Welcome to the class of Advanced Topics in Information Retrieval !

Min ZHANG （张敏）
z-m@tsinghua.edu.cn

Course basic information

Instructor

Instructor: Min ZHANG （张敏）

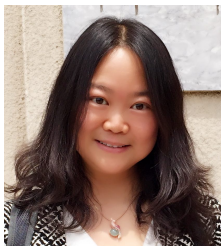
- Associate Professor
- IR group, DCST, Tsinghua Uni.
- Office: FIT 1-506A Tel: 62798279
- Email: z-m@tsinghua.edu.cn
- http://www.thuir.cn/group/~mzhang

TA: ZHANG Fan （张帆）

- Email: frankyzf94@gmail.com
- Tel: 18671829106
- Office: FIT 1-506

3

A little bit about Min…

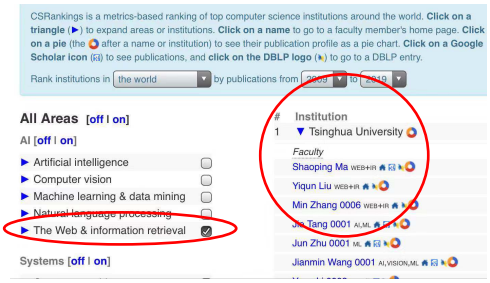


- Experiences
  - Dept. of CST, Tsinghua Unviersity
    - 1995~1999 (Bachelor), 1999~2003 (Ph.D.)
  - Associate Professor in THUIR group, CST Dept.
  - Visiting scholar/researcher in DFKI (Germany), Kyoto University, City University of HongKong, MSRA and NUS.
- Research Interests
  - Information Retrieval & Recommendation,
  - User modeling and behavior analysis,
  - Machine learning applications.

4

A little bit about Min…

- Selected Achievements and Awards
  - Published 100+ papers on important conferences & journals
    - ~3500 Citations(on Feb. 2019, Google Scholar)， H-index 32
  - Multiple Top Perform. in TREC and NTCIR since 2002.
  - One of the top ranked World-wide IR researchers
  - Awards:
    - Beijing Science & Technology Progress Award, 1st prize, 2016
    - State of Minist. of Radio, Film &Television, Guangxi Science &Technology, etc.
    - Excellent Young Faculty Teaching Award, Tsinghua Uni.



A little bit about Min…

- Academic Activities and Industry Connections
  - Vice Director of the state key lab of intelligent technology and systems (AI institute)
  - Vice Director of MOE MSRA Key Lab on Web and multimedia, Tsinghua University
  - Associate Editor of Transaction on Information Systems (TOIS) (CCF A journal) and reviewers for multiple top journals
  - Program Chairs: WSDM2019 workshop, SIGIR 2019 Tutorials, SIGIR2018 short papers, WSDM 2017, AIRS 2016
  - Area Chair, Senior PC or PC for
    - IJCAI, SIGIR, WWW, KDD, WSDM, CIKM, ACL……
  - Committee members of China Associations
    - Chinese Information Processing Committee (CCF)
    - Information Retrieval Committee (CIPS)
    - Machine Learning Committee (CAAI)

Course Arrangements

- Friday afternoon (13:30-14:15, 14:20-15:05), VI-6B206
- The first 13~14 weeks
  - Lectures by the teacher on selected topics
  - With "Tea Time" discussions on the news/progresses on Web and IR industry/research (by one student)
    - Send me the email by Wed. noon 11:59am to apply a tea time show on Thursday's course.
    - WolframAlpha, Adidas miCoach Dark net SE Man vs. Machine Facebook Scandal
- The last 2~3 weeks -- A workshop
  - Seminar presented by the students
  - 10~20 mins' talk + 5~10 mins' QA (tentative, depending on #students)
  - Awards and celebration to "the Best Project"

7

Part I-1. Lectures on Web IR Fundamentals (subject to modifications)

Week	Date	Topic	Content
1	3. 1	Introduction	Course intro, IR history
2	3. 8	IR key tech (I)	General procedure. Key tech: crawler , Index
3	3. 15	IR key tech (II)	Content-based Ranking Models: Term weighting; Boolean, VSM, probability(classical), LM
4	3. 22	IR key tech (III)	Evaluation: methodology; metrics (pre, recall, MAP, MRR, NDCG); Consistency analysis
5	3. 29	IR key tech (IV): Web Link Analysis	Link Analysis: Counting degree, HITS, PageRank, TrustRank, Spreading Activation, Anchor
6	4. 5	**Holiday**	
7	4. 12	Recommendation	Recommendation

