

Information Retrieval

Weike Pan

The slides are **adapted from those provided by Prof. Hinrich Schütze** at University of Munich (<http://www.cis.lmu.de/~hs/teach/14s/ir/>).

Chapter 8 Evaluation in information retrieval

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.1 Information retrieval system evaluation

Measures for a search engine

- How **fast** does it **index**
 - e.g., number of bytes per hour
- How **fast** does it **search**
 - e.g., latency (延迟) as a function of queries per second
- What is the **cost** per **query**?
 - in dollars

8.1 Information retrieval system evaluation

- All of the preceding criteria are **measurable** (可以衡量的): we can quantify the speed/size/money
- However, the key measure for a search engine is **user happiness**
 - **Speed** of response
 - **Size** of index
 - Uncluttered (整洁的) **UI**
 - Most important: **relevance** (相关性)
 - Actually, maybe even more important: **it's free**

8.1 Information retrieval system evaluation

Who is the user? (1/2)

- Who is the user we are trying to make happy?
- Web search engine: **searcher** (搜索引擎用户). Success: Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: **advertiser** (广告商). Success: Searcher clicks on advertisement. Measure: click-through rate (CTR).

8.1 Information retrieval system evaluation

Who is the user? (2/2)

- Ecommerce: **buyer** (购买者). Success: Buyer buys something. Measures: time to purchase, fraction of “conversions” (转化率) of searchers to buyers.
- Ecommerce: **seller** (销售者). Success: Seller sells something. Measure: profit per item sold.
- Enterprise: **CEO**. Success: Employees are more productive because of effective search. Measure: profit of the company.

8.1 Information retrieval system evaluation

Most common definition of user happiness: **Relevance**

- User happiness is equated with (等同于) the **relevance** of search results to the query.
- But how do you measure the relevance?
- Standard methodology in information retrieval consists of three elements.
 - A **benchmark** document collection
 - A **benchmark** suite of queries
 - An **assessment** of the relevance of each query-document pair

8.1 Information retrieval system evaluation

Relevance: query vs. information need (1/2)

- **Information need i**: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- **Query q**: [red wine white wine heart attack]
- **Consider document d'**: At the heart of his speech was an attack on the wine industry lobby (游说团) for downplaying the role of red and white wine in drunk driving.
- d' is **an excellent match** for query q ...
- d' is **not relevant** to the information need i

8.1 Information retrieval system evaluation

Relevance: query vs. information need (2/2)

- User happiness can only be measured by **relevance to an information need**, not by relevance to queries.
 - → query intent classification
- Our terminology is sloppy (草率的) in the textbook: we talk about **query-document** relevance judgments even though we mean **information-need-document** relevance judgments.

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.2 Standard test collections

What we need for a benchmark

- **A collection of documents**: Documents should be **representative** of the documents we expect to see in reality.
- **A collection of information needs** (often incorrectly called **queries**): Information needs should be **representative** of the information needs we expect to see in reality.
- **Human relevance assessments**: We need to hire/pay “judges” or assessors to do this
 - Expensive, time consuming
 - Judges should be **representative** of the users we expect to see in reality

8.2 Standard test collections

First standard relevance benchmark: Cranfield

- Pioneering: first **testbed** allowing precise quantitative measures of information retrieval effectiveness
 - Late 1950s, UK
 - 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all (query, document) pairs
 - Too small, too untypical for serious IR evaluation today

8.2 Standard test collections

Second-generation relevance benchmark: TREC

- **TREC** = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST, 美国国家标准与技术研究院)
- TREC is actually a set of several different relevance benchmarks.

8.2 Standard test collections

Information Retrieval Benchmark

- <http://www.bigdatalab.ac.cn/benchmark/bm/Domain?domain=Information%20Retrieval>
- <https://www.microsoft.com/en-us/research/publication/letor-benchmark-collection-research-learning-rank-information-retrieval/>

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.3 Evaluation of unranked retrieval sets

Precision and recall (1/2)

- **Precision(精确率)** is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- **Recall(召回率)** is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

8.3 Evaluation of unranked retrieval sets

Precision and recall (2/2)

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$\underline{P = TP / (TP + FP)}$$

$$\underline{R = TP / (TP + FN)}$$

8.3 Evaluation of unranked retrieval sets

Precision/recall tradeoff

- You can increase recall by returning more documents.
- Recall is a **non-decreasing function** of the number of docs retrieved.
- A system that returns all documents has 100% recall.
- Similarly, it is usually easy to get high precision for very low recall.

