EIE3280 Project

Course Content Search Engine

Zeng Zhuoru 119010417
Chen Xuanwen 120090582
Zhuang Yan 120090303
Peng Yiwei 119010245

Contents

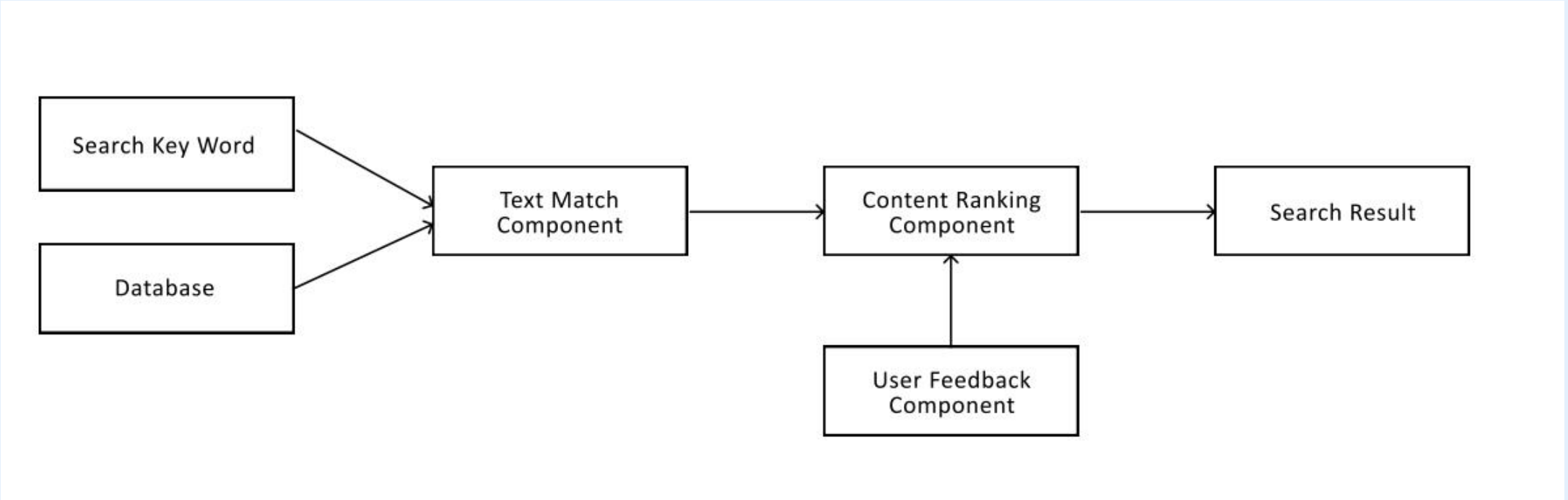
Introduction

- Motivation: To help students review the courses better.
- Background: There is an classial educational search engine designed by Kamal EL GUEMMAT, which can translate the key words into SPARQL queries. And use SPARQL to rank data.
- Challange: This project is a new scene of the educational search engine.
- Contribute: Build the database of EIE3280, from chapter 1 to chapter 8;