8

Part I-2. Lectures on Advanced Topics (subject to modifications)

Week	Date	Topic	Content
8	4. 19	Advanced Topics (I)	User Behavior Analyses
9	4. 26	Advanced Topics (II)	Challenges: anti-spam
10	5. 3→5. 5 Sunday	Advanced Topics (III)	Challenges: Multi-source fusion, UI
11	5. 10	Advanced Topics (IV)	Challenges: Scale
12	5. 17	Evaluation Rethinking	What to evaluate, How to evaluate
13	5. 24	Social IR Or Invited Industry Talk	Social network & Human computation, Crowd computing
14	5. 31	Visual IR Or Invited Industry Talk	Visual IR

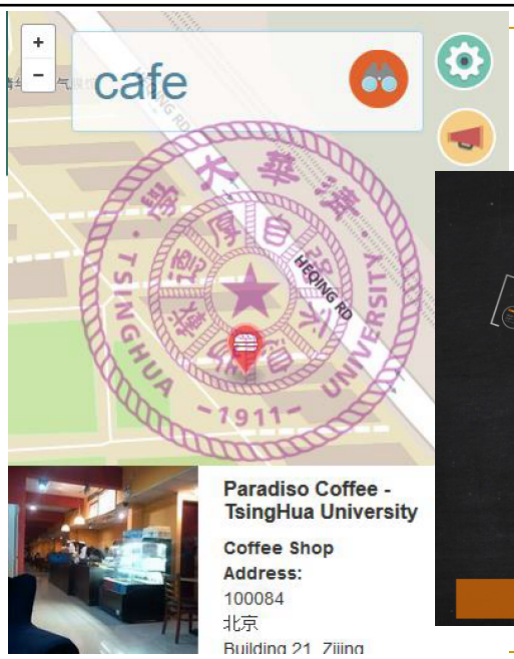
Part III. Project and Course workshop

- III-1. Design& implement a prototype Search Engine
  - Two students as a team
  - You can use general IR toolkits to build the prototype SE.
  - Both general and domain-specific SEs are acceptable.
    - A specific one with some interesting functions might be more attractive to your audiences
  - A demo system should be released online (accessible in THU net) at least 1 week before the presentation.
    - Final score including both pre-test and the live show.
- III-2. Submit a paper on this project (1 paper per student)

Week	Date	Topic	Content
15	**6.7→ 6.10 Mon**	Course Workshop	SE prototype design & implementation
16	6.14	Course Workshop	SE prototype design & implementation


Pokémon
Video games search engine
Search mode: Title only
Lam Corporation - June 2013 - Information - Suggestion