8.3 Evaluation of unranked retrieval sets

A combined measure: F

- F allows us to tradeoff precision against recall.

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1 - \alpha}{\alpha}$$

$\alpha \in [0, 1]$ and thus $\beta^2 \in [0, \infty]$

- Most frequently used: balanced F with $\beta = 1$ or $\alpha = 0.5$
 - This is the harmonic mean of P and R:

$$\frac{1}{F} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{R} \right)$$

8.3 Evaluation of unranked retrieval sets

Example for precision, recall, F1

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- $P = 20 / (20 + 40) = 1/3$
- $R = 20 / (20 + 60) = 1/4$
- $F_1 = 2 \frac{1}{\frac{1}{3} + \frac{1}{4}} = 2/7$

8.3 Evaluation of unranked retrieval sets

Accuracy (准确率)

- Why do we use complex measures like precision, recall, and F?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table (列联表) in the previous page, $\text{accuracy} = (TP + TN) / (TP + FP + FN + TN)$.

8.3 Evaluation of unranked retrieval sets

Exercise

- Compute precision, recall, F1 and accuracy for this result set:

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

8.3 Evaluation of unranked retrieval sets

Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
 - You then get **99.99% accuracy** on most queries.
- Searchers on the web (and in IR in general) **want to find something** and have a certain **tolerance** for junk.
- It's better to return some bad hits **as long as you return something**.
 - We use precision, recall, and F for evaluation, not accuracy.

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

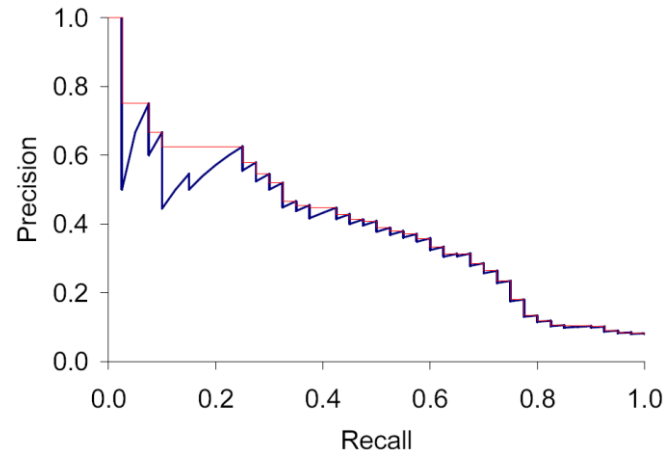
8.4 Evaluation of ranked retrieval results

Precision-recall curve

- Precision/recall/F1 are measures for **unranked sets**.
- We can easily turn set measures into measures of ranked lists.
 - Just compute the set measure for **each “prefix”**: the top 1, top 2, top 3, top 4 ... results, e.g., Precision@k and Recall@k, $k=1,2,3,\dots$
 - Doing this for precision and recall gives you a precision-recall curve.

8.4 Evaluation of ranked retrieval results

A precision-recall curve



- Each point (in blue) corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \dots$). Precision@ k usually decreases with larger k ; Recall@ k usually increases with larger k .
- Interpolation (插值, in red): Take maximum of all future points.
- Why use interpolation: the area under the red curve better represents the overall ranking performance than the area under the blue curve, because a user is usually willing to look at more stuff if both precision and recall get better.