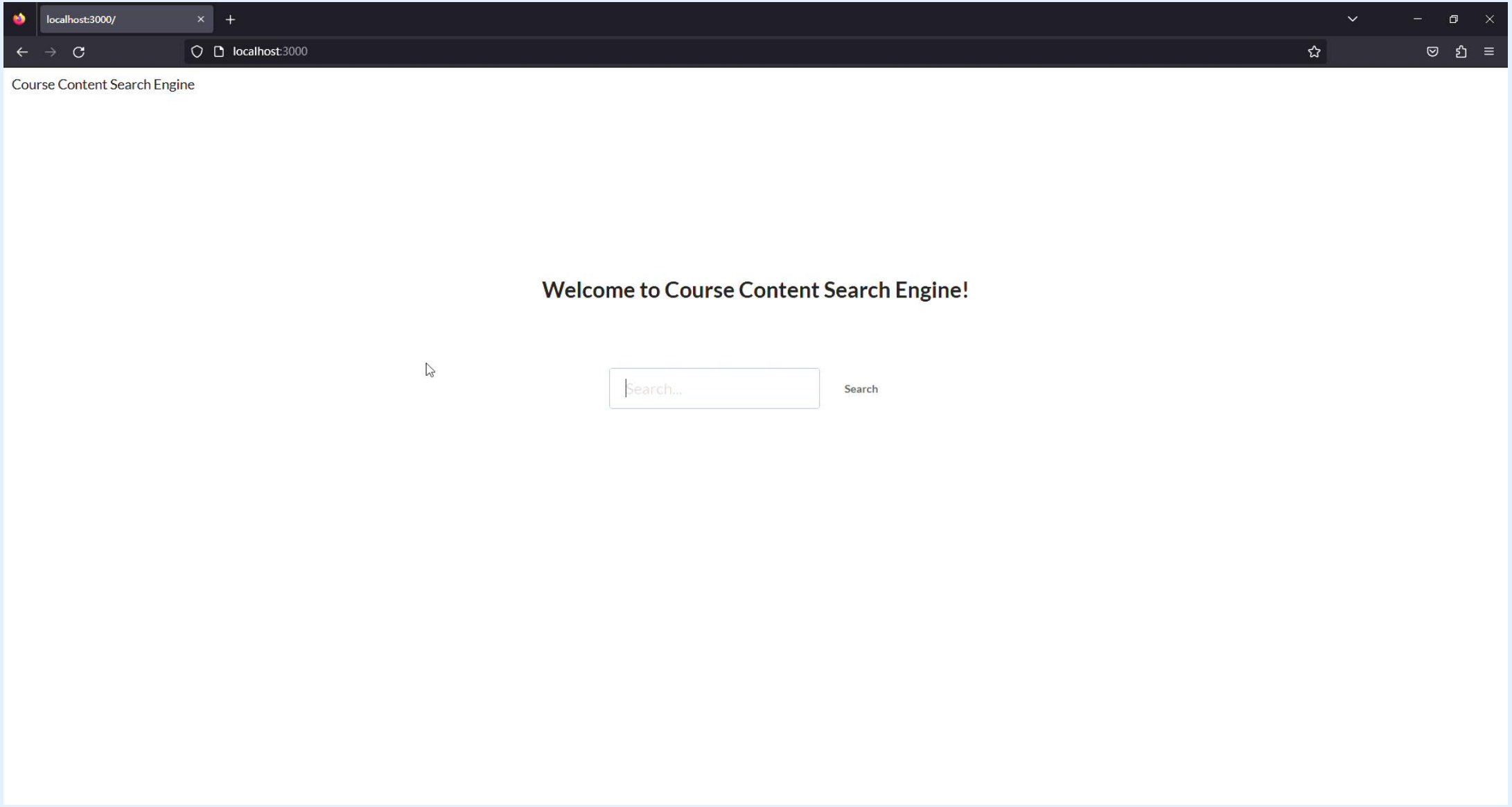
  The application of the importance score, relevance score;

  Build website applications.

