

# Journey of Information Retrieval to Information Retrieval Tools - IR&IRT A Review

*Tapan P Gondaliya*

*Hiren D Joshi*

## Abstract

*We can say that after thousands of year's public have to realize and know the value of archiving and finding the information. With the invention of computers, it becomes possible to store a large amount of data and finding the useful information from them. Basically Information retrieval field's was dawn in year 1950. [1] Meaning of term Information retrieval can be a very wide. Information Retrieval is one kind of activity that main goal is to obtaining the data resources according to the information needed from a collection of information resources. In this kind of activity search is mainly based on the full text or metadata and other content based indexing. [2] Information Retrieval is mainly used for searching some particular result from the large amount of data. Best example of the information retrieval is searching a strings in particular web search engine. In this paper we first of all describe the Information retrieval & its process cycle after that different model of IR and in next phase of this paper we describe the three different Information Retrieval Tools and comparative study of that tool.*

**Keywords:** Database Systems, Information Retrieval, Information Retrieval Model, Information Retrieval Process Cycle, Information Retrieval System, Information Retrieval System, Information Retrieval Tools, Lucene, Solr, Terrier

## 1. Information Retrieval



**Figure 1: Information Retrieval**

Information Retrieval is the kind of technology that deals with retrieval of unstructured data, like textual based documents in response to a query or topic statement. [4] Information retrieval is finding a data

or information of an unstructured nature that satisfies an information need from within large collections. According to different authors different meaning of information retrieval here we defines some of the definition of the information retrieval according to authors.

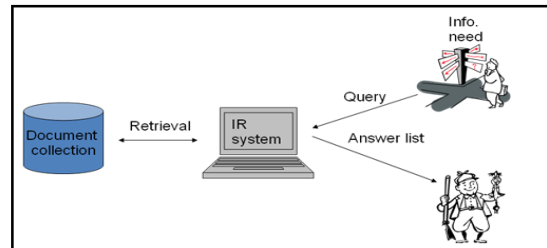
According to Calvin Mooers Information retrieval is the methods were user of information is able to convert his need for information into actual list of documents in storage. It is the finding or discovery process with respect to stored information. [5][6] According to Science and Technology Dictionary, information retrieval is the technique and process of searching, recovering, and interpreting information from large amounts of stored data. [7] According to the Wikipedia IR (Information retrieval) is the type of activity for obtaining information resources related to an information need



11<sup>th</sup> International CALIBER-2017  
Anna University, Chennai, Tamil Nadu 02-04 August, 2017  
© INFLIBNET Centre, Gandhinagar, Gujarat

from a collection of information resources. And here Searches can be mainly based on metadata or on full-text indexing.[8] McGill / Salton 1983 gave this definition An Information retrieval system is a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations. Definition According to the Britannica Concise Encyclopaedia an Information retrieval is recovery of information, especially in a database stored in a computer.

Goal = find documents relevant to an information need from a large collection of documents [12]



**Figure 2: Information Retrieval System**

**What kind of Information basically Retrieves**

**Types of Information**



**2. Information Retrieval System**

Information retrieval system is a built in user interface that is basically used for the searching or retrieving some useful data in the bunch of different data sources according to the user wants. Information retrieval systems originally treated documents as a collection of words. Web search engines are one of the best examples of the Information Retrieval System for example Google, Yahoo and Bing.[11]

**3. IR System Vs DB System**

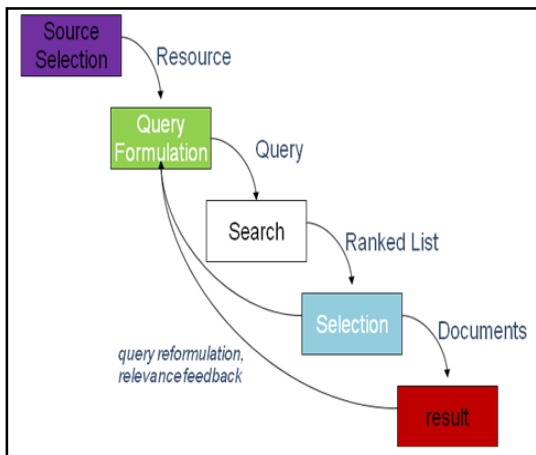
Information retrieval system is totally different than the database system. Let's see how can both are different.