11

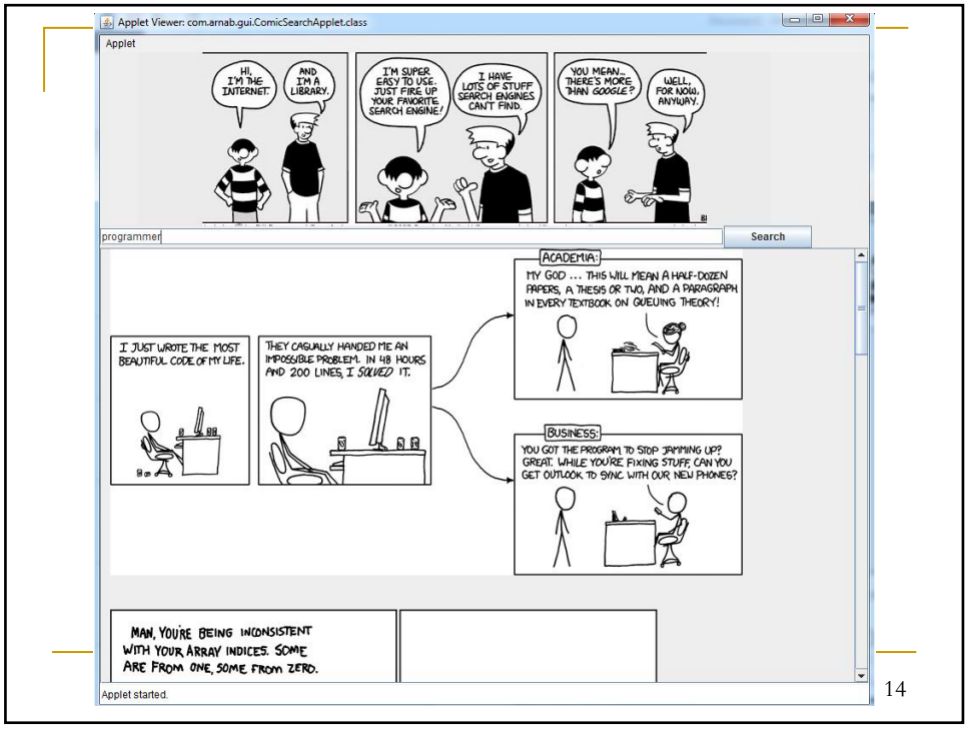

cafe
TSINGHUA UNIVERSITY
HEQING RD
-1911-
Paradiso Coffee - TsingHua University
Coffee Shop
Address:
100084
北京
Building 21, Zijing


Comic Searcher

12


Applet Viewer: com.arnab.gui.ComicSearchApplet.class
Applet
HI, I'M THE INTERNET.
AND I'M A LIBRARY.
I'M SUPER EASY TO USE. JUST FIRE UP YOUR FAVORITE SEARCH ENGINE!
I HAVE LOTS OF STUFF SEARCH ENGINES CAN'T FIND.
YOU MEAN... THERE'S MORE THAN GOOGLE?
WELL, FOR NOW, ANYWAY.
roses red
Search
WHIRRRRRR
hey, a package!
BOOM

13


Applet Viewer: com.arnab.gui.ComicSearchApplet.class
Applet
HI, I'M THE INTERNET.
AND I'M A LIBRARY.
I'M SUPER EASY TO USE. JUST FIRE UP YOUR FAVORITE SEARCH ENGINE!
I HAVE LOTS OF STUFF SEARCH ENGINES CAN'T FIND.
YOU MEAN... THERE'S MORE THAN GOOGLE?
WELL, FOR NOW, ANYWAY.
programmer
Search
I JUST WROTE THE MOST BEAUTIFUL CODE OF MY LIFE.
THEY CASUALLY HANDED ME AN IMPOSSIBLE PROBLEM. IN 48 HOURS AND 200 LINES, I SOLVED IT.
ACADEMIA:
MY GOD ... THIS WILL MEAN A HALF-DOZEN PAPERS, A THESIS OR TWO, AND A PARAGRAPH IN EVERY TEXTBOOK ON QUEUING THEORY!
BUSINESS:
YOU GOT THE PROGRAM TO STOP JAMMING UP? GREAT. WHILE YOU'RE FIXING STUFF, CAN YOU GET OUTLOOK TO SYNC WITH OUR NEW PHONES?
MAN, YOU'RE BEING INCONSISTENT WITH YOUR ARRAY INDICES. SOME ARE FROM ONE, SOME FROM ZERO.
Applet started.

14

Supports fuzzy search (maximum edit distance of 3)

LaTeX Symbol Search

esplon Search

Mouse over equations to display their TeX commands

Symbol in TeX	TeX Command
E, ϵ and ε	`E`, `\epsilon` and `\vareps`

Can search content

LaTeX Symbol Search

the set of all real numbers Search

Mouse over equations to display their TeX commands

Symbol in TeX	TeX Command	Name / Read as / Category	Explanation
$\mathbb{R}$ $\mathbf{R}$	`\mathbb{R}` `\mathbf{R}`	real numbers; R; the (set of) real numbers; the reals; numbers	$\mathbb{R}$ means the set of real numbers.

15

theamazingemojisearcher.herokuapp.com/?q=happy

theamazingemojisearcher.herokuapp.com/emoji/3

The Amazing Emoji Searcher

Home / face with tears of joy

Face with tears of joy

Description

A laughing emoji which at small sizes is often mistaken for being tears of sadness. In fact , this emoji is laughing so much that it is crying tears of joy. Tears are coming from both eyes, not due to sadness, but overwhelmed with happiness or laughter.

