A

COURSE FILE REPORT

ON

"*INFORMATION RETRIEVAL SYSTEM*"

III B-Tech II Semester

Submitted by

Dr.E.SURESH BABU

Associate Professor

Department

of

COMPUTER SCIENCE & ENGINEERING

CMR ENGINEERING COLLEGE

(UGC Autonomous)

KANDLAKOYA (V), MEDCHAL (M), R.R.DIST.

A.Y 2023-2024

CMR ENGINEERING COLLEGE

UGC AUTONOMOUS

(Approved by AICTE - New Delhi. Affiliated to JNTUH and Accredited by NAAC & NBA)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE INSTRUCTOR NAME: Dr.E.SURESH BABU **ACADEMIC.YEAR:2023-24**

SUBJECT NAME: INFORMATION RETRIEVAL SYSTEM **CLASS ROOM NO: B-207**

E-MAIL ID: sureshbabu.edem@cmrec.ac.in

B.Tech : III YEAR II SEM **CONTACT NO:**9030368361

SEM START AND SEM END DATES: 29-01-2024 TO 08-06-2024

CONTENTS OF COURSE FILE:

1. **Department vision & mission**
2. **List of PEOs , POs & PSOs**
3. **List of COs(Action verbs as per Bloom's Taxonomy)**
4. **Syllabus Copy and Suggested/Reference Books**
5. **Individual Time Table**
6. **Session Plan/Lesson Plan**
7. **Session execution log**
8. **Lecture Notes (Hand Written)**
9. **Assignment Questions along with sample Assignments Scripts**
10. **Mid exam Question Papers along with sample Answers Scripts**
11. **Scheme of Evaluation**
12. **Mapping of COs with POs and PSOs**
13. **COs, POs, PSOs justification**
14. **Attainment of COs, POs and PSOs (Excel Sheet)**
15. **Previous Year Question Papers**
16. **Power point presentations (PPTs)**
17. **Innovative Teaching Methods**
18. **References (Textbook/Websites/Journals)**

1. DEPARTMENT VISION & MISSION

Vision:

To produce globally competent and industry-ready graduates in Computer Science & Engineering by imparting quality education with the know-how of cutting-edge technology and holistic personality.

Mission:

1. To offer high-quality education in Computer Science & Engineering in order to build core competence for the graduates by laying a solid foundation in Applied Mathematics and program frameworkwith a focus on concept building.
2. The department promotes excellence in teaching, research, and collaborative activities to prepare graduates for a professional career or higher studies.
3. Creating an intellectual environment for developing logical skills and problem-solving strategies, thus developing, an able and proficient computer engineer to compete in the current global scenario.

2. LIST OF PEOs, POs AND PSOs

2.1 Program Educational Objectives (PEO):

PEO 1: Excel in professional career and higher education by acquiring knowledge of mathematical computing and engineering principles.
PEO 2: To provide an intellectual environment for analyzing and designing computing systems for technical needs.
PEO 3: Exhibit professionalism to adapt current trends using lifelong learning with legal and ethical responsibilities.
PEO 4: To produce responsible graduates with effective communication skills and multidisciplinary practices to serve society and preserve the environment.

2.2. Program Outcomes (POs):

Engineering Graduates will be able to satisfy these NBA graduate attributes:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions
5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice
9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings
10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions
11. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments
12. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

2.3 Program Specific Outcomes (PSOs):

PSO1: Professional Skills and Foundations of Software development: Ability to analyze, design and develop applications by adopting the dynamic nature of Software developments.
PSO2: Applications of Computing and Research Ability: Ability to use knowledge in cutting edge technologies in identifying research gaps and to render solutions with innovative ideas.

3. List of Cos

CO1	**Define** the basic concepts of information retrieval systems metrics.
CO2	**Explain** indexing techniques and data structures to information retrieval.
CO3	**Calculate** the different weightage techniques to search statement in different databases.
CO4	**Classify** the various techniques of thesaurus generation and binding the search statement to the item
CO5	**Apply** the different searching techniques algorithms and evaluate search engines using binding techniques.

REVISED Bloom's Taxonomy Action Verbs

Definitions	I. Remembering	II. Understanding	III. Applying	IV. Analyzing	V. Evaluating	VI. Creating
Bloom's Definition	Exhibit memory of previously learned material by recalling facts, terms, basic concepts, and answers.	Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, giving descriptions, and stating main ideas.	Solve problems to new situations by applying acquired knowledge, facts, techniques and rules in a different way.	Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations.	Present and defend opinions by making judgments about information, validity of ideas, or quality of work based on a set of criteria.	Compile information together in a different way by combining elements in a new pattern or proposing alternative solutions.
Verbs	• Choose • Define • Find • How • Label • List • Match • Name • Omit • Recall • Relate • Select • Show • Spell • Tell • What • When • Where • Which • Who • Why	• Classify • Compare • Contrast • Demonstrate • Explain • Extend • Illustrate • Infer • Interpret • Outline • Relate • Rephrase • Show • Summarize • Translate	• Apply • Build • Choose • Construct • Develop • Experiment with • Identify • Interview • Make use of • Model • Organize • Plan • Select • Solve • Utilize	• Analyze • Assume • Categorize • Classify • Compare • Conclusion • Contrast • Discover • Dissect • Distinguish • Divide • Examine • Function • Inference • Inspect • List • Motive • Relationships • Simplify • Survey • Take part in • Test for • Theme	• Agree • Appraise • Assess • Award • Choose • Compare • Conclude • Criteria • Criticize • Decide • Deduct • Defend • Determine • Disprove • Estimate • Evaluate • Explain • Importance • Influence • Interpret • Judge • Justify • Mark • Measure • Opinion • Perceive • Prioritize • Prove • Rate • Recommend • Rule on • Select • Support • Value	• Adapt • Build • Change • Choose • Combine • Compile • Compose • Construct • Create • Delete • Design • Develop • Discuss • Elaborate • Estimate • Formulate • Happen • Imagine • Improve • Invent • Make up • Maximize • Minimize • Modify • Original • Originate • Plan • Predict • Propose • Solution • Solve • Suppose • Test • Theory

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing, Abridged Edition. Boston, MA: Allyn and Bacon.