**Table 1: IR System Vs DB System**

IR System	DB System
1) Don't deal with Transaction updates including Concurrency Control and recovery.	1) Deal With Transaction updates like a Concurrency control and recovery.
2) IR system deal with Un Structured data or Semi Structured data Like a Picture, Video, Audio, Text	2) Data Base System only Deal with Structured Data. Like a textual data or information.
3) No Schema Used	3) Schema Used
4) Matching Sometimes relevant	4) Matching Exact. Always correct in a formal sense.
5) SQL Query language is used	5) Natural Language used (Used Keywords )
6) Most Familiar Example of the IR System is Search Engines like a Google, Yahoo, Bings	6) Most familiar example of the Data Base System is Oracle, Dbase, Sql

#### 4. Information Retrieval Process Cycle

Basically an information retrieval process start when a particular user enters a query into the system for finding some particular information or data like a Image, Files, Video, Audio etc. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. In this queries are the formal statements or key words that user wants from the IR System for example search strings in web search engines. Web search engines are the most familiar example of IR systems.



**Figure 3: Information Retrieval Process Cycle**

In information retrieval process first step is the sources selection in this steps system decides that where I take the data after that query formulation will be done in this steps system decide and create a structure of the query and then searching will be done in searching it will give the all relevant data as per the user query and after user will select the data as per his requirement. Let see the real life example. Suppose one user need to find the simple answer of the simple question.



**Figure 4: Example of Information Retrieval System**

#### 5. Information Retrieval Models

Few years ago Information Retrieval systems was kind of Boolean systems which was allowed users to specify their information need using a complex Combination of Boolean ANDs, ORs, NOTs. In this kind of the systems have a several disadvantages from the user point of view for example No inherent notion of document ranking and it's very tough for a user to searching a good relevant documents. Even in Boolean system there are no such features of document relevance ranking or order by dates. Even though it has been shown by the research community that Boolean systems are less effective than ranked retrieval systems. [1] There are different kinds of information Retrieval Model available for the purpose of finding relevant data. First one is

Boolean model second is Vector space model and last but not least Probabilistic model.

**5.1 Boolean Model**

Simply Boolean retrieval model is a model which is used for information retrieval. In this model we can pose a query in the form of Boolean expression term, terms are combined with different operators like AND, OR, NOT.[12] This model is one of the oldest as well as the simplest model in the field of information retrieval. This model is based on the Boolean algebra logic and the classical set theory. [19]In this model documents and query are indicated as a set of index term. This model is also used for different Boolean operations like AND, OR, NOT while in query formulation. [17][20] In this model there are some disadvantages like only exact matching possible, partial matching is not possible, Query language expressive but more complicated, Retrieval document is also not ranked. [17]

**Table 2: Boolean model Advantages & Disadvantages.**

Advantages	Disadvantages
1. Boolean model is Simple & Oldest Model.	1. Query formulation is difficult for most of the user.
2. Can be very effectively Implemented.	2. Difficulty when the size increases with collection size
3. Work well when you know exactly what you're looking for because this model is used for exact matching.	3. Index vocabulary same as query vocabulary
4. Predictable and easy to explain.	4. Retrieval documents not ranked.