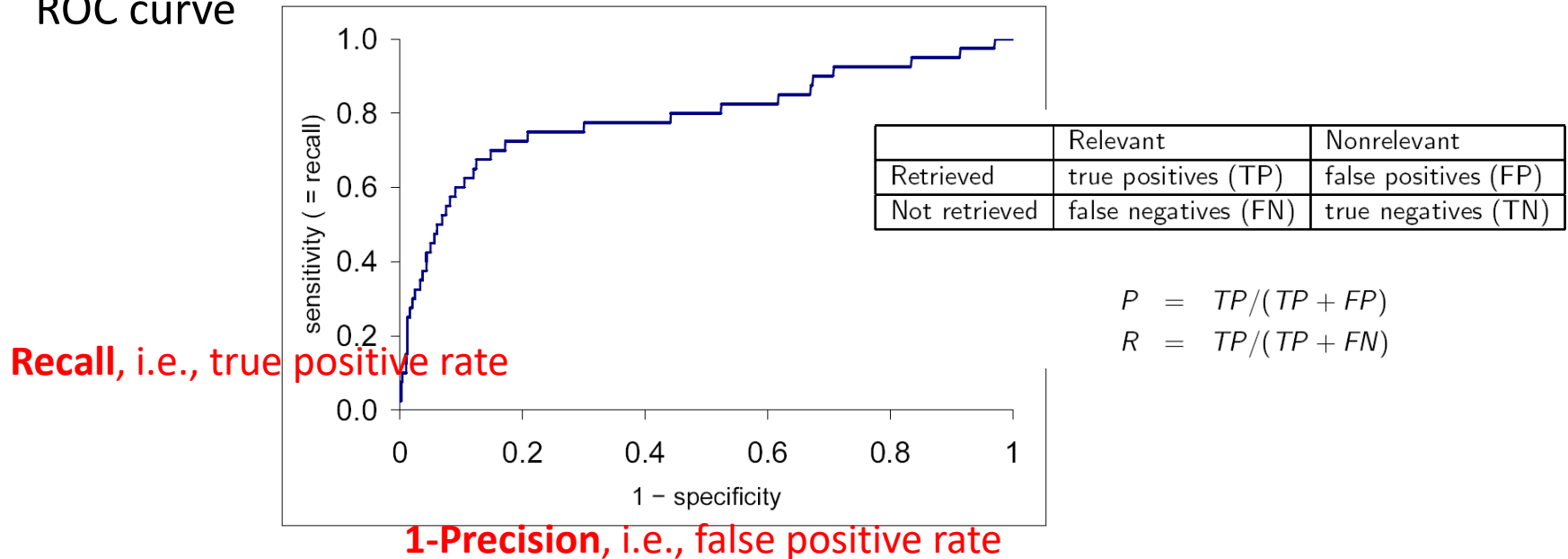
8.4 Evaluation of ranked retrieval results

11-point interpolated average precision

	Recall	Interpolated Precision
	0.0	1.00
• 11-point average: 0.425	0.1	0.67
	0.2	0.63
	0.3	0.55
• This measure measures performance at <u>all recall levels</u> .	0.4	0.45
	0.5	0.41
	0.6	0.36
	0.7	0.29
	0.8	0.13
	0.9	0.10
	1.0	0.08

8.4 Evaluation of ranked retrieval results

ROC curve



- For the ROC curve (receiver operating characteristic curve), we are only interested **in the small area in the lower left corner** (because when 1-Precision=1, i.e., Precision=0, we can always have Recall=1)

8.4 Evaluation of ranked retrieval results

Variance of measures like precision/recall

- For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).
- Indeed, it is usually the case that the variance of the same system across queries is much larger than the variance of different systems on the same query.
- That is, there are easy information needs and hard information needs.

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.5 Assessing relevance

Kappa measure (1/2)

- Relevance assessments by two judges are usable if they are **consistent**.
- How can we measure this **consistency** between two judges?
→ **Kappa measure**, how much two judges agree or disagree

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- $P(A)$ = proportion of time that two judges agree, i.e., **observed agreement**
- $P(E)$ = what agreement would we get by chance, i.e., **expected agreement**

8.5 Assessing relevance

Kappa measure (2/2)

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Observed agreement

$$P(A) = (300 + 70)/400 = 370/400 = 0.925$$

Expected agreement

$$P(E) = (80/400) \times (90/400) + (320/400) \times (310/400) = 0.665$$

Kappa statistic

$$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$$

Values in the interval $[2/3, 1.0]$ are seen as **acceptable**.

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.6 A broader perspective: System quality and user utility

Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use **precision at top k**, e.g., $k = 10$...
- Search engines also use non-relevance-based measures.
 - E.g., **click-through rate (CTR) on first result**

8.6 A broader perspective: System quality and user utility

A/B testing

- Purpose: **Test a single innovation**
- Prerequisite: You have a large search engine up and running.
 - Have most users use the **old system**
 - Divert (转移) **a small proportion of traffic** (e.g., 1%) to the new system that includes the innovation
 - Evaluate with an “automatic” measure like CTR on first result
- **Probably the evaluation methodology that large search engines trust most**

8.6 A broader perspective: System quality and user utility

Marginal relevance

- We've defined relevance for an isolated (query, document) pair.
- Alternative definition: **marginal relevance**
- The **marginal relevance** of the document d_k at position k in the result list is the **additional information it contributes** over and above the information that was contained in documents d_1, \dots, d_{k-1} .

