Multimedia oral examination - Miscellaneous Techniques behind IR

1.Page Rank

(1)Why pagerank? the motivation?

Retrieval results based on term frequence are not satisfactory. Because it is a tough issue and scalability is a big burden.

Users are not able to express what they want by keywords only and keywords are informative which means different for different people. Keywords are very few and many pages share similar similarity score.

(2)What is pagerank

Pagerank use hyper-links to rerank and improve the search quality. The idea of pagerank is that pages being linked to by other pages should be important and ranked higher. So PageRank assigns a weight to each document, which is defined recursively and depends on the number and weight of pages link to it. A page that is linked to by many pages with high PageRank receives a high rank itself.

(3)How to builg a pagerank model.

Pagerank is produced based on an adjacent matrix of web pages. First,Initialize a pagerank for each page based on uniform distribution,then update pagerank for each page iteratively until convergence.

The PageRank value is updated by the sum of the weights assigned from pages linked to it. The weight of each page that can be assigned is its pagerank divided by the number of pages it points to. convergence condition maybe the magnitude of the update is less than a threshold.

(4)Damping factor.

PageRank's formula can be optimized by a damping factor。It allows pages have zero inward links have non-zero pagerank values and do not affect the pagerank values of other wedpages.

Physical meaning of the damping factor is the probability that user will continue clicking on links during surfing.

(5)Tricks to promote your webpage based on Page Rank.

Based on the idea of Page Rank,the rank of a page can be promoted by askind webpages that has higher pagerank to link to it. Search and click-in website with Google from different places is also helpful.

Google rocbot can fix these tricks. For example,ignore hyperlinks that share the same color with the background

---

2.Evaluation on IR performance

IR performance can be evaluated by some metrics,such as TP,FP,TN,FN,recall,Precision,F-measure,Average Precision,mean Average Precision and so on.

TP refers to True Positive,which means the number of relevant documents retrieved. FN refers to False Negative,which means the number of relevant documents missed. FP refers to False Positive,which means the number of irrelevant documents retrieved. TN refers to True Negative which means the number of irrelevant documents not retrieved. Recall is the fraction of retrieved relevant documents over the total amount of relevant documents,caculated by TP divided by the total amount of ground truth.

Precision is the fraction of retrieved relevant documents over the retrieved documents,caculated by TP divided by K in top-K evalution.

F1 score is the harmonic average of the precision and recall.The reciprocal of F1 is equal to the sum of the reciprocals of recall and precision.

In practice, users are more sensitive to precision and have no knowledge about recall

AP and mAP are more suitable.

Mean average precision for a set of queries is the mean of the average precision scores for each query.

---

3. web crawler

(1)What is web crawler and how to do it?

Crawler is used for updating web content. It starts with a list queue of URL seeds ,repeat visiting,downing,parsing and adding new hyperlinks to the list until reaches a terminal condition.

(2)Some duties of web crawler.

Cache DNS record to map between URL and IP address. Parsing web pages to extract semantic meaning. Handle exceptions such as 404 Error.

(3)Some strategies of web crawler.

hottest sites should be crawled frequently and allocate more computing resources. Typical strategies are Breadth-first search ,Performance and Quality based scheduling.