**5.2 Vector Space Model**

This model is an algebraic type of model and in vector space model text is represented by a vector of terms and terms are typically words and phrases for example index terms. [1][14] In This Model Index terms are assigned positive and non-binary weights. The index terms in the query are also weighted. Query and the document in the vector form is as under [15][17]

**Table 3: Vector Space Model Advantages & Disadvantages**

Advantages	Disadvantages
1. Simplest model based on linear algebra.	1. Theoretically assumes terms are statistically independent.
2. This model is Term weights not binary.	2. In this model long document is poorly represented.
3. Its Allows computing a continuous degree of similarity between queries and documents.	3. Keyword must be precisely match document terms. Substrings result in a false positive match.
4. Model also allows ranking documents according to their possible relevance.	
5. Partial matching possible in this model	

**5.3 Probabilistic Model**

The most important part of this kind of the model is attempt to rank documents by their probability of relevance given a query. [17][20] Probabilistic model archives almost correct match in a set of documents. [21] Some of the advantages of this kind of model are simple query formulation possible,

straightforward relevance ranking, Sound mathematical as well as the theoretical mode & very effective model then the other model of information retrieval.

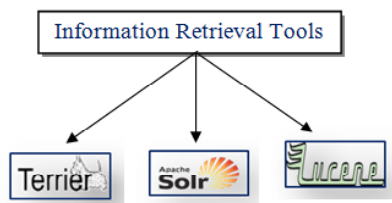
**Table-4 Probabilistic Model Advantages & Disadvantages.**

Advantages	Disadvantages
1. Straight forward relevance ranking	1. Unrealistic assumptions because term independence
2. Simple query formulation	2. Probabilities difficult to estimate
3. Sound mathematical & theoretical mode	
4. Effective model	

In above part of the paper we discussed the definitions of Information Retrieval, we differentiate IR System Vs Database System and then we describe the Goal, Process cycle and different models of Information Retrieval. Now in this paper very next phase we will discuss the different types of the Information retrieval system tools and its comparative study.

**6. Information Retrieval Tools**

There are lots of tools available for the information retrieval but here we discuss mainly 3 tools first one is Terrier second Solr and last but not least Lucene.

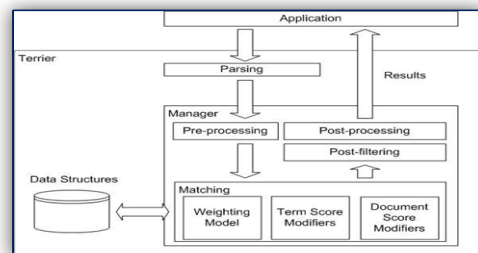


**Figure 5: Different Types of Information Retrieval Tools**

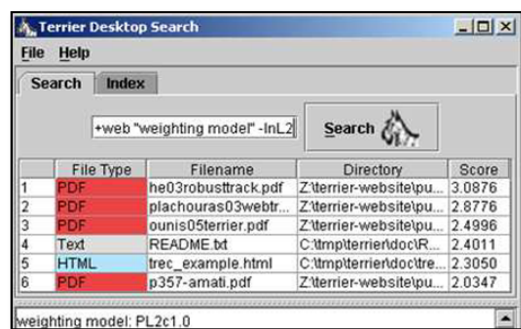
**6.1 Terrier**

**TERabyte RetrIEveR[24]**

Terrier is the one of the high performance as well as the scalable search engine. [22] This tools introduced in 2006 at Glasgow University and developed by Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. [22][23] Terrier is a one of the open source product and used in rapid development of the large scale information retrieval application. [23][24] Terrier is a highly flexible, efficient, very effective platform for IR research, readily deployable on large-scale collections of documents. [24] [22] Terrier is mainly written in the Java Language and work on different operation system like windows, Mac, Linux, and also provides a stat of Term art indexing as well as retrieval functionalities.



**Figure 6: Architecture of Terrier**



**Figure 7: Document Search in Terrier Tools**

In above architecture of Terrier tools is mainly communication with the manager, in this manager actually runs desired matching module and after that matching assign scores to the particular documents as well. Scores assign using weighting model and document scores modifiers. [22] Weighting Model is instantiated and document scores for the query are then computed. To improve the scores, query moves to post processing. [26] In this module also used (DFR) domness framework approach that basically supplies parameter free probabilistic model. [22][26][27] In this architecture Post filtering is the final step in Terrier’s retrieval process, where a bulk of filters can remove already retrieved documents, which do not satisfy a given condition. [26]