Outline

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading

8.7 Results snippets

How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
 - This description is crucial.
 - The user can often identify good hits (= relevant hits) based on the description.
 - No need to actually view any document.

8.7 Results snippets

Document description in result list

- Most commonly: doc title, URL, some metadata ... and a summary
- How do we “compute” the summary?

8.7 Results snippets

Summaries

- Two basic kinds: (i) static (ii) dynamic
- A **static summary** of a document is always the same, regardless of the query that was issued by the user.
- **Dynamic summaries are query dependent**. They attempt to explain why the document was retrieved for the query at hand.

8.7 Results snippets

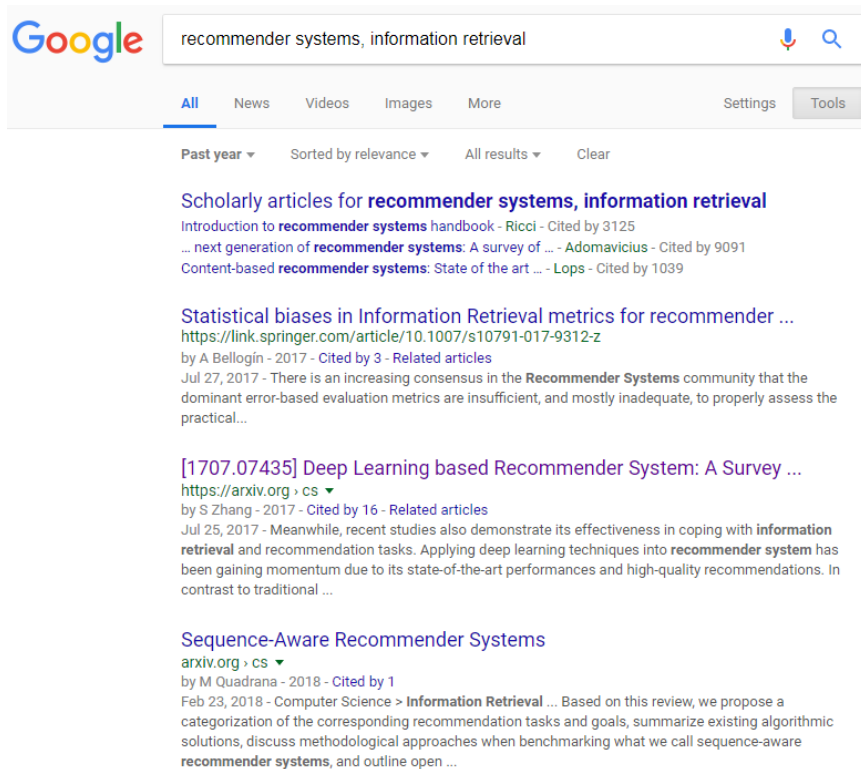
Static summaries

- In typical systems, the static summary is a subset of the document.
 - Simplest heuristic: **the first 50 or so words** of the document
 - More sophisticated: extract from each document **a set of “key” sentences**
 - Simple NLP heuristics to score each sentence
 - Summary is made up of top-scoring sentences.
 - ...
- Most sophisticated: complex **NLP** to **synthesize/generate** a summary
 - For most IR applications: not quite ready for prime time yet

8.7 Results snippets

Dynamic summaries

- Present one or more “windows” or **snippets** within the document that contain several of the query terms.



Google search results for the query "recommender systems, information retrieval". The search bar shows the query and a microphone icon. Below the search bar are tabs for All, News, Videos, Images, and More. The results are sorted by relevance. The first result is "Scholarly articles for recommender systems, information retrieval" with a snippet about a handbook. The second result is "Statistical biases in Information Retrieval metrics for recommender ..." with a snippet about a survey. The third result is "[1707.07435] Deep Learning based Recommender System: A Survey ..." with a snippet about a survey. The fourth result is "Sequence-Aware Recommender Systems" with a snippet about a review.

recommender systems, information retrieval

All News Videos Images More Settings Tools

Past year Sorted by relevance All results Clear

Scholarly articles for **recommender systems, information retrieval**
Introduction to **recommender systems** handbook - Ricci - Cited by 3125
... next generation of **recommender systems**: A survey of ... - Adomavicius - Cited by 9091
Content-based **recommender systems**: State of the art ... - Lops - Cited by 1039

Statistical biases in Information Retrieval metrics for recommender ...
<https://link.springer.com/article/10.1007/s10791-017-9312-z>
by A Bellogin - 2017 - Cited by 3 - Related articles
Jul 27, 2017 - There is an increasing consensus in the **Recommender Systems** community that the dominant error-based evaluation metrics are insufficient, and mostly inadequate, to properly assess the practical...

