https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

Indian Institute of Information Technology, Sri City

Characteristic vectors representing code are often high dimensional, *e.g.*, there are more than two hundreds different syntactic element types in Java ASTs..

– Lee et al. From their paper, "Instant Code Clone Search".

Project Evaluation Plan

- Submit the following to google classroom. Deadline: Dec 9th, 9 PM.
  - A 2 to 4 page technical report on your project. Four samples are shared with you.
  - A 15-minute presentation (10 to 15 minutes talk + 5 to 10 minutes Q & A).
- Give working demonstration during the project evaluation days. Detailed roll-no-wise schedule will be published closer to the day of the evaluation. Check your almanac for project evaluation days.
- In this presentation and the report, you will be evaluated for the following:
  - Your general understanding of IR concepts.
  - Demonstration of in-depth knowledge in at least one IR concept/idea.
  - Application of at least one IR concept/idea through your project.
  - A working system.
  - How is your system useful?
- Note that you may have to explain your understanding to me as well as to a different faculty member who may not know Information Retrieval.

Exercise

- For a query
  - q0: IR in IIITS
- Assume the following are relevant documents:
  - d1: IIITS is great for IR
  - d2: I study IR at IIITS
  - d3: IR is great
- Assume **it**, **is, are, for, at, to** and **in** are stopwords.
- Assume the following are non-relevant documents:
  - d4: IIITS competes in ICPC
  - d5: It is great to play
- Apply Rocchio method to reformulate the query. Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

Quiz

- In the previous example, what happens when $\alpha = 0$, $\beta = 1$, $\gamma = 1$ *?*
- In the previous example, what happens when $\alpha = 1$, $\beta = 0$, $\gamma = 0$ *?*
- In the previous example, what happens when $\alpha = 0$, $\beta = 0$, $\gamma = 0$ *?*
- In the previous example, what happens when $\alpha = 1$, $\beta = 2$, $\gamma = 2$ *?*

Pseudo (Blind) Relevance Feedback

- No User Judgment.
- Assume that the top-k ranked documents are relevant.

Initial query = “cheap CDs cheap DVDs extremely cheap CDs”.
d_1 = “CDs cheap software cheap CDs”.
d_2 = “cheap thrills DVDs”.
What would the revised query vector be after pseudo relevance feedback if top-1 document is considered as relevant?
Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

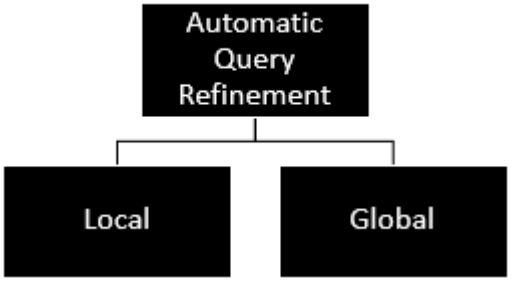
- May lead to query drift.

Indirect (Implicit) Relevance Feedback

- No asking for judgments from users.
- No automatic feedback such as assuming top-k documents as relevant.

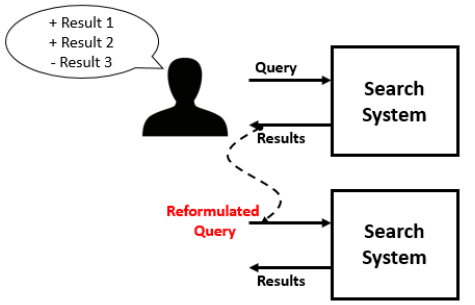
Clickstream Mining

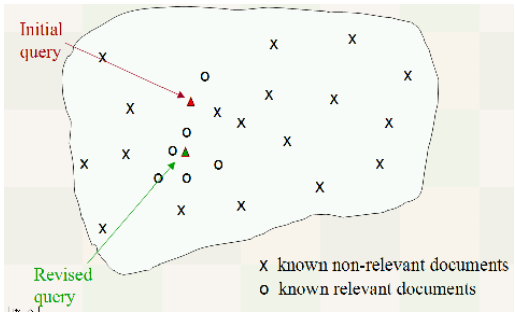
Recap



Dealing with Local Query Refinement

Relevance Feedback





	cheap	CDs	DVDs	extremely	software	thrills
q_0	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1
$q_m = q_0 + 0.75*d_1 - 0.25*d_2$						
q_m	4.25	3.5	0.75	1	0.75	0

Negative weight does not make sense. So, leave them as zero.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Pseudo (Blind) Relevance Feedback
Assume top-k is relevant.

Indirect (Implicit) Relevance Feedback
Using clickstreams, query logs, etc.

Global (User/Result-Independent) Query Refinement

- Automatic Thesaurus Generation
  - Fast = rapid
  - Tall = height?
  - Sound = noise?
  - Restaurant = Hotel = Motel?
- How to handle domain specific phrases?
- Slangs!
- …

How to automate the thesaurus generation?

Co-occurrence Analysis

MAINFRAMES

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of

MAINFRAMES

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers

Source: https://web.stanford.edu/~jurafsky

Co-occurrence Analysis

- Term-Document Matrix
  - How often does individual terms appear in a document?
- Term-Term Matrix
  - How often terms co-occur?

Quiz: Which books are similar?

	Book1	Book2	Book3	Book4
cricket	400	10	355	3
football	5	5	4	4
hockey	9	330	10	200
tennis	2	6	12	4

Co-occurrence Analysis

- Two documents are similar if the document vectors are similar.

	Book1	Book2	Book3	Book4
cricket	400	10	355	3
football	5	5	4	4
hockey	9	330	10	200
tennis	2	6	12	4

Book1 and Book3 seem to be on cricket.
Book2 and Book4 are about hockey.

Co-occurrence Analysis

- Two terms are similar if the term vectors are similar.

	Book1	Book2	Book3	Book4
boundary	400	310	355	389
four	515	225	390	400
movie	9	4	8	1
film	2	6	9	2

Remember, context is important!

Magic with Matrices

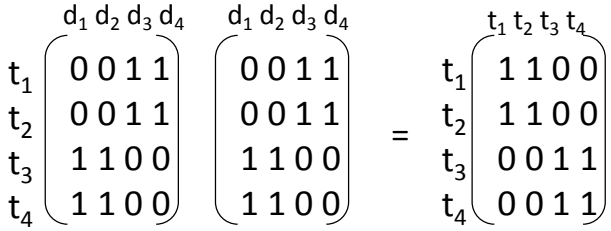
Transpose

- If A is as given below, what is A^T?

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \qquad \mathbf{A^T} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

Co-occurrence Analysis

- Assume a Boolean term-document matrix A.
- What does AA^T mean?



- Usually, weighted length-normalized tf in a sliding window is used to count co-occurrence.

Thank You