modes & appropriately incorporated suggested views through this paper would facilitate the future user and purify the data retrieval system spontaneity.

## 11. REFERENCES

[1] Foote, Jonathan (1999). "An overview of audio information retrieval". *Multimedia Systems* (Springer).

[2] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press

[3] Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research". Informing Science.

[4] Frakes, William B. (1992). Information Retrieval Data Structures & Algorithms. Prentice-Hall, Inc

[5] Joel Stubinz, Samuel Whighli, 2004. Information retrieval system design for high effectiveness. www.cs.rmit.edu.au/~jz/sci/p3.pdf

[6] Dr. Gregory B. Newby, Kevin Gamiel 2002. Secure Information Sharing and Information Retrieval Infrastructure with Grid IR. In NSF/NIJ Symposium on Intelligence and Security Informatics.

[7] Manning, C. D., Raghavan, P., and Schütze, H. 2008. no., pp. Muict.polppolservice.com.-dl-acm.org.

[8] Cooper, W. S. -, 1971 Information storage and retrieval – Elsevier GW Furnas, S Deerwester, ST Dumais- in information retrieval, 1988 - dl.acm.org.

[9] Korfhage, Robert R. (1997). Information Storage and Retrieval. Wiley. pp. 368 pp.

[10] Macleod, I. 1991. Text retrieval and the relational model. J. Am. Soc. Inf. Sci. 42, 3, 155-165.

[11] Bob Blakely, Ellen Mc Dermott, Dan Geer. 2002. Information security is information risk management. In *NSPW'01*.