Related to

- Smiling Face With Open Mouth And Smiling Eyes
- Cat Face With Tears Of Joy
- Crying Face
- Loudly Crying Face
- Smiling Face With Open Mouth And Tightly-Closed Eyes

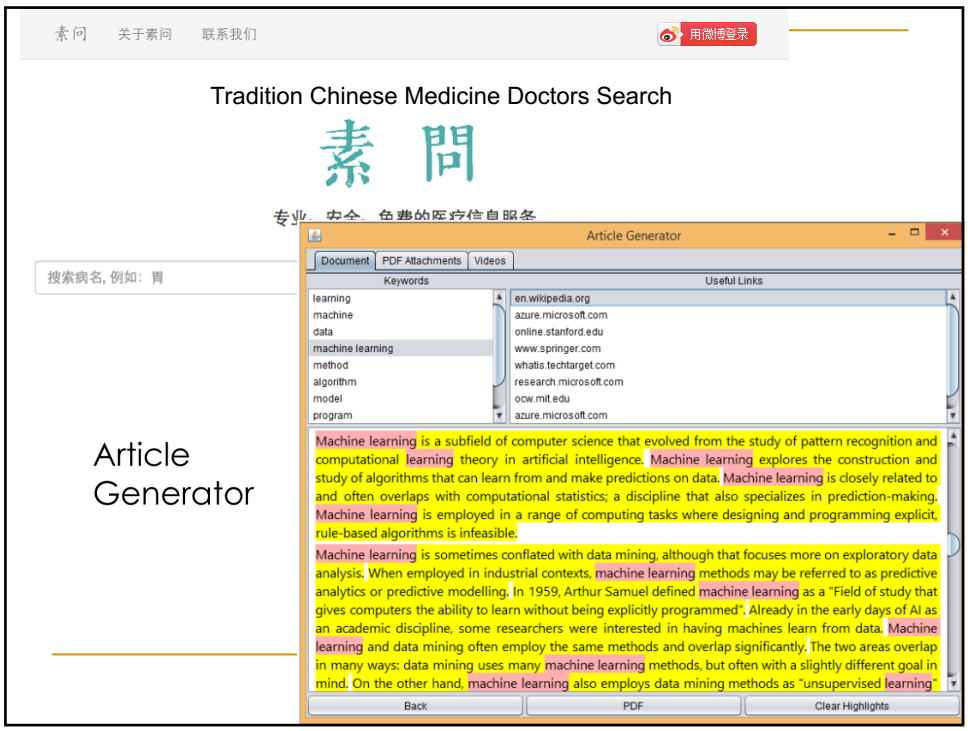
Also known as

- Laughing Emoji
- Laughing Crying Emoji
- Happy Tears Emoji
- Laughing Tears Emoji
- LOL Emoji

Different versions

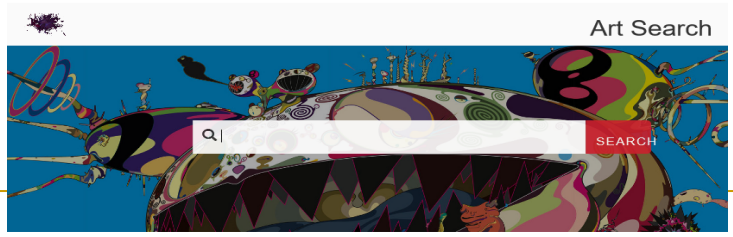
iPhone Android Twitter Gmail Windows Black and white