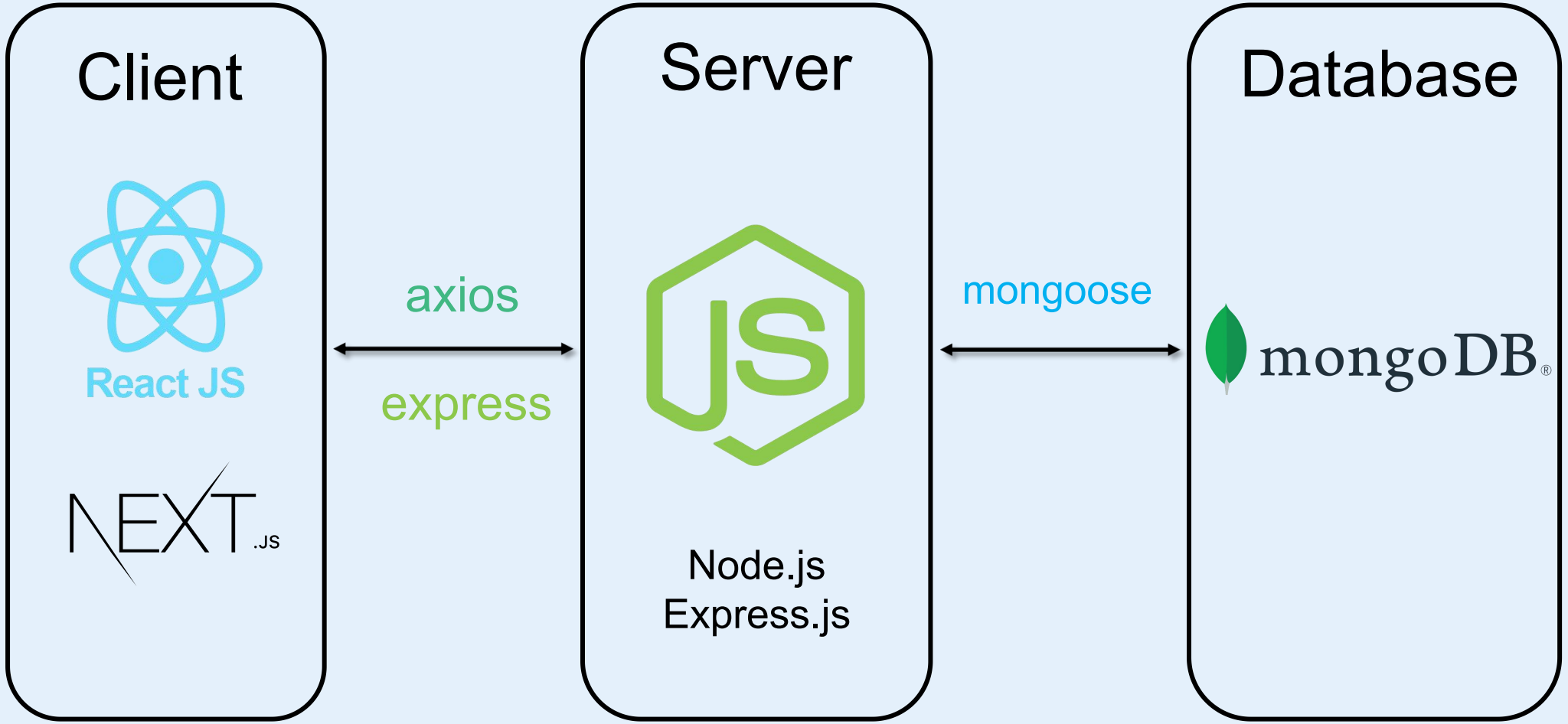
Components of the system



Example Video



Website App Implementation



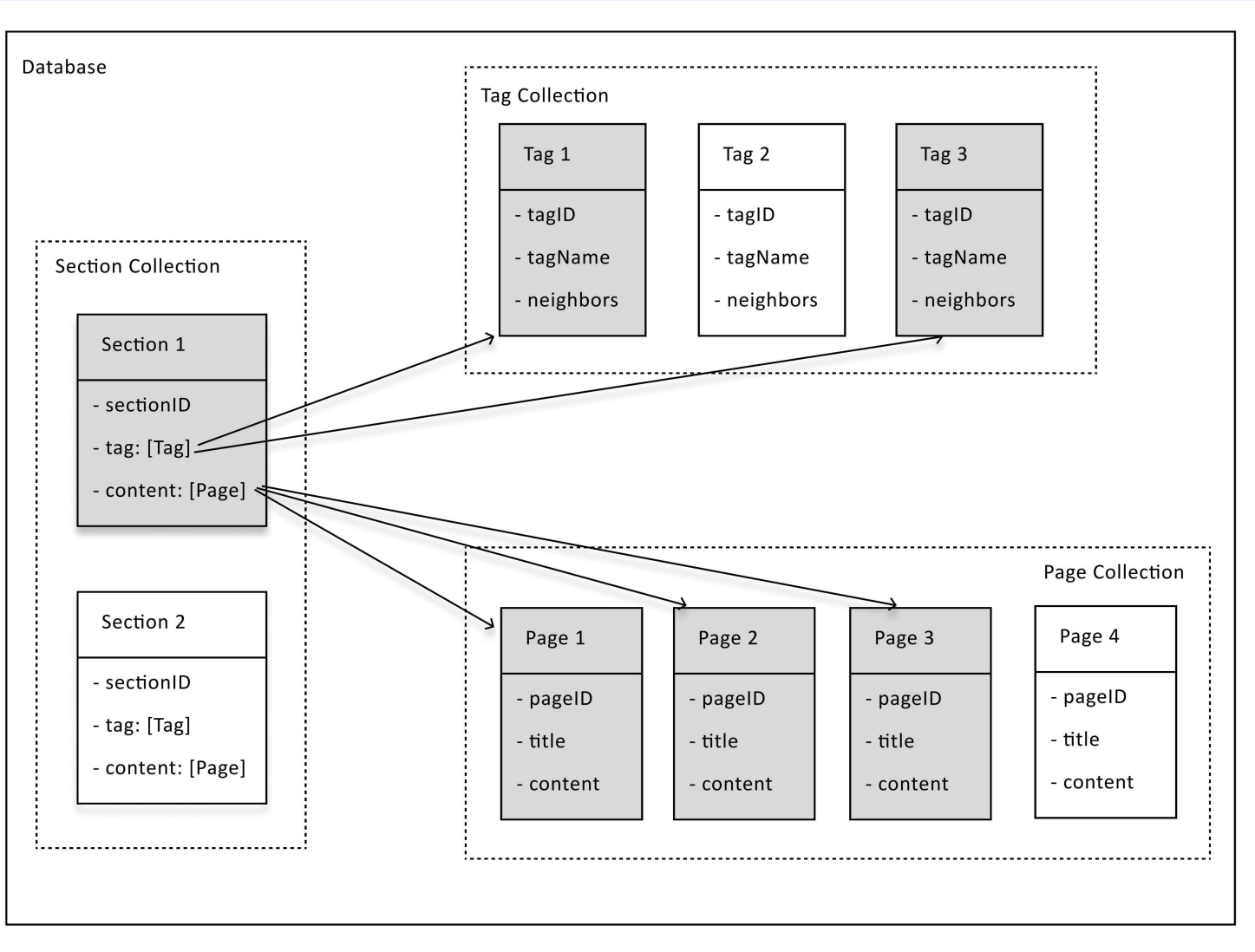