6.2 Solr

Solr is an open source enterprise search platform of Apache software foundation. Developed by Yonik Seeley at CNET Networks in year 2004. [28] Solr is basically written in Java language and work in cross platform as well. [29][30] Solr runs as a standalone full-text search server and also provides different features like full text search, hit highlighting, real time indexing, dynamic clustering, database integration, NoSql features and last but not least rich document handling. [28][29] Solr is the one of the biggest tools for the information retrieval system and used in many real life applications likes’ media, e-commerce, job portal and careers site, enterprise search and social media search. [32]

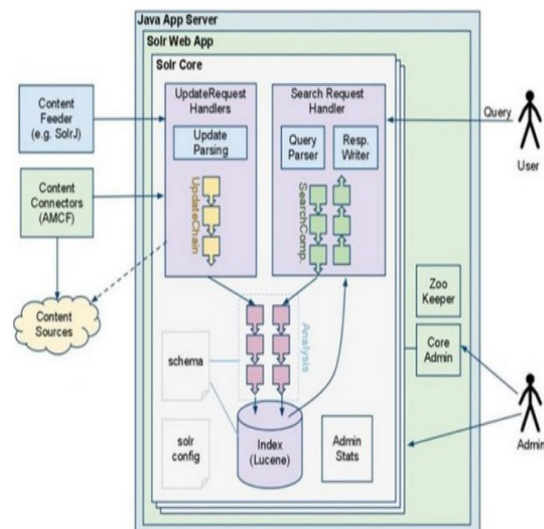


Figure 7: Architecture of Solr



Figure 8: Search String in Solr Tools

In above solr architecture when the user enter the search button search query is proposed Request handler in above figure, basically search request handler is defines the logic to be used when Solr processes a request. Solr supports a variety of request handlers. In search request handler includes



response writer and query parser. Query parser:-To process a search query of request handler. Query parser is responsible for parsing the textual query and converting them into corresponding lucene query object. Different query parser has different syntax as well. Response writer:-component builds the query response object in the required format for the last presentation, mainly XML/JSON object is returned. In Solr has the cloud based architecture. Solr is a kind of system in which data is organized into multiple instances, or in the form of shards, that can be hosted on multiple machines, with replicas providing redundancy for both scalability and fault tolerance. ZooKeeper is a server that helps to manage the whole structure so that both indexing and search requests routed are in sync. [33] Solar solution is used in many real life applications likes yp.com, McClatchy—leading newspaper publisher, zappos.com, smithonian, dig.com, buy.com. [32]Let see the screenshot of the solr.

### 6.3 Lucene

Lucene is highly flexible, scalable open source IR software tools founded in 1999 by Doug Cutting. Lucene is also a product as well as part of apache software foundation. Basically this is written in JAVA language and ported to other programming language like Delphi, Perl, C#, C++, Python, Ruby, and PHP. [35][36] It will work in cross platform. Mainly lucene is used for search and index kind of activity. This tools is high performing and suitable for any application that requires for full text indexing and searching capability, Lucene has been widely recognized for its utility in the implementation of Internet search engines and local, single-site searching.[35][37] Lucene has efficient & precise search tools. It retrieves the documents query based

on their ranking. Its also provides different types of queries like Phrase Query, Wildcard Query, Range Query, Fuzzy Query, Boolean Query.[38] Lucene is also used in many popular web sites like Wikipedia, LinkedIn, Monster.com, and FDA. [32]

