Answering Student Programming Questions using Domain-specific Search

Kelvin Meyer and Hussein Suleman

Digital Libraries Laboratory
Department of Computer Science @ School of IT
University of Cape Town

July 2020

Problem 1 – Within year repetition

Adam Gabriels : Hi Stephan, please could the missing solutions to the June exams be uploaded, by Monday morning, at least. That will have given most a decent time frame to attempt the questions.

May 30 17:33 Reply + − 12 Likers

Michelle Kuttel : Adam - this query was answered by Maria below.

Ivan Asem : I still don't have my theory test mark

May 3 16:36 Reply + − 4 Likers

Michelle Kuttel : See comment below by Yamiko: " You should receive your result by end of day Monday"

May 3 18:08 + − 1 Liker

Lebohang Mbele : Monday has past and I still don't have my theory results madam.

May 5 02:27 + − 1 Liker

Tayla Thomas : Is it possible for more past Test 3's to be uploaded please?

May 4 09:19 Reply + − 10 Likers

Michelle Kuttel : See reply above

May 5 10:02 + −

Problem 2 – Across year repetition

Storm Johnson : Are the marks that are recorded for our assignments the highest, or are they the average?
Apr 13 16:26 Reply + −

Stephan Jamieson : If you mean the marks associated with multiple submissions of an assignment, it's the highest that's taken.
Apr 13 16:27 + −

Una Singo : May someone please help my understand why we should describe a function's purpose with a docstring when we could use "#"notes?
May 26 12:40 Reply + − 1 Disliker

Hussein Suleman : docstrings are in a standard format so everyone knows where to look and what to look for. by extension, the interpreter can also associate each docstring with the correct function.
standardization is the key thing.
May 27 09:39 + − 6 Likers

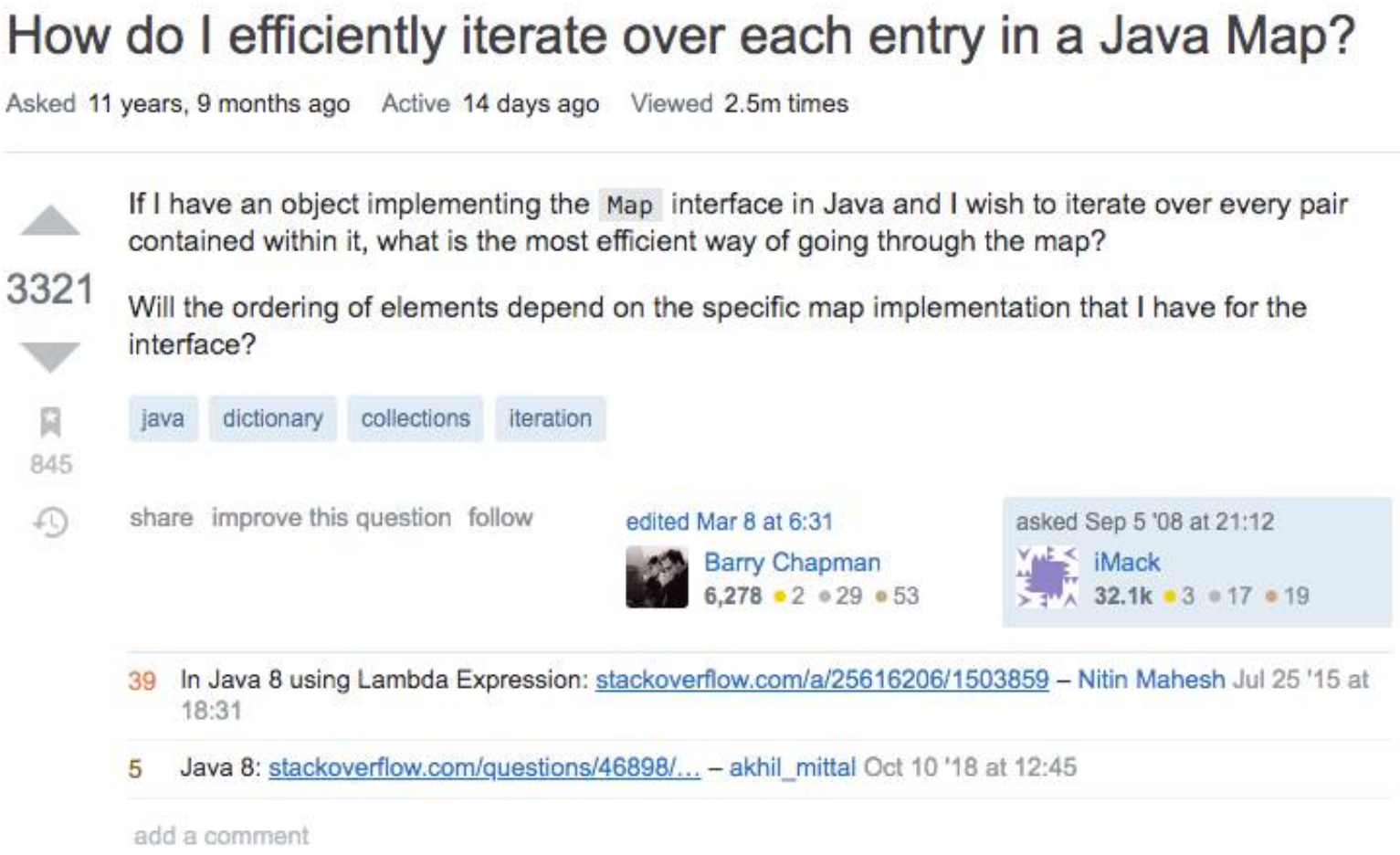
Xihluke Nkuna : what does it mean when your program works correctly but when you test it with large numbers it says aborted, it to A marker it says killed...palindromeprime
May 2 01:55 Reply + −