Action Words for Bloom's Taxonomy					
Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
define	explain	solve	analyze	reframe	design
identify	describe	apply	compare	criticize	compose
describe	interpret	illustrate	classify	evaluate	create
label	paraphrase	modify	contrast	order	plan
list	summarize	use	distinguish	appraise	combine
name	classify	calculate	infer	judge	formulate
state	compare	change	separate	support	invent
match	differentiate	choose	explain	compare	hypothesize
recognize	discuss	demonstrate	select	decide	substitute
select	distinguish	discover	categorize	discriminate	write
examine	extend	experiment	connect	recommend	compile
locate	predict	relate	differentiate	summarize	construct
memorize	associate	show	discriminate	assess	develop
quote	contrast	sketch	divide	choose	generalize
recall	convert	complete	order	convince	integrate
reproduce	demonstrate	construct	point out	defend	modify
tabulate	estimate	dramatize	prioritize	estimate	organize
tell	express	interpret	subdivide	find errors	prepare
copy	identify	manipulate	survey	grade	produce
discover	indicate	paint	advertise	measure	rearrange
duplicate	infer	prepare	appraise	predict	rewrite
enumerate	relate	produce	break down	rank	role-play
listen	restate	report	calculate	score	adapt
observe	select	teach	conclude	select	anticipate
omit	translate	act	correlate	test	arrange
read	ask	administer	criticize	argue	assemble
recite	cite	articulate	deduce	conclude	choose
record	discover	chart	devise	consider	collaborate
repeat	generalize	collect	diagram	critique	collect
retell	give examples	compute	dissect	debate	devise
visualize	group	determine	estimate	distinguish	express
	illustrate	develop	evaluate	editorialize	facilitate
	judge	employ	experiment	justify	imagine
	observe	establish	focus	persuade	infer
	order	examine	illustrate	rate	intervene
	report	explain	organize	weigh	justify
	represent	interview	outline		make
	research	judge	plan		manage
	review	list	question		negotiate
	rewrite	operate	test		originate
	show	practice			propose
	trace	predict			reorganize
	transform	record			report
		schedule			revise
		simulate			schematize
		transfer			simulate
		write			solve
					speculate
					structure
					support
					test
					validate

4. Syllabus copy

UNIT-I

Introduction: Definition, Objectives, Functional Overview, Relationship to DBMS, Digital libraries and Data Warehouses,

Information Retrieval System Capabilities - Search, Browse, Miscellaneous.

UNIT - II

Cataloging and Indexing: Objectives, Indexing Process, Automatic Indexing, Information Extraction, Data Structures: Introduction, Stemming Algorithms, Inverted file structures, N-gram data structure, PAT data structure, Signature file structure, Hypertext data structure and XML data structures,Hidden Markov Models

UNIT -III

Automatic Indexing: Classes of automatic indexing, Statistical indexing, Natural language, Concept indexing, Hypertext linkages.

Document and Term Clustering: Introduction, Thesaurus generation, Item clustering, Hierarchy of clusters

UNIT – IV

User Search Techniques: Search statements and binding, Similarity measures and ranking, Relevance feedback, Selective dissemination of information search, Weighted searches of Boolean systems, Searching the Internet and hypertext –

Information Visualization: Introduction, Cognition and perception, Information visualization technologies.

UNIT - V

Text Search Algorithms: Introduction, Software text search algorithms, Hardware text search systems.

Multimedia Information Retrieval: Spoken language audio retrieval,non-speech audio retrieval,graph retrieval image retrieval,graph retrieval image retrieval,video retrieval.

References (Text books/websites/Journals)

4.1Text books

TEXT BOOKS

T1. Information Storage and Retrieval Systems: Theory and Implementation By Kowalski, Gerald, Mark Maybury Kluwer Academic Press, 2000.

T2.Modern Information Retrival By Ricardo Baeza-Yates, Pearson Education, 2007.

T3.Information Retrieval: Algorithms and Heuristics By David A Grossman and Ophir Frieder, 2nd Edition, Springer International Edition, 2004.

REFERENCE BOOKS

R1.Information Retrieval Data Structures and Algorithms By William B Frakes, Ricardo Baeza-Yates, Pearson Education, 1992.

R2.Information Storage & Retieval By Robert Korfhage – John Wiley & Sons.

R3.Introduction to Information Retrieval By Christopher D. Manning and Prabhakar

Raghavan, Cambridge University Press, 2008

5. INDIVIDUAL TIME TABLE ()

Time	9:10AM-10:10AM	10:10 AM-11:00AM	11:00 AM-11:50AM	11:50AM-12:40PM	12:40PM-1:20PM	1:20PM-2:20PM	2:20PM-3:10PM	3:10PM-4:00PM
Period Day	I	II	III	IV		V	VI	VII
MON							III-D IRS	
TUE		IIID IRS						
WED			III-D IRS					
THU	III-D IRS							III-D IRS
FRI			III-D IRS					
SAT								

6. Session plan

SUBJECT (LESSON) PLAN

S.NO	Topic (JNTU syllabus)	Sub-Topic	NO. OF LECTURES REQUIRED	Suggested Books	Remarks
		UNIT - I			
1	**Introduction:**	Definition, Objectives	**L1**	**T3,R1**	
2		Functional Overview	**L2**	**T3,R1**	
3		Relationship to DBMS	**L3**	**T3,R1**	
4		Digital libraries and Data Warehouses	**L4**	**T3,R1**	
5		Information Retrieval System Capabilities	**L5**	**T3,R1**	
6		SearchMiscellaneous	**L6**	**T3,R1**	
7		Search Capabilities	**L7**	**T3,R1**	
9		Miscellaneous	**L8**	**T3,R1**	**I UNIT COMPLETED**
		UNIT - II			
10	**Cataloging and Indexing:**	Objectives, Indexing Data Structures	**L9-**	**T3,T3,R2**	
11		Automatic Indexing	**L11**	**T3,T3,R2**	
12		ProcessInformation Extraction	**L12**	**T3,T3,R2**	
13		Introduction, Stemming Algorithms	**L13**	**T1,T3,R2**	
14		Inverted file structures	**L14**	**T1,T3,R2**	
15		N-gram data structure	**L15**	**T1,T3,R2**	