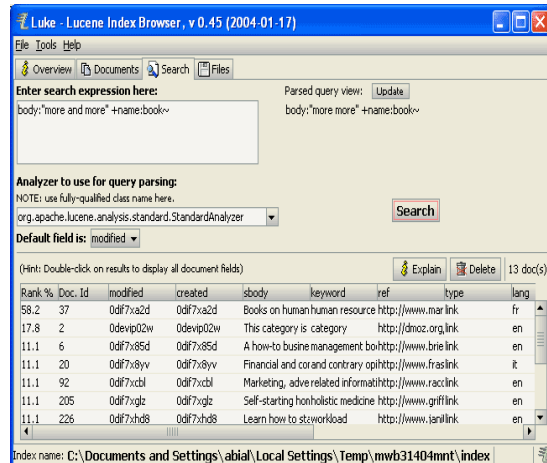


Figure 9: Searching Text in Lucene Tools

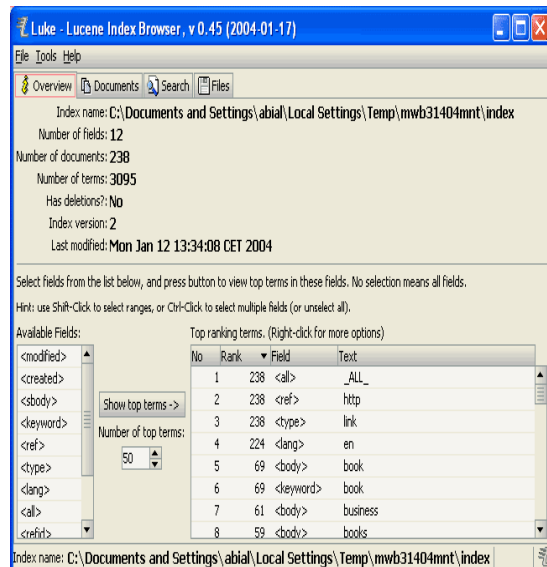
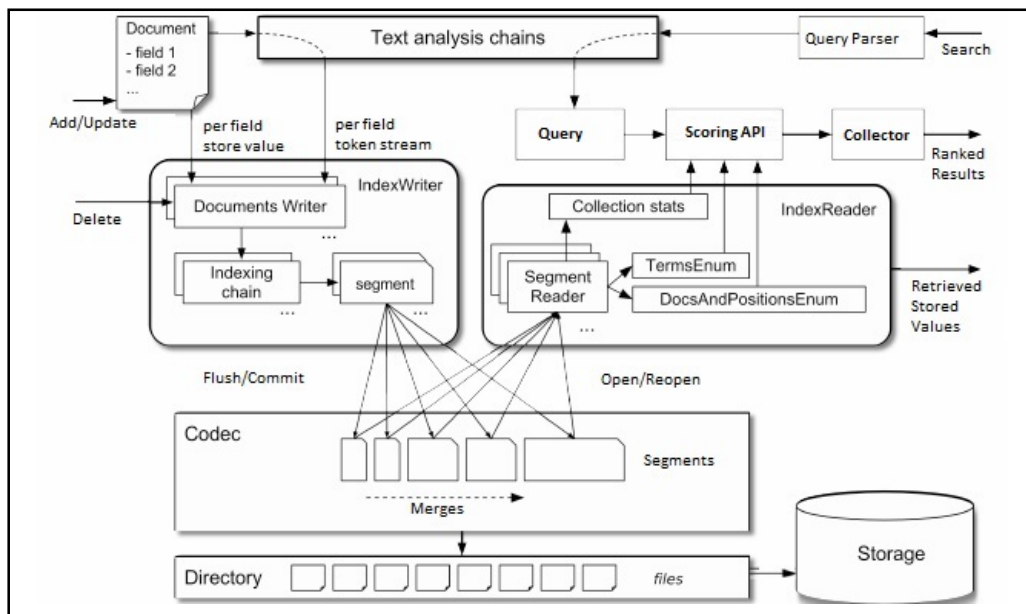


Figure 10: Overview Tab in Lucene Tools

Here above is the screenshot of the searching the simple textual data in Lucene tool. And second screenshot is overview tab of Lucene.