Database

Data Source

- Text data
- From EIE3280 Chapter 1-8 Course Slides

Database Structure

- Pages
- Sections
- Tags



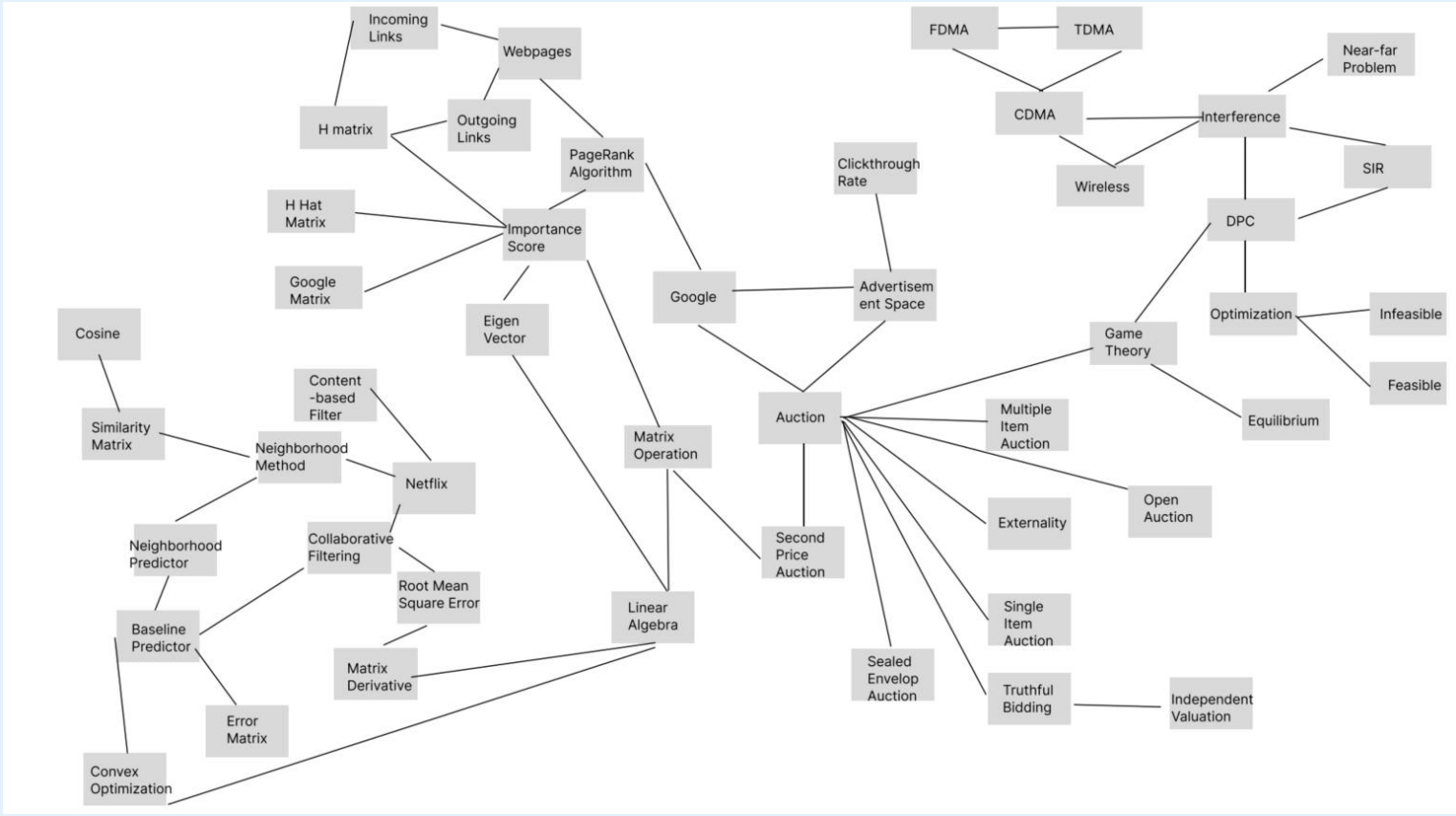