16		PAT data structure	**L16**	**T1,T3,R1**	
17		Signature file structure Hypertext data structure	**L17**	**T1,T3,R3**	**II UNIT COMPLETED**
		UNIT -III			
18	**Automatic Indexing: Document and Term Clustering:**	Classes of automatic indexing	**L18**	**T3,R3**	
19		Statistical indexing	**L19-L21**	**T3,R3**	
20		Natural language	**L22-L23**	**T3,R3**	
21		Concept indexing	**L24**	**T3,R3**	
22		Hypertext linkages	**L25-L28**	**T3,R3**	
23		Introduction	**L29**	**T3,R3**	
24		Thesaurus generation	**L30-L31**	**T3,R3**	
25		Item clustering	**L32-L33**	**T3,R3**	
26		Hierarchy of clusters	**L34-L35**	**T3,R3**	**III UNIT COMPLETED**
		UNIT -IV			
27	**User Search Techniques:**	Search statements and binding	**L36**	**T3,R3**	
28		Similarity measures and ranking	**L37-L38**	**T3,R3**	
29		Relevance feedback	**L39**	**T3,R3**	
30		Selective dissemination of information search	**L40**	**T3,R3**	
31		Weighted searches of Boolean systems	**L41-L42**	**T3,R3**	
32		Searching the Internet and hypertext	**L43-L44**	**T3,R3**	

33	**Information Visualization:**	Introduction, Cognition and perception	**L45**	**T1,T3, R2**	
34		Information isualization technologies	**L46**	**T1,T3, R2**	**V UNIT COMPLETED**
		UNIT -V			
35	**Text Search Algorithms& Multimedia Information Retrieval**	Introduction	**L47**	**T1,T3, R2**	
36		Hardware, Software text search algorithms	**L48**	**T1,T3,R2**	
37		text search systems	**L49**	**T1,T3,R2**	
		Spoken language audio retrieval	**L50**	**T1,T3,R2**	
		non-speech audio retrieval	**L51**	**T1**	
		graph retrieval image retrieval	**L52**	**T1**	
		graph retrieval image retrieval	**L53**	**T1**	
		video retrieval	**L54**	**T1**	

METHODS OF TEACHING:

M1:Lecture Method	**M11:Tutorial**
M2:Demo Method	**M12:Assignment**
M3:Guest Lecture	**M13:Industry Visit**
M4:Presentation/PPT	**M14:Project Based Learning**
M5:Mind Map	**M15:Mnemonics**
M6:ATL Lab	**M16:Laboratory Improvement Future Trends**

M7:Group Learning	**M17:Collaborative Learning**
M8:One minute Paper	**M18:Think Pair Share**
M9 :Case Study	**M19:NPTEL Video Lectures**
M10:Flipped Classes	**M20:Innovative Assignment**

7. Session execution log

S no	Unit	Scheduled started date	Completed date	Remarks
1	I	29-01-2024	22-02-2024	COMPLETED
2	II	23-02-2024	21-03-2024	COMPLETED
3	III	21-03-2024	18-04-2024	COMPLETED
4	IV	19-04-2024	02-05-2024	COMPLETED

5	V	03-05-2024	31-05-2024	COMPLETED

8. Lecture notes

9. Assignment Questions

Academic year: 2023-2024

Year/SEM: **III B.TECH II-SEM (R20)**

Subject: **INFORMATION RE TRIEVAL SYSTEM (IRS)**

ASSIGNMENT-I

1. Define IRS? Briefly explain the functional overview of IRS **(CO1)**
2. Explain the Browse capabilities of IRS **(CO1)**

3. a) Define objectives of indexing and Explain the automatic indexing Process. **(CO2)**

   b) What is stemming? Explain the porter stemming algorithm **(CO2)**

4. a) Explain the N-Gram algorithm **(CO2)**

   b) Explain the Inverted file structure **(CO2)**

5. What is the vector weighting ? Explain the inverse document frequency with example? **(CO2)**

Academic year: 2023-2024

Year/SEM: **III B.TECH II- SEM**

Subject: **INFORMATION RETRIEVAL SYSTEM (IRS)**

ASSIGNMENT -II

1. a)Develop the term cluster using following existing clusters with 7 terms CO3

   Class1(term1,term2)
   Class2(term3,term4)
   Class3(ter5,term6)

   b) Explain the one pass assignments with example

2. a)Explain different similarity measures & ranking and with example calculate similarity between the item & search statement? CO4

b) Explain the selective dissemination of information search.

CO4

3. Explain the following visualization techniques. CO4

a)Perspective Wall
b)CONE tree
c) Envision interface

4.. a)Demonstrate Boyre-Moore Algorithm for the following scenario, explain each step. **CO4**
String to be searched: abcac
Input String: ababdcabcdacabcac

b)What are hardware text search algorithms? Explain them in detail. **CO4**

5..Explain the following **CO4**

a)Non speech audio retrieval

b)Multimedia Retrieval.