**Figure 11: Architecture of Lucene**




In above Lucene architecture first of all lucene conducts its searches by means of queries entered by the users, this queries processed by the Lucene Query Parser. Basically query parser is a kind of lexical analyzer which interpreted the user search string or information as queries which may be understand by the lucene. Queries are mainly creation of two parts in lucene one is the terms and second one is the operators. This feature is useful for the users to find the specific fields such as authors, titles. Lucene uses Boolean operators like to form complex queries. Boolean operated likes AND, OR, NOT, +, -. [41] Lucene stores all of the important information about the documents upon which its searches are conducted in files called indexes. Indexes contains a sequence of documents, which

themselves consist of sequences of fields. Fields are named sequences of terms, and terms are called simple strings. In lucene architecture Indexes stores the information regarding terms in order to improve search efficiency. Rather than listing which terms are contain in each documents, in lucene indexes list which documents contain each term it is called inverted indexing. Field in the indexed it may be “stored” in this case the text is inverted or in another case “indexed” in this case the not inverted index. In lucene fields may be tokenized & placed into the index as individual tokens as well. [41] Analysis chains consist of character filters, tokenizes and series of token filters that modify the original token stream. Custom token attributes can be used for passing bits of per-token information between the elements of the chain. [37] Lucene includes a total of five character filtering Implementations, 18



tokenization strategies and 97 token filtering implementations and covers 32 different languages. [37] Above token streams performing different functionalities like a tokenization by rules based pattern based and dictionaries, specialized token filter for numeric value and date, stop word remover, creation of word level N-Grams or creation of character as well [37] In Lucene architecture indexes may be divided into segments, each index is acts as a fully independent index. Segments searched by either themselves or together with other indexes, including other segments. In index each documents assigned as a document number starting from zero for first document added. Index stores a list of all terms which make up the fields, along with the number of documents which contain the term, pointers to each term's frequency and proximity data within the Term dictionary. **Frequency data** stores the document numbers of each document containing the term along with the frequency with which the term appears and **Proximity data** stores the positions where the term appears in document. Index segment is also containing normalization values for each field that is a part of the calculation for the score used to rank search results. Term vectors for every fields which generally store term text and term frequency, and a list of deleted documents. [41] In next phase of this paper we compare the above three different IR Tools with its different terms and factors.

Table 5: Terrier Vs Solr Vs Lucene

Comparative Study of Information Retrieval Tools			
Tools Name			
Authors	<ul style="list-style-type: none"> <li>→ Ian Gurns</li> <li>→ Gianni Amati</li> <li>→ Vassilis Plachouras</li> <li>→ Ben He</li> <li>→ Craig Macdonald</li> <li>→ Christina Lioma</li> </ul>	→ Tomik Seeley	→ Doug Cutting
Year of	2006 At Glasgow University	2004 At CNET Networks	1999 Apache Software Foundation
Written In	Java	Java	Java
License By	Mozilla	Apache	Apache
Features	<ul style="list-style-type: none"> <li>→ Free Open Source</li> <li>→ Flexible, Effective, Efficient Search Engine</li> <li>→ Multi-Lingual</li> <li>→ Cross-Platform</li> </ul>	<ul style="list-style-type: none"> <li>→ Free &amp; Open Source</li> <li>→ Full Text Search, Hit Highlighting</li> <li>→ Real Time Indexing</li> <li>→ Rich Doc. Handling</li> <li>→ Dynamic Clustering</li> <li>→ Cross Platform</li> <li>→ Multi-Lingual</li> <li>→ Highly Scalable</li> </ul>	<ul style="list-style-type: none"> <li>→ Free &amp; Open Source</li> <li>→ Highly Flexible &amp; Scalable</li> <li>→ Full Text Indexing &amp; searching capability</li> <li>→ Cross platform</li> <li>→ Provides different queries like Parse, Wild card, Boolean queries.</li> </ul>
Uses	<ul style="list-style-type: none"> <li>→ Used in rapid development of large scale IR application</li> <li>→ Searching</li> <li>→ Indexing</li> </ul>	<ul style="list-style-type: none"> <li>→ Full text search</li> <li>→ Searching &amp; Indexing</li> <li>→ Automated Failure and Recovery</li> <li>→ Fault Tolerant</li> </ul>	<ul style="list-style-type: none"> <li>→ Full text indexing</li> <li>→ Searching</li> <li>→ TML Parsing</li> <li>→ Internet search engines &amp; local,</li> <li>→ Single-site searching</li> </ul>
Used BY	<ul style="list-style-type: none"> <li>→ Desktop Search</li> <li>→ Tec Terrier</li> </ul>	<ul style="list-style-type: none"> <li>→ yp.com</li> <li>→ McClatchy</li> <li>→ zappos.com</li> <li>→ Smithsonian</li> <li>→ dig.com</li> <li>→ buy.com</li> </ul>	<ul style="list-style-type: none"> <li>→ Wikipedia</li> <li>→ LinkedIn</li> <li>→ Monster</li> <li>→ FDA</li> </ul>