Michelle Kuttel : It's taking too long...
May 2 08:19 + −

Xihluke Nkuna : What's the solution?
May 2 10:56 + − 1 Liker

Michelle Kuttel : Are you sure that you don't have infinite recursion?
May 2 21:58 + − 1 Liker

Problem 3 - References



How do I efficiently iterate over each entry in a Java Map?

Asked 11 years, 9 months ago Active 14 days ago Viewed 2.5m times

3321

If I have an object implementing the `Map` interface in Java and I wish to iterate over every pair contained within it, what is the most efficient way of going through the map?

Will the ordering of elements depend on the specific map implementation that I have for the interface?

java dictionary collections iteration

845

share improve this question follow

edited Mar 8 at 6:31
Barry Chapman
6,278 • 2 • 29 • 53

asked Sep 5 '08 at 21:12
iMack
32.1k • 3 • 17 • 19

39 In Java 8 using Lambda Expression: stackoverflow.com/a/25616206/1503859 – Nitin Mahesh Jul 25 '15 at 18:31

5 Java 8: stackoverflow.com/questions/46898/… – akhil_mittal Oct 10 '18 at 12:45

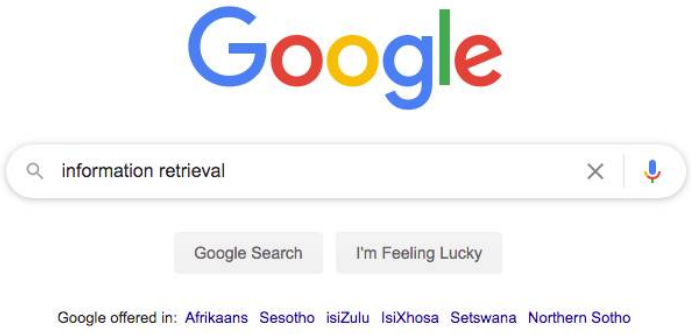
add a comment

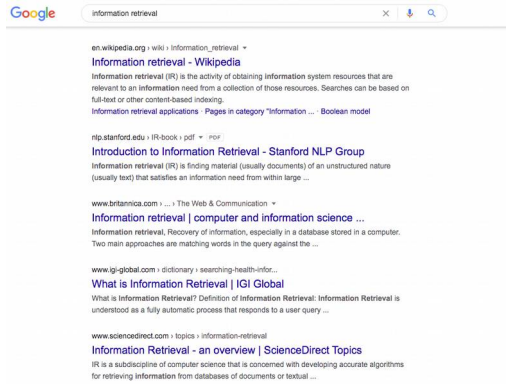
Key Research Question

Can we use
Information Retrieval
to answer
student programming questions?

Information Retrieval (IR)

- Sub-discipline of Computer Science concerned with providing users with **relevant** information to meet their **information needs**.
- Search Engines are one example.





But isn't this just a Database?

- No.
- IR systems can search through structured (e.g., databases), semi-structured (e.g., XML), and unstructured (e.g., unformatted plain text) data.
- Queries are often imprecise.
- Results most likely to be relevant are higher up.