[1707.07435] Deep Learning based Recommender System: A Survey ...
<https://arxiv.org> - CS
by S Zhang - 2017 - Cited by 16 - Related articles
Jul 25, 2017 - Meanwhile, recent studies also demonstrate its effectiveness in coping with **information retrieval** and recommendation tasks. Applying deep learning techniques into **recommender system** has been gaining momentum due to its state-of-the-art performances and high-quality recommendations. In contrast to traditional ...

Sequence-Aware Recommender Systems
arxiv.org - CS
by M Quadrana - 2018 - Cited by 1
Feb 23, 2018 - Computer Science > **Information Retrieval** ... Based on this review, we propose a categorization of the corresponding recommendation tasks and goals, summarize existing algorithmic solutions, discuss methodological approaches when benchmarking what we call sequence-aware **recommender systems**, and outline open ...



Baidu search results for the query "推荐系统 信息检索". The search bar shows the query and a camera icon. Below the search bar are tabs for 网页, 新闻, 贴吧, 知道, 音乐, 图片, 视频, 地图, 文库, and 更多. The results are sorted by relevance. The first result is "全文检索系统-大数据基础服务-定制解决方案" with a snippet about a solution. The second result is "搜索推荐系统是如何实现的?-知乎" with a snippet about a question. The third result is "第一篇:你不一定了解的'推荐系统'-穆晨-博客园" with a snippet about a blog post. The fourth result is "推荐系统属于自然语言处理范畴还是信息检索范畴-百度知道" with a snippet about a question. The fifth result is "推荐系统 基于内容的推荐-CSDN博客" with a snippet about a blog post.

推荐系统 信息检索

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多

百度为您找到相关结果约551,000个 搜索工具

全文检索系统-大数据基础服务-定制解决方案
阿里云和Elastic联合发布,基于全文检索系统及商业版X-pack插件深度优化,为大数据实时搜索和分析量身打造,提供企业级权限管控,安全监控,自动报表等功能。
www.aliyun.com 2018-04 - 587条评价 - 广告

搜索推荐系统是如何实现的?-知乎
2017年12月5日 - 在作为互联网产品的连接器:连接人、信息、服务之间的桥梁,搜索和推荐有其各自的特点。它们都是数据挖掘技术、信息检索技术、计算统计学等悠久学科的智...
<https://www.zhihu.com/question...> - 百度快照

第一篇:你不一定了解的"推荐系统"-穆晨-博客园
2017年6月12日 - 为了解决信息过载问题,截至至今,人们经历了分类目录、搜索引擎、推荐系统三个阶段。下文将以信息检索为例,说明人们检索方式(或者可以理解为解决信息过载方式)的演化...
<https://www.cnblogs.com/muchen...> - 百度快照

推荐系统属于自然语言处理范畴还是信息检索范畴-百度知道
2个回答 - 最新回答: 2013年09月06日 - 7人觉得有用
最佳答案: 自然语言处理中有关语言本身的研究,信息检索都不研究 详细一点来说,分词、语法、句法、词义等,信息检索都不会涉及 自然语言主要是偏研究的,重点是建模...
更多关于推荐系统 信息检索的问题>>
<https://zhidao.baidu.com/quest...> - 百度快照

推荐系统 基于内容的推荐-CSDN博客
2016年5月26日 - 基于内容的推荐 基于内容的信息推荐的理论依据主要来自于信息检索和信息过滤,所谓的基于内容的推荐方法就是根据用户过去的浏览记录来想用户推荐用...
<https://blog.csdn.net/u0112639...> - 百度快照

8.7 Results snippets

Generating dynamic summaries

- Where do we get these other terms in the snippet from?
 - We **cannot construct a dynamic summary from the positional inverted index** – at least not efficiently.
 - We need to cache documents.
 - Note that the cached copy can be outdated
 - Don't cache very long documents – just cache a short prefix

Summary

- 8.1 Information retrieval system evaluation
- 8.2 Standard test collections
- 8.3 Evaluation of unranked retrieval sets
- 8.4 Evaluation of ranked retrieval results
- 8.5 Assessing relevance
- 8.6 A broader perspective: System quality and user utility
- 8.7 Results snippets
- 8.8 References and further reading