10.Mid exam question papers

Academic year: 2023-2024

Year/Sem: **III B.TECH II-SEM Sec:B&D**

Subject: **INFORMATION RETRIEVAL SYSTEM (IRS)** **Reg:R20**

MID-I

SET-I

Short answer questions

1. Define Precision and write the formule**(CO1)**
2. What is the vector weighting **(CO1)**
3. Define the proximity. **(CO1)**

4. What is XML Data structures. **(CO2)**
5. Define the statistical indexing. **(CO2)**

Long answer questions

6. Define IRS? Briefly explain the functional overview of IRS **(CO1)** **5M**

(OR)

7. Explain the search capabilities of IRS **(CO1)**
8. What is stemming? Explain the porter stemming algorithm with example **(CO2)**

(OR)

9. Define objectives of indexing and Explain the automatic indexing Process. **(CO2)**
10. Explain the probabilistic weighting index **(CO2)**

(OR)

11. Explain the inverse document frequency with example? **(CO2)**

MID-II

Academic year: 2023-2024

Year/SEM: **III B.TECH II- SEM**

Subject: **INFORMATION RETRIEVAL SYSTEM (IRS)**

Short answer questions

1. Define the thesaurus? **(CO3)**
2. Definition cognition. **(CO4)**
3. What is software text search. **(CO4)**
4. Explain multimedia information retrieval. **(CO4)**
5. Explain the text streaming Architecture. **(CO3)**

Long answer questions

6. Creating **ITEM** Clusters in complete term relation method using four algorithms (**single link,Star,netwrk diagram & cliques techniques**) for below vector matrix. Threshold value is 10 **CO3**

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Item 1	0	4	0	0	0	2	1	3
Item 2	3	1	4	3	1	2	0	1
Item 3	3	0	0	0	3	0	3	0
Item 4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

OR

7. Explain the one pass assignments with example **CO3**

8. Explain the following visualization techniques. CO4

a)Perspective Wall
b)Envision interface

OR

9.Explain the selective dissemination of information search **CO4**

10. Demonstrate Boyre-Moore Algorithm for the following scenario, explain each step. **CO4**
String to be searched: abcac
Input String: ababdcabcdacabcac

OR

11. What are hardware text search algorithms? Explain them in detail. **CO4**

11.Scheme of evaluation

MID-I

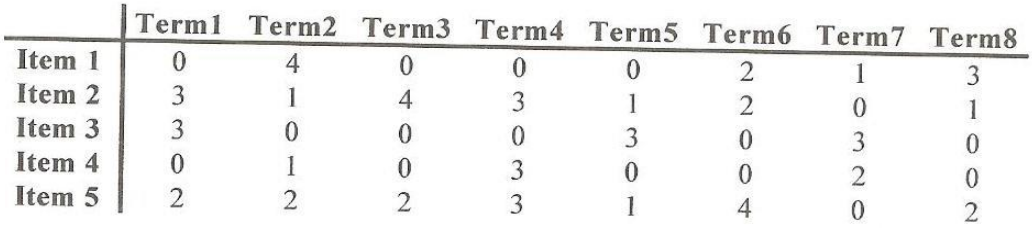
Q.No	Question	Sub division marks	Total marks
	SHORT ANSWER QUESTIONS		
1	Define Precision and write the formule**(CO1)**	**2M**	**2M**
2	What is the vector weighting **(CO1**	**2M**	**2M**

3	Define the proximity. **(CO1)**	**2M**	**2M**
4	What is XML Data structures. **(CO2)**	**2M**	**2M**
5	Define the statistical indexing. **(CO2)**	**2M**	**2M**
	LONG ANSWER QUESTIONS		
6.	Define IRS? Briefly explain the functional overview of IRS **(CO1)**	**1M** **4M**	**5M**
7	Explain the search capabilities of IRS **(CO1)**	**5M**	**5M**
8	What is stemming? Explain the porter stemming algorithm with example **(CO2)**	**1M** **4M**	**5M**
9	Define objectives of indexing and Explain the automatic indexing Process. **(CO2)**	**1M** **4M**	**5M**
10	Explain the probabilistic weighting index **(CO2)**	**5M**	**5M**
11	Explain the inverse document frequency with example? **(CO2)**	**5M**	**5M**

MID-II

Q.No	Question	Sub division marks	Total marks
	SHORT ANSWER QUESTIONS		
1	Define the thesaurus? **(CO3)**	**2M**	**2M**
2	Definition cognition. **(CO4)**	**2M**	**2M**
3	What is software text search. **(CO5)**	**2M**	**2M**
4	Explain multimedia information retrieval. **(CO5)**	**2M**	**2M**
5	Explain the text streaming Architecture. **(CO3)**	**2M**	**2M**

LONG ANSWER QUESTIONS

6.	6. Creating **ITEM** Clusters in complete term relation method using four algorithms (**single link,Star,netwrk diagram & cliques techniques**) for below vector matrix. Threshold value is 10 **CO3** 	**5M**	**5M**
7	Explain the one pass assignments with example **CO3**	**5M**	**5M**
8	Explain the following visualization techniques. a)Perspective Wall b)Envision interface CO4	**2M** **3M**	**5M**
9	Explain the selective dissemination of information search **CO4**	**2M** **3M**	**5M**
10	Demonstrate Boyre-Moore Algorithm for the following scenario, explain each step. **CO5** String to be searched: abcac Input String: ababdcabcdacabcac	**5M**	**5M**
11	What are hardware text search algorithms? Explain them in detail. **CO5**	**5M**	**5M**

12.Mapping of Cos with Pos and PSOs

Course Outcomes (CO)	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2

CO1	2	2	2		3								2	-
CO2		2	3		2								3	-
CO3	2	3	2		2								2	2
CO4	2	2	3	2										2
CO5		2	3	2									2	

13.COs, POs, PSOs Justification

CO1	**Define** the basic concepts of information retrieval systems metrics.
CO2	**Explain** indexing techniques and data structures to information retrieval.
CO3	**Calculate** the different weightage techniques to search statement in different databases.
CO4	**Classify** the various techniques of thesaurus generation and binding the search statement to the item
CO5	**Apply** the different searching techniques algorithms and evaluate search engines using binding techniques.
PSO1	**Professional Skills and Foundations of Software development:** Ability to analyze, design and develop applications by adopting the dynamic nature of Software developments.
PSO2	**Applications of Computing and Research Ability:** Ability to use knowledge in cutting edge technologies in identifying research gaps and to render solutions with innovative ideas.