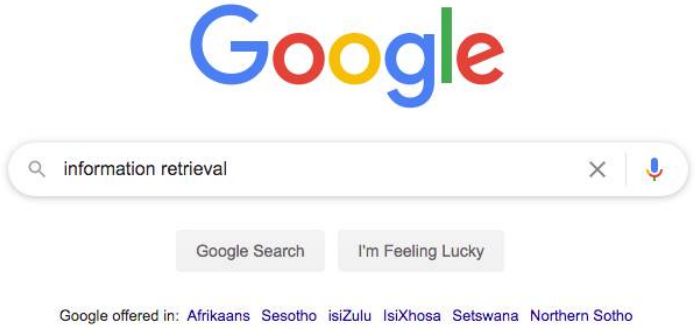
How do IR algorithms work?

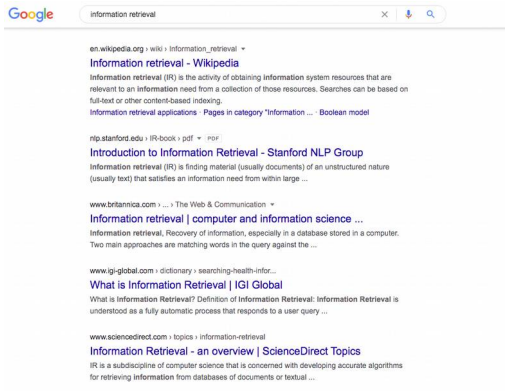
- **Filtering**
  - Removing those results that do not match.
    - Use inverted files for very efficient filtering.
  - E.g., All those documents that do not contain the words “information” or “retrieval”
- **Ranking**
  - Sorting the results by probable relevance.
  - Sort by **aboutness** – documents more about “information retrieval” must be higher up.

Ranked Retrieval

- Ranking of results is based on combining factors such as:
  - Term Frequency – how many times a search term appears in a document
  - Inverse Document Frequency – how rare a search term is
  - Document length
  - PageRank – how authoritative a page is
  - Document source
  - …

Web Search to Students Questions





Storm Johnson : Are the marks that are recorded for our assignments the highest, or are they the average?

If you mean the marks associated with multiple submissions of an assignment, it's the highest that's taken.

Domain-Specific Search

- Is there domain-specific information that can improve ranking?
  - i.e., provide better results?

- We tested authority of posters.
  - If an answer came from a staff member, it will be ranked more highly (AUTH).

    vs.
  - It did not matter who answered (NORM).

Experimental Setup

- **Apache SOLR** was used as IR toolkit.
- All threaded discussions from 2013 course used as document collection.
- Stopwords (common words) were ignored.
- 32 questions were selected from 2014 course to serve as queries.
- Top 30 results were retrieved from SOLR for each query for NORM and for AUTH.

Relevance Judgements

- For each query:
  - There were 30 results for NORM and 30 for AUTH.
  - Duplicates in the list were removed.
  - 5 users provided relevance judgements for each result (1=irrelevant, 5=highly relevant).

Metrics for IR

- **Recall@N** – what percentage of the relevant documents were retrieved in the top N results.
- **Precision@N** – what percentage of the top N results are relevant.
- **Mean Average Precision (MAP)** – Average of precision over different N values.
- **Normalised Discounted Cumulative Gain (NDCG)** – sum of rank-moderated matching scores, adjusted relative to ideal ranking.

Results

	P@3	R@3	P@5	R@5
NORM	0.3021	0.4732	0.3	0.5427
AUTH	0.3021	0.4911	0.3	0.5617

	NORM	AUTH
MAP	0.3793	0.3882
NDCG@3	0.5955	0.6034
NDCG@5	0.6139	0.6211
NDCG@30	0.8292	0.834

Discussion

- IR can provide students with answers in programming forums.
- @3 and @5 is because results could be presented like Google's query suggestions.
- Authority adds a small increase in performance.
  - What about other factors? Student participation level? Thread structure?

Integration in practice

- In all forums and chat rooms:
  - When a student asks a question:
    - The question is sent to the IR system as a query.
    - Ranked results are displayed to the student.
    - If the student finds the answer, they are done.
      - No repeated question asked.
    - If there is no answer, the new question can be posted.

Conclusions

- IR techniques in discussion forums can:
  - Give students immediate answers.
  - Reduce work for staff.
  - Create a knowledge-base across multiple years.
  - Exploit domain-specific features for better results.
- We need to use statistical techniques and information-centric approaches for smarter teaching.

questions, comments, ...

More information, papers, etc. on
http://dl.cs.uct.ac.za/

enkosi
hamba kakuhle
thank you and go well