Graph of Tags

Graph Built by Tags

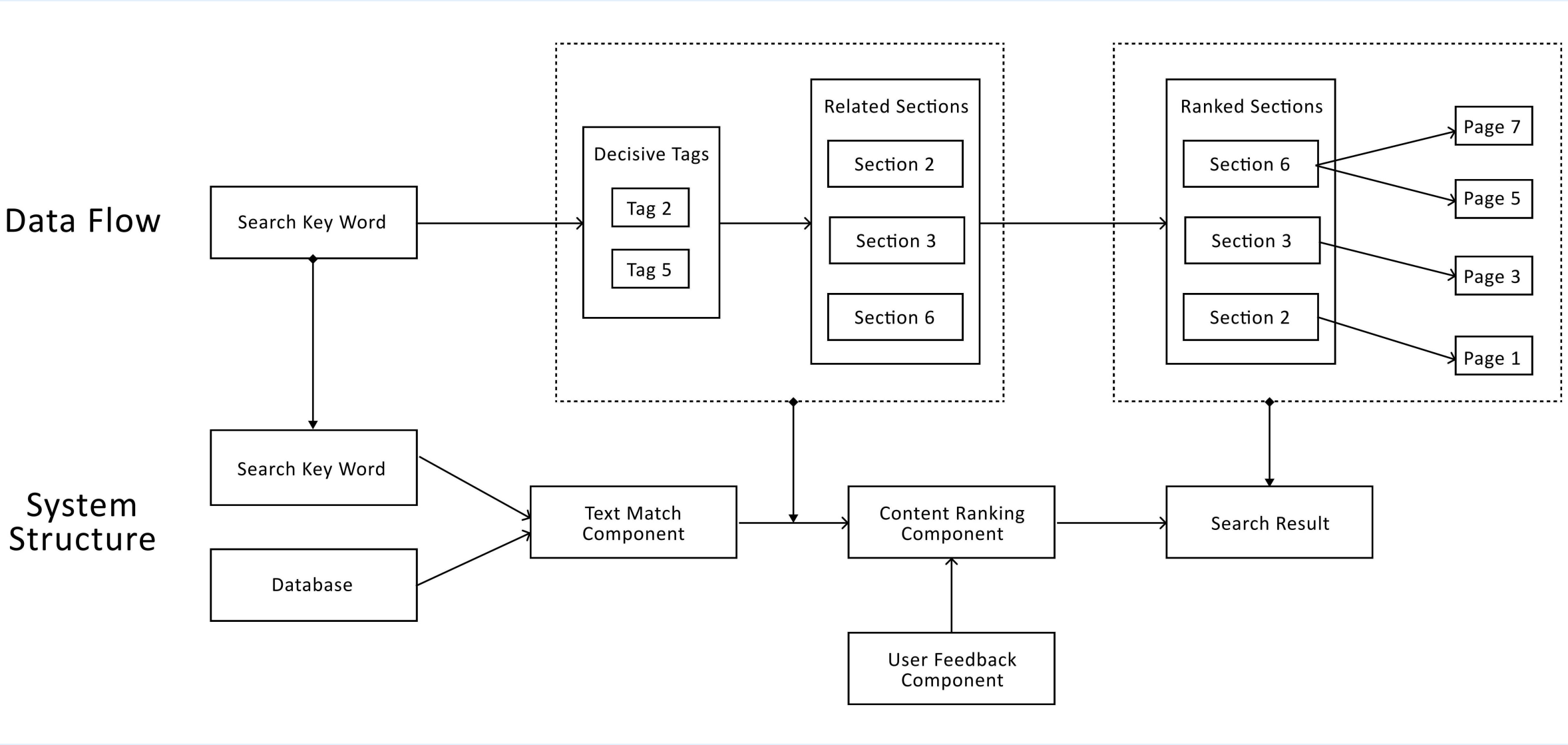
- Each tag is a node
- Relationship between tags are links
- Undirected graph