Justification:

CO1: Define the basic concepts of information retrieval systems metrics.
Correlated with PO1 moderately: Understanding and applying foundational knowledge in mathematics, science, and engineering is crucial for developing efficient IR systems. Engineers need to comprehend algorithms, data structures, and statistical models to design and optimize IR systems that can handle large datasets and deliver accurate results.
Correlated with PO2 moderately: Analyzing complex engineering problems is essential for identifying the key challenges in IR systems, such as indexing, querying, and relevance ranking. Engineers must be able to break down these problems and apply analytical techniques to devise effective solutions.
Correlated with PO3 moderately:. Designing solutions for complex problems is directly related to developing and implementing IR systems. Engineers must create algorithms and system architectures that meet user needs, ensuring that the IR system retrieves relevant and accurate information efficiently.
Correlated with PO5 highly: Utilizing modern tools and technologies is vital for implementing and testing IR systems. Engineers must be proficient in using software tools, programming languages, and platforms to develop, evaluate, and optimize IR systems, ensuring they **are robust and scalable.**
Correlated with PSO1 moderately: **Foundational Knowledge**: A strong understanding of software development principles, including algorithms, data structures, and system design, is crucial for developing effective IR systems. These skills ensure that engineers can create efficient, scalable, and maintainable software solutions. • **Professional Skills**: Proficiency in modern software development tools and practices, such as version control, testing, and continuous integration, is essential for implementing and maintaining IR systems. These professional skills ensure that software development processes are streamlined, collaborative, and adhere to industry standards. • **Problem-Solving Abilities**: Developing IR systems often involves addressing complex challenges, such as handling large datasets, optimizing query performance, and ensuring accurate relevance ranking. A solid foundation in software development equips engineers with the problem-solving skills necessary to tackle these challenges effectively. • **Interdisciplinary Knowledge**: The development of IR systems requires knowledge across multiple domains, including computer science, information theory, and user experience design. A comprehensive foundation in software development provides the interdisciplinary knowledge needed to integrate these domains seamlessly. • **Adaptability**: The field of information retrieval is continually evolving, with new algorithms, technologies, and methodologies emerging regularly. A strong foundation in software development ensures that engineers can adapt to these changes, stay current with the latest advancements, and continuously improve their IR systems.

CO2: Explain indexing techniques and data structures to information retrieval.

Correlated with PO2 moderately: Analyzing complex problems in information retrieval requires understanding the underlying data structures and indexing techniques that optimize search and retrieval processes.

Students need to be able to decompose the problem of document retrieval into manageable parts, such as indexing, query processing, and relevance ranking.

Example: Analyzing the efficiency of different indexing techniques (e.g., inverted index vs. signature files) for a given dataset.

Correlated with PO3 highly:. Designing effective IR systems involves creating and implementing indexing structures that enhance retrieval speed and accuracy.

Students must be able to design solutions that meet specified needs, considering the trade-offs between different indexing techniques and data structures.

Example: Designing a search engine that uses a combination of B-trees for efficient range queries and tries for fast prefix searches.

Correlated with PO5 moderately: Modern IR systems rely on advanced tools and technologies for implementing and managing indexing and search processes.

Students must be proficient in using software tools, programming languages, and platforms to develop, test, and deploy IR systems.

Example: Implementing an inverted index using modern programming languages and libraries, and evaluating its performance using benchmarking tools.

Correlated with PSO1 highly:

Foundational Knowledge: Understanding and implementing efficient indexing techniques and data structures is crucial for developing high-performance IR systems. This foundational knowledge ensures that software developers can design systems that quickly and accurately retrieve relevant documents from large datasets.

Example: Learning to implement and optimize inverted indexes helps developers build search engines that can handle extensive data with low latency.

Professional Skills: Proficiency in modern software development tools and practices is essential for creating robust IR systems. Developers need to use version control, testing frameworks, and continuous integration tools to ensure that their indexing and retrieval systems are reliable and maintainable.

Example: Using tools like Git for version control and Jenkins for continuous integration ensures that the IR system can be developed collaboratively and deployed seamlessly.

Problem-Solving Abilities: Developing IR systems involves addressing complex challenges such as handling large-scale data, optimizing query performance, and ensuring relevance ranking. Strong problem-solving skills enable developers to analyze these challenges and devise

effective solutions.

Example: Analyzing the trade-offs between different indexing techniques (e.g., inverted index vs. suffix arrays) allows developers to choose the best approach for their specific use case.

Interdisciplinary Knowledge: The development of IR systems requires knowledge across multiple domains, including computer science, information theory, and user experience design. A comprehensive foundation in software development ensures that developers can integrate these domains effectively.

Example: Combining knowledge of data structures with user experience principles helps developers design search interfaces that are both efficient and user-friendly.

Adaptability: The field of information retrieval is constantly evolving, with new algorithms, technologies, and methodologies emerging regularly. A strong foundation in software development equips developers with the ability to adapt to these changes, stay current with the latest advancements, and continuously improve their IR systems.

Example: Staying updated with the latest developments in machine learning for IR enables developers to integrate new ranking algorithms that enhance search accuracy.

CO3: Calculate the different weightage techniques to search statement in different databases.

Correlated with PO1 moderately: Weightage Technique: Assigning weights based on the domain knowledge relevance of terms in the query. For example, terms related to the core concepts of the database subject area may be weighted higher.

Justification: Engineers need a solid understanding of the subject domain to accurately assign weights that reflect the importance of terms in the query.

Correlated with PO2 highly: **Weightage Technique**: Utilizing statistical analysis techniques to assign weights based on the frequency of terms within documents and across the database. This could involve techniques like TF-IDF (Term Frequency-Inverse Document Frequency).

Justification: Analyzing the problem of query relevance requires understanding how often terms appear in relevant documents versus the entire database.

Correlated with PO3 moderately: **Weightage Technique***:* Implementing custom weighting algorithms that consider not only term frequency but also document structure (e.g., title vs. body text), document age, or other metadata.

Justification: Designing effective search solutions involves developing algorithms that accurately

prioritize relevant documents based on a combination of factors beyond simple term occurrence.