Made with love by Richard Luong. Available on Github



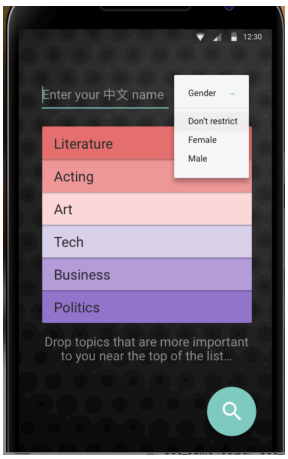
Article Generator

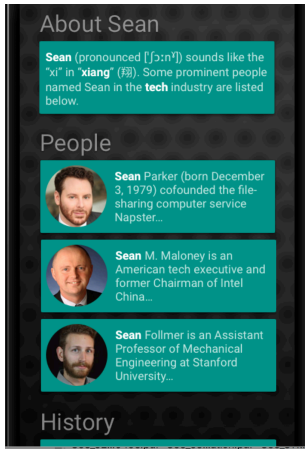
Web Art Search



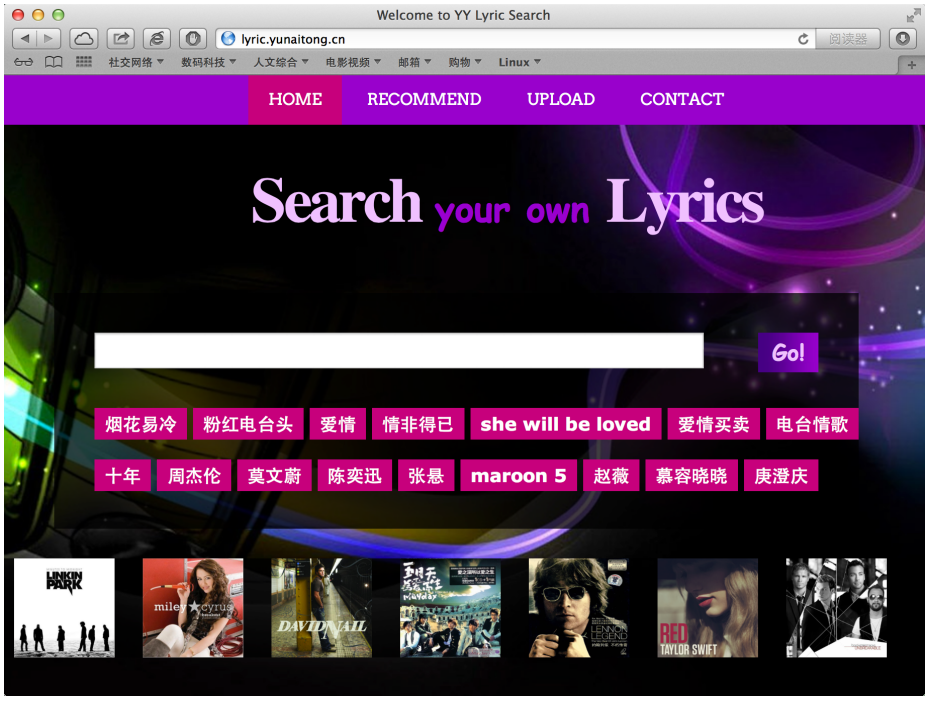
18

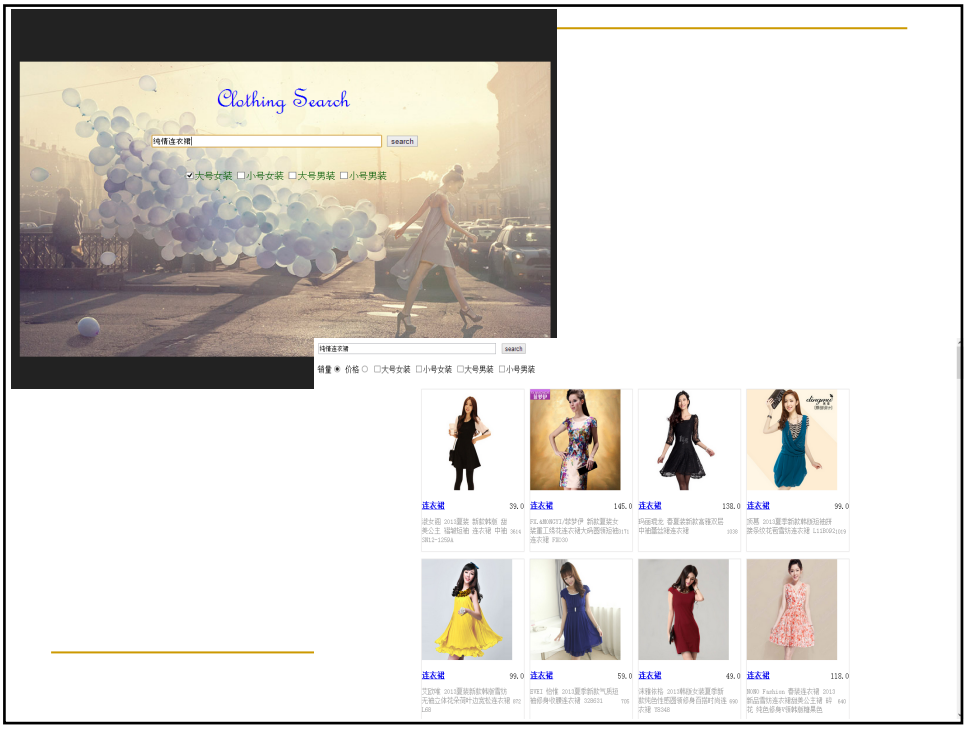
Bilingual Name Search Engine





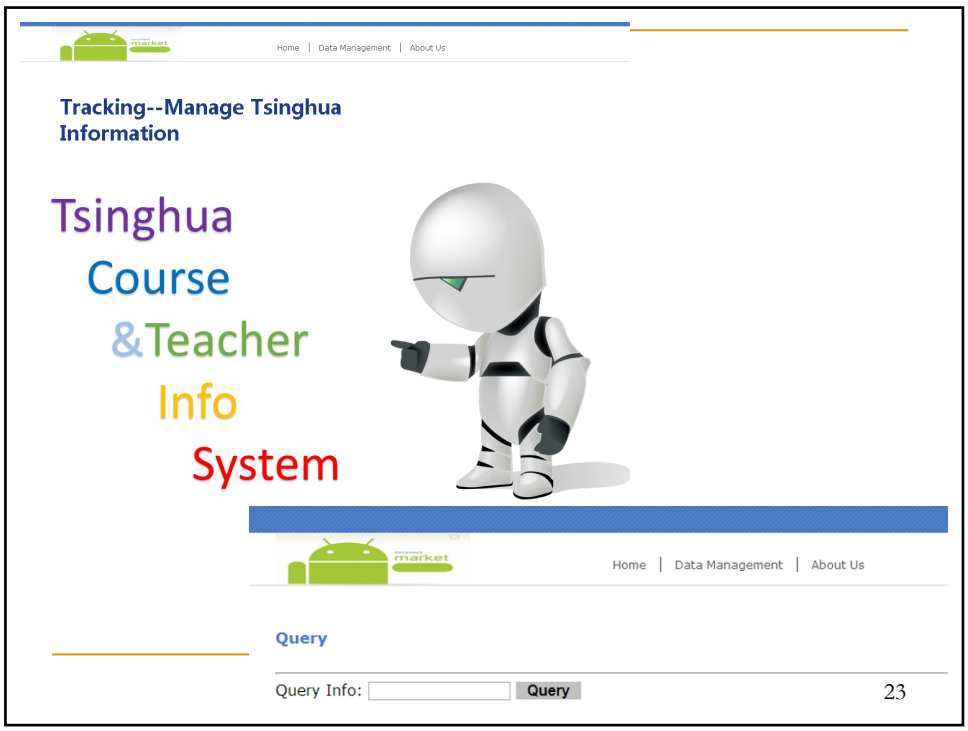
19




Clothing Search
search
大号女装 小号女装 大号男装 小号男装


CWePS
王伟
Search
Tag Cluster
Visualization
All Persons
王伟 0
曾国藩
京东 (6)
王伟 1
美国 (4)
王伟 2
中国科学院 (3)
王伟 3
昆山
财政部 (3)
王伟 4
鹿鼎记 (2)
王伟 5
海南
中华网 (2)
王伟 6
新安
中国共产党 (2)
This cluster with 4 documents
1 王伟 坠机前美国侦察机上拍摄的录像[视频]_历史频道_凤凰网
>2012年4月1日，我们可以看到，王伟驾驶着低速性能很差的歼-8II，接近飞行速度很慢的美军EP-3。，双方当时的距离十分近。
http://news.ifeng.com/history/zhongguoxiandaishi/detail_2012_04/01/1
2 王伟 _明星库_娱乐频道_凤凰网
>凤凰网娱乐频道艺人资料库为您提供最新、最全的 王伟 资料、新闻、照片、视频。
http://app.ent.ifeng.com/star/3730
9 海空卫士 王伟：尽职爱国的飞行员- 搜狐视频
> 1 Apr 2013
http://tv.sohu.com/20130402/n371449684.shtml
25 [视频]海空卫士 王伟："9时05分"至"9时07分"_新闻频道_央视网
>2013年4月1日，他的父母回忆说，王伟 想当飞行员"简直到了痴迷的程度"。他的老家湖州有个空军.机场，只要飞机从城市上空飞过，哪怕是吃饭，王伟 也要丢下碗筷，.
http://news.cntv.cn/2013/04/01/VIDE1364831560361674.shtml