Graph Abstraction

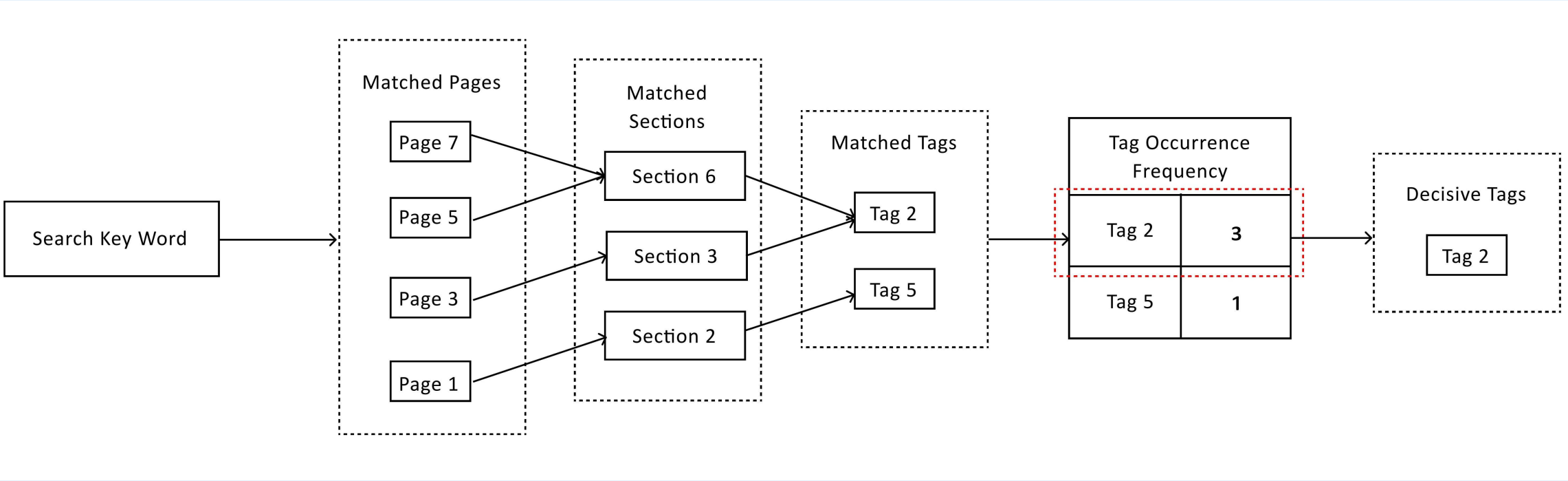
- Links between tags indicate relationships between pages / sections



Decisive Tags as a Medium