Correlated with PO5 moderately: **Weightage Technique**: Leveraging advanced search engine capabilities such as semantic analysis, natural language processing (NLP), or machine learning algorithms to dynamically adjust weights based on context and user behavior.

Justification: Modern tools enable the integration of sophisticated techniques that enhance the accuracy and efficiency of search statements by adapting weights to match user expectations and current database trends.

- **Correlated with PSO1 moderately**: **Foundational Knowledge**: Understanding weightage techniques such as TF-IDF and semantic analysis requires a strong foundation in software development principles and algorithms. This knowledge ensures that engineers can implement efficient and effective search algorithms that meet user expectations.
- **Professional Skills**: Proficiency in software development tools and practices enables engineers to design and deploy complex search algorithms, integrating advanced techniques like NLP and machine learning to enhance search accuracy and relevance.
- **Problem-Solving Abilities**: Developing search algorithms involves addressing complex challenges such as handling large datasets, optimizing performance, and ensuring scalability. Software development skills empower engineers to analyze these challenges and implement innovative solutions.
- **Interdisciplinary Knowledge**: Effective search algorithms require integration of knowledge from various domains, including computer science, linguistics (for NLP), and statistics (for TF-IDF). A comprehensive foundation in software development facilitates seamless integration of these domains.
- **Adaptability**: The field of search algorithms is constantly evolving with advancements in NLP, machine learning, and data analytics. Professional skills in software development enable engineers to adapt to these changes, continuously improving search functionalities.

Correlated with PSO2 moderately: **Research Abilities**: Proficiency in computing applications and research enables professionals to conduct in-depth analyses of databases and their contents. This proficiency ensures accurate identification and prioritization of relevant terms through techniques like TF-IDF and semantic analysis.

- **Computational Expertise**: Expertise in computing allows for the implementation of advanced algorithms and models, such as machine learning-based ranking systems, that optimize search effectiveness and user satisfaction in research contexts.

- **Analytical Skills**: Strong research abilities empower computing professionals to critically evaluate search algorithms and techniques, ensuring that they are aligned with the specific requirements and challenges of different databases and research domains.

- **Innovative Solutions**: Applying computing applications and research ability fosters the development of innovative solutions for improving search statement weighting techniques, thereby enhancing the efficiency and effectiveness of information retrieval processes.

CO4: **Classify** the various techniques of thesaurus generation and binding the search statement to the item
Correlated with PO1 moderately: Thesaurus generation techniques require domain expertise to accurately compile or extract relevant terms, ensuring that the search system reflects the specialized terminology and concepts of the database domain.
Correlated with PO2 moderately:. Analyzing user search behavior and content interactions informs the refinement of thesaurus entries, improving search effectiveness by incorporating synonyms and related terms that align with user preferences.
Correlated with PO3 highly: Designing algorithms and systems that integrate thesaurus generation techniques enhances search capabilities, supporting efficient query expansion and semantic enhancement based on structured knowledge representations.
Correlated with PO4 moderately: Investigating and resolving semantic ambiguities and complex search challenges through advanced thesaurus generation techniques involves applying computational and linguistic analyses to optimize search precision and relevance.
Correlated with PSO2 moderately: **Research Abilities**: Thesaurus generation techniques leverage computing skills to systematically analyze and organize terms, enhancing search capabilities by incorporating specialized and evolving research terminology. • **Computing Proficiency**: Proficiency in computing allows for the development of automated algorithms and semantic structures that optimize the organization and retrieval of information based on comprehensive thesaurus entries. • **Problem-Solving Skills**: Applying research abilities and computing expertise supports the resolution of complex semantic challenges, such as synonymy and polysemy, improving query precision and relevance in diverse research domains. • **Innovative Solutions**: By integrating advanced thesaurus generation techniques, computing professionals can innovate and optimize search functionalities, enhancing the efficiency and effectiveness of information retrieval systems in research contexts.

CO5: Apply the different searching techniques algorithms and evaluate search engines using binding techniques.
Correlated with PO2 moderately: Application of these searching techniques involves analyzing user queries and system performance metrics to refine algorithms and enhance search precision.
Correlated with PO3 highly: Designing and implementing algorithms based on these techniques requires expertise in software development to optimize performance and scalability.
Correlated with PO4 moderately: : Investigating and solving complex information retrieval challenges using advanced searching techniques involves applying analytical skills and computational methods to improve search engine effectiveness.
Correlated with PSO1 moderately: **Computing Applications**: Application of these searching techniques and algorithms enhances computing applications by developing robust search engines that effectively process and retrieve information based on user queries. • **Research Abilities**: Proficiency in these techniques involves conducting research to innovate and optimize search algorithms, improving retrieval accuracy and user satisfaction. • **Problem-Solving Skills**: Solving complex information retrieval challenges using advanced searching techniques requires problem-solving skills to optimize algorithms and adapt them to diverse search scenarios.

14. Attainment of Cos, Pos and PSOs

15.Previous question papers

Code No: 155BV

R18

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech III Year I Semester Examinations, March - 2021
INFORMATION RETRIEVAL SYSTEMS
(Computer Science and Engineering)

Time: 3 Hours **Max. Marks: 75**

Answer any five questions
All questions carry equal marks
- - -

1.a) Define Information retrieval systems and write about importance of IRS.
b) What are the capabilities of Information retrieval systems? Describe. [7+8]

2.a) What are digital libraries and data warehouses? Write their significance.
b) Explain the history and objectives of indexing. [8+7]

3. What is stemming? Explain porter stemming algorithm. [15]

4.a) Explain about Inverted file structure with example.
b) Discuss in detail about XML and Hypertext data structures. [7+8]

5. Differentiate between the process of information extraction and document indexing. [15]

6. Discuss about hierarchy of clustering. Compare and contrast term clustering and item clustering. [15]

7. How cognition and perception are significant in information visualization? What are the technologies used in information visualization? Discuss. [15]