22


Home | Data Management | About Us
Tracking--Manage Tsinghua Information
Tsinghua
Course
&Teacher
Info
System
Home | Data Management | About Us
Query
Query Info:
Query
23


Gonow
Gonow
世界公园
顶90%
10%踩
24

Evaluation (Subject to modifications)

- Homework (0%~15%)
- Workshop (55%~70%), evaluated by
  - the other students (30%)
  - + the teacher and TA (30%)
- Paper (~20%)
- QA activities (~10%)
  - Activities in QA sessions of seminars
  - Question answering in classes
- Bonus:
  - The best project on the workshop
  - Tea time presentations on weekly classes
  - *Active thinking and discussions are highly encouraged !*

25

What you will learn

- Overview and In-sight of Search (and Recommendation)Technologies
  - Beat 90% general SE users (but still not an expert, to be honest).
- Obtaining knowledge for (web) information processing and big data analysis (not only SE and IR), Such as
  - Evaluation strategies
    - Consistencies, correlation, and metrics (precision, recall, AUC, NDCG, ……)
  - Important classical and state-of-arts models
  - Better understanding on challenges
- Getting to know about new techniques in Web information services
  - Search, Recommender Systems, User modeling, Social analysis, Visual processing …
- Be used to always keep an eye on the new trends of Internet and AI

26

References

- We're not having an official textbook
  - There isn't one with good coverage of all & only the topics we'll discuss
  - A changing field, advanced topics
- A list of references:
  - Books
    - W. Bruce Croft, Donald Metzler, Trevor Strohman, **Search Engine: information retrieval in practice**
    - Christopher D. Manning, Prabhakar Raghavan ,Hinrich Schütze, **Introduction to information retrieval**
    - I. Witten, A. Moffat, and T. Bell, **Managing Gigabytes**
  - Proceedings of Conferences
    - SIGIR, WWW, IJCAI, WSDM, CIKM, TREC, NTCIR ...
  - Very important: Web resources, Search engines

27

What's IR?

Figure Copyright by TREC

What is Information Retrieval (IR)?

- Narrow-sense:
  - IR = Search Engine Technologies (i.e. IR =
    - Google, Yahoo, Bing, Ask, Baidu, Sogou, …
    - Library info search, enterprise search, in-site search, destop search…
    - PicSearch, Greplin, Blekko, SkyScanner, KooXoo, Qunar, …

29

What's IR? (cont.)

- Broad-sense: IR ~ **Information Management**
  - General problem: how to manage information?
  - How to **find** useful information? (retrieval & recommendation)
    - Beyond search engine:
    - e.g. in news feed, movie, travel, e-commerce, financial… scenarios
    - e.g. in social media platform, e.g. Twitter, Facebook, YouTube, WeChat, Weibo, Zhihu, ……
  - How to **organize** information? (classification & filtering)
    - e.g., automatically assign email to different folders
  - How to **discover** information (or even knowledge) from the data? (mining)
    - e.g., discover correlation of events

30

What's IR? (cont.)

- Goal:
  - Find documents *relevant* to an information need from a large document set
- And now:
  - Beyond relevance
  - Multi-modal documents
  - Users' (implicit) information need
  - Heterogeneous environment

Figure Copyright by TREC

31

IR is Hard!

- Under/over-specified query
  - Ambiguous: "buying CDs" (certificate deposit? or compact disc?)
  - Incomplete: what kind of CDs?
  - What if "CD" is never mentioned in document?
- Vague semantics of documents
  - Ambiguity: word-sense, structural
    - e.g. "bank"
  - Incomplete: Inferences required
    - E.g. "windows" "apple"
- A difficult task even for human beings!
  - Only 80% agreement in human judgments

32

IR is "Easy"!

- IR CAN be easy in a particular case
  - Ambiguity in query/document is RELATIVE to the database
  - So, if the query is SPECIFIC enough, just one keyword may get all the relevant documents
- PERCEIVED IR performance is usually better than the actual performance
  - Users can NOT judge the completeness of an answer
  - E.g. Web Search vs. Machine Translation