## 7. Conclusion

In this paper first of all authors explained the information retrieval and information retrieval system with its process cycle and example. After then that describes how can differ the information retrieval system then the database system. In next phase authors explained three important information retrieval models and its advantages and disadvantages. In final stage of the paper we describe the information retrieval tools with its architecture and its description. In the last compare these tools with its different terms.

## References

1. Amit Singhal ,2001, “Modern Information Retrieval: A Brief Overview”, IEEE, TREC 2001
2. Website, Wikipedia, ‘Information Retrieval’, [https://en.wikipedia.org/wiki/Information\\_retrieval](https://en.wikipedia.org/wiki/Information_retrieval)
3. Minoru Etoh, Xing Xie, Wang-Chien Lee, Qiang Yang, (FEB-2010) “Introduction to Mobile Information Retrieval”, IEEE Computer Society, 1541-1672/10
4. Ed Greengrass, 30 Nov 2000, “Information Retrieval: A Survey”, Ebook
5. Birger Hjørland, Information Retrieval <http://www.iva.dk/>
6. Mooers C. N., 1951, Zetocoding applied to mechanical organization of knowledge, American Documentation, 2, 20-32.
7. Meenakshi Sinha , “Information Retrieval and its Legal Impact on the Society” , Hidayatullah National Law University Raipur, [http://www.legalserviceindia.com/articles/in\\_ret2.htm](http://www.legalserviceindia.com/articles/in_ret2.htm)
8. Website, Wikipedia, “Information Retrieval”, [http://en.wikipedia.org/wiki/Information\\_retrieval](http://en.wikipedia.org/wiki/Information_retrieval)
9. <http://boston.lti.cs.cmu.edu/classes/11-744/>
10. Lecture Notes, Ms. K APARNA, “Information Retrieval System”, [http://www.iare.ac.in/sites/default/files/lecture\\_notes/irs%20notes\\_0.pdf](http://www.iare.ac.in/sites/default/files/lecture_notes/irs%20notes_0.pdf)
11. Silberschatz, Korth, Sudarshan, SEP 2005, “Information Retrieval”, Database System Concepts, Chapter 19, 5th Edition, <https://www.cse.iitb.ac.in/~sudarsha/db-book/slide-dir/ch19.pdf>
12. Jian-Yun Nie, “Introduction to Information Retrieval”, Morgan Schooley, University of Montreal, Canada.
13. Malathi Murugan, Oct 2013 “Query formulation process”, Technology
  - a. <http://www.slideshare.net/malathimurugan/query-formulation-process>
14. Gerard Salton, A. Wong, C. S. Yang Nov 1975, “A vector space model for information retrieval”, Communications of the ACM, 18 (11): 613–620
15. [https://en.wikipedia.org/wiki/Vector\\_space\\_model](https://en.wikipedia.org/wiki/Vector_space_model)
16. [https://en.wikipedia.org/wiki/SMART\\_Information\\_Retrieval\\_System](https://en.wikipedia.org/wiki/SMART_Information_Retrieval_System)
17. Priyanka Mesariya, Nidhi Madia, Abhishek Kumar, March 2016, “Document Ranking using Customizes Vector Method – A Review”, IJCSMC, Vol-5, Issue-3
18. Premalatha, R., S. Srinivasan, 2014 "Text processing in information retrieval system using