8.a) What are hardware text search algorithms? Explain them in detail.
b) Explain audio, video and graph retrieval methods of MIR. [7+8]

---ooOoo---

Code No: 155BV

R18

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech III Year I Semester Examinations, September - 2021
INFORMATION RETRIEVAL SYSTEMS
(Computer Science and Engineering)

Time: 3 hours **Max. Marks: 75**

Answer any five questions
All questions carry equal marks

1.a) List and explain the Data Warehouses.
b) Discuss the relationship between information retrival systems and database management systems. [7+8]

2.a) Difference between the concept of a "Digital Library" and an Information Retrieval System.
b) Give a brief note on Browse Capabilities. [8+7]

3. List and explain the various search algorithms on the PAT tree. [15]

4.a) With the help of a neat diagram, explain the Hidden Markov Models.
b) Explain the similarities between term stemming algorithms and n-grams. Describe how they affect precision and recall. [7+8]

5. What is automatic indexing? List and explain the various types of automatic indexing. [15]

6. Discuss the hypertext linkages and natural languages. [15]

7. Explain the Selective Dissemination of Information Search. [15]

8.a) Illustrate the Software Text Search Algorithms.
b) Give a brief note on Hardware Text Search Systems. [7+8]

---oo0oo---

Code No: 155BV

R18

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech III Year I Semester Examinations, February - 2022
INFORMATION RETRIEVAL SYSTEMS
(Computer Science and Engineering)

Time: 3 hours **Max. Marks: 75**

Answer any five questions
All questions carry equal marks

1.a) With the help of a neat diagram, explain the text normalization process.
b) Discuss the Search Capabilities in detail. [7+8]

2.a) What new areas of information retrieval research may be important to support a Digital Library? Explain.
b) Write the difference between data retrieval and information retrieval. [8+7]

3. Differentiate human indexing and automatic indexing and list the advantages and disadvantages of automatic indexing. [15]

4.a) Explain the History and Objectives of Indexing.
b) Describe the Hypertext and XML Data Structures. [7+8]

5. Prove that a term could not be found in multiple clusters when using the single link technique. [15]

6. What are the tradeoffs in the use of zoning as part of the indexing process? Explain.[15]

7.a) Explain about weighted searches of Boolean systems.
b) Explain about cognition and perception in information visualization. [7+8]

8.a) Discuss the Non-Speech Audio Retrieval.
b) Explain the software text search algorithms in detail. [7+8]

---oo0oo---

Code No.: CS624PE | R20 | H.T.No. | | | 8 | R | | | | | |

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–II–Semester End Examinations (Supply) - January- 2024
INFORMATION RETRIEVAL SYSTEM
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks and may have a, b, c as sub questions.

PART-A **(20 Marks)**

1. a)	List the capabilities of Information Retrieval Systems.	[2M]
b)	Define Data warehouse.	[2M]
c)	What is meant by Inverted File Structure?	[2M]
d)	Define PAT Data Structure.	[2M]
e)	What is Automatic Indexing?	[2M]
f)	Define Concept Indexing.	[2M]
g)	Define Hyper Text.	[2M]
h)	Define Cognition.	[2M]
i)	Give a note on concept of text search.	[2M]
j)	What is Imagery Retrieval?	[2M]

PART-B **(50 Marks)**

2.	Define Information retrieval system. Explain the objectives of the Information retrieval system.	[10M]
	OR	
3.	Explain the functional overview of Information Retrieval System.	[10M]
4.	What is stemming? Explain the porter stemming algorithm with example.	[10M]
	OR	
5.a)	Explain about Multimedia Indexing.	[5M]
b)	Explain the following data structures with examples. i) N-Gram ii) PAT	[5M]
6.	Differentiate between Manual Clustering and Automatic Term clustering. Explain with suitable examples.	[10M]
	OR	
7. a)	Explain the process of Thesaurus generations.	[5M]
b)	Define the Clustering. What are the guidelines for clustering?	[5M]
8.	Write short notes on the following with examples.	[10M]
a)	Similarity Measures	
b)	Ranking Algorithms	
	OR	
9.	Explain the main aspects of information visualization technology.	[10M]
10. a)	List out the differences between Boyer-Moore text search algorithm and Knuth-Morris algorithm.	[5M]
b)	Differentiate hardware versus software text search algorithms.	[5M]
	OR	
11.	Explain the GESCAN and Fast Data Finder hardware text search machines?	[10M]

****

Code No.: CS624PE | R20 | H.T.No. | | | 8 | R | | | | | |

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–II–Semester End Examinations (Regular) - May- 2023
INFORMATION RETRIEVAL SYSTEM
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks and may have a, b, c as sub questions.

PART-A **(20 Marks)**

1. a) Explain miscellaneous capabilities. [2M]
   b) Define the IRS. [2M]
   c) What is signature file structure? [2M]
   d) List Hidden markov models. [2M]
   e) Explain automatic indexing. [2M]
   f) Define Similarity measure. [2M]
   g) What is weighted search? [2M]
   h) Define cognition. [2M]
   i) What is software text search? [2M]
   j) Explain multimedia information retrieval. [2M]

PART-B **(50 Marks)**

2. Explain about search and browse capabilities. [10M]

OR

3. Discuss in detail about functional overview. [10M]

4. Write about objectives of indexing. [10M]

OR

5. Explain in detail Hypertext and XML data structures. [10M]

6. Write about Thesaurus generation with example. [10M]

OR

7. Explain Statistical indexing with example. [10M]

8. Illustrate about different Similarity measure and Ranking. [10M]

OR

9. Discuss about cognition and perception. [10M]

10. Explain about video retrieval in detail. [10M]

OR

11. Write about spoken language and audio retrieval. [10M]

****

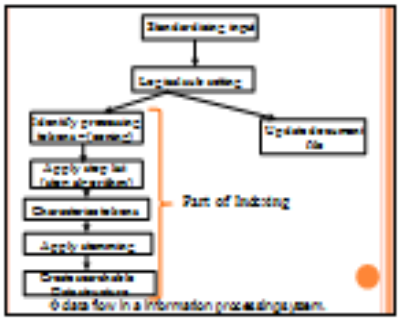
16.Power point presentations

UNIT-III

- Automatic Indexing
  - Classes of Automatic Indexing
    - Statistical Indexing
    - Natural Language
    - Concept Indexing
  - Hypertext Linkages

CLASSES OF AUTOMATIC INDEXING

- Automatic indexing is the process of analyzing an item to extract the information to be permanently kept in index.
- This process is associated with the generation of the searchable data structures associated with an item.
- The left hand side of the figure includes identifying processing tokens, apply stop list algorithm , characterize tokens, apply stemming and creating searchable data structure is part of indexing process.



6 data flow in a information processing system.

- All systems go through the initial stage of zoning and identify the processing tokens to create index.
- Some systems automatically divide the document up into fixed length passages or localities, which become the item unit that is indexed.
- Filters such as stop list algorithm and stemming algorithms -to reduce the processing tokens.
- An index is the data structure created to support search strategy.
- Search strategy – classified as statistical , natural language, and concept.

- Statistical strategy – covers the broadest range of indexing techniques and common in commercial systems .
- Basis for statistical is – frequency of occurrence of processing tokens(words/phrases) within documents and within data base.
- The words/phrases are the domain of searchable values.
- Statistics that are applied to the event data are probabilistic, Bayesian, vector spaces, neural net.

- Statistic approach – stores a single statistic, such as how often each word occurs in an item – used for generating relevance scores after a Boolean search.
- Probabilistic indexing – stores the information that are used in calculating a probability that a particular item satisfies (i.e., is relevant to) a particular query.
- Bayesian and vector approaches – store the information that are used in generating a relative confidence level of an item's relevancy to a query.
- Neural networks – dynamic learning structures – comes under concept indexing – that determine concept class.
- Natural language approach perform the similar processing token identification as in statistical techniques - additional level of parsing of the item (present, past, future action) enhance search precision.

- Concept indexing uses the words within an item to correlate to concepts discussed in the index item.
- When generating the concept classes automatically, there may not be a name applicable to the concept but just a statistical significance.
- Finally, a special class of indexing can be defined by creation of hypertext linkages.
- These linkages provide virtual threads of concepts between items versus directly defining the concept within an item.
- Each technique has its own strengths and weaknesses.

STATISTICAL INDEXING

- Uses frequency of occurrence of events to calculate the number to indicate potential relevance of an item.
- The documents are found by normal Boolean search and then statistical calculation are performed on the Hit file, ranking the output(eg. The term-frequency algorithm)

1. Probability weighting
2. Vector Weighting
   1. Simple Term Frequency algorithm
   2. Inverse Document Frequency algorithm
   3. Signal Weighting
   4. Discrimination Value
   5. Problems with the weighting schemes and vector model
3. Bayesian Model

---

1. PROBABILISTIC WEIGHTING

- Probabilistic systems attempt to calculate a probability value that should be invariant to both calculation method and text corpora(large collection of written/spoken texts).
- The probabilistic approach is based on direct application of theory of probability to information retrieval systems.
- Advantage: uses the probability theory to develop the algorithm.
- This allows easy integration of the final results when searches are performed across multiple databases and use different search algorithms.
- The use of probability theory is a natural choice – the basis of evidential reasoning (drawing conclusions from evidence).

- This is summarized by PRP(probability ranking principle) and Plausible corollary (reasonable result)
- PRP – hypothesis – if a reference retrieval system's response to each request is a ranking of the documents in order of decreasing probability of usefulness to the user, the overall effectiveness of the system to its users is best obtainable on the basis of the data available.
- Plausible corollary : the techniques for estimating the probabilities of usefulness for output ranking in IR is standard probability theory and statistics.
- probabilities are based on binary condition the item is relevant or not.
- IRS the relevance of an item is a continuous function from non- relevant to absolutely useful.
- Source of problems: in application of probability theory come from lack of accurate data and simplified assumptions that are applied to mathematical modeling.
- cause the results of probabilistic approaches in ranking items to be less accurate than other approaches.

- Advantage of probabilistic approach is that it can identify its weak assumptions and work to strengthen them.
- Ex : logistical regression
- Approach starts by defining a model 0 system.
- In retrieval system there exits a query q i and a document term d i which has a set of attributes (Vl…Vn) from the query (e.g., counts of term frequency in the query), from the document (e.g., counts of term frequency in the document) and from the database (e.g., total number of documents in the database divided by the number of documents indexed by the term).
- The logistic reference model uses a random sample of query-document term triples for which binary relevance judgments has been made from a sample.

17. Innovative teaching Methods

1.Take your own document calculate the term frequencies for the all unique terms.

2.take set of documents apply the term clustering method create the clusters.

18. Websites or URLs e- Resources

1) http://forum.jntuworld.com/showthread.php?22532-Information-Retrieval-Systems-%28IRS%29-Notes-Study-Materials
2) *www.old-site.clsp.jhu.edu/ws2005/calendar/documents/LavrenkoJuly5.**PPT***
3) www.iro.umontreal.ca/~nie/IFT6255/Introduction-IR.**ppt**
4) *www.comminfo.rutgers.edu/~tefko/Courses/610/Lectures/IntroIR.**ppt***

5) http://lecturesppt.blogspot.in/search/label/InformationRetrievalandDataMining
6) http://computer.howstuffworks.com/internet/basics/search-engine.htm
7) http://web.simmons.edu/~benoit/lis466/syllabus.html
8) https://support.google.com/websearch/answer/136861?hl=en
9) http://verdi.unisg.ch/www/edis.nsf/wwwDisplayIdentifier/3830/$FILE/dis3830.pdf

a. Journals

1.International Journal of Information Retrieval Research (IJIRR)

2.International Journal of Multimedia Information Retrieval

NATIONAL

1.Bates'BibliographyofWorks on Information Seeking, Indexing, and Information Retrieval System Design

2) International Journal of Information and Communication Technology Research