- vector space model", International Conference on Information Communication and Embedded Systems (ICICES 2014), IEEE
19. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, 1999, "Modern Information Retrieval", ACM Press.
20. Joydip Datta, Dr. Pushpak Bhattacharyya, April 2016, "Ranking in Information Retrieval", M.Tech Seminar Report, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay
21. Raman, Shivangi, Vijay Kumar Chaurasiya, Swaminathan Venkatesan, 2012, "Performance comparison of various information retrieval models used in search engines", Communication, Information & Computing Technology (ICCICT), IEEE Conference.
22. Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma, Aug-2006, "Terrier: A High Performance and Scalable Information Retrieval Platform", ACM SIGIR'06 Workshop on Open Source Information Retrieval, Seattle, Washington, USA.
23. Website, "terrier", Wikipedia [https://en.wikipedia.org/wiki/Terrier\\_Search\\_Engine](https://en.wikipedia.org/wiki/Terrier_Search_Engine)
24. Iadh Ounis, Christina Lioma, Craig Macdonald, Vassilis Plachouras, Feb-2007, "Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web", CEPIS (Council of European Professional Informatics Societies, Vol- VIII, issue No. 1
25. Website, "terrier", <http://terrier.org/>
26. Anshita Talsania, Prof. Sandip Modha, Prof. Hardik Joshi, Dr. Amit Ganatra, Dec-2015,

- "Automated Story Illustrator", Fire-2015, DAIICT, Gandhinagar
27. G. Amati, 2003, "Probabilistic Models for Information Retrieval based on Divergence from Randomness", PhD thesis, Department of Computing Science, University of Glasgow
28. Website, "Solr", Wikipedia, [https://en.wikipedia.org/wiki/Apache\\_Solr](https://en.wikipedia.org/wiki/Apache_Solr)
29. Website, "Solr", Official Site [lucene.apache.org/solr/](http://lucene.apache.org/solr/)
30. Website, "Solr", db Engines, <http://db-engines.com/en/system/Solr>
31. Website, "Solr", Cwiki.apache, <https://cwiki.apache.org/confluence/display/solr/Running+Solr>
32. A Lucid Imagination White Paper, Jan-2014, "The Case for Lucene/Solr: Real World Search Applications", Lucid Imagination
33. Rajani Maski, 2013, "Using Apache Solr for Ecommerce Search Applications", Happiest Minds, IT Services
34. Grant Ingersoll, May 2007, "Search smarter with Apache Solr", Part 1: Essential features and the Solr schema, IBM Corporation.
35. Website, "Lucene", Wikipedia, <https://en.wikipedia.org/wiki/Lucene>
36. Website, "Luke-Lucene", getopt, <http://www.getopt.org/luke/>
37. Andrzej Bialecki, Robert Muir, Grant Ingersoll, Lucid Imagination, Aug 2012, "Apache Lucene 4", SIGIR 2012 Workshop on Open Source Information Retrieval.

38. Mamatha Balipa, Balasubramani R, OCT-2015, “ Search Engine using Apache Lucene”, International Journal of Computer Applications(IJCA), Volume 127, No.9
39. Website, “apache-lucenesearch”, IBM <http://www.ibm.com/developerworks/library/os-apache-lucenesearch/index.html>
40. Website, “Lucene Architecture” <http://sebol.webs.com/architectureoverview.htm>
41. Jhon Whissel, Dec 2009, “Information retrieval using lucene and wordnet”, Thesis, University of Akron
42. Website, Lucene, <http://ostatic.com/lucene/screenshot>
43. Tapan P. Gondaliya, Hiren D. Joshi, Hardik Joshi, Nov- 2014, “Source Code Plagiarism Detection ‘SCPDet’: A Review”, IJCA, Vol-105, No. 17

---

**About Authors**

**Mr. Tapan P Gondaliya**, IT/Network Department and Risk Management System, SKSE Securities Limited, Rajkot  
Email: [tapan.gondaliya@gmail.com](mailto:tapan.gondaliya@gmail.com)

**Dr. Hiren Joshi**, Professor, Department of Computer Science, Rollwala Computer Center, Gujarat University, Ahmadabad.  
Email: [hiren.joshi@baou.edu.in](mailto:hiren.joshi@baou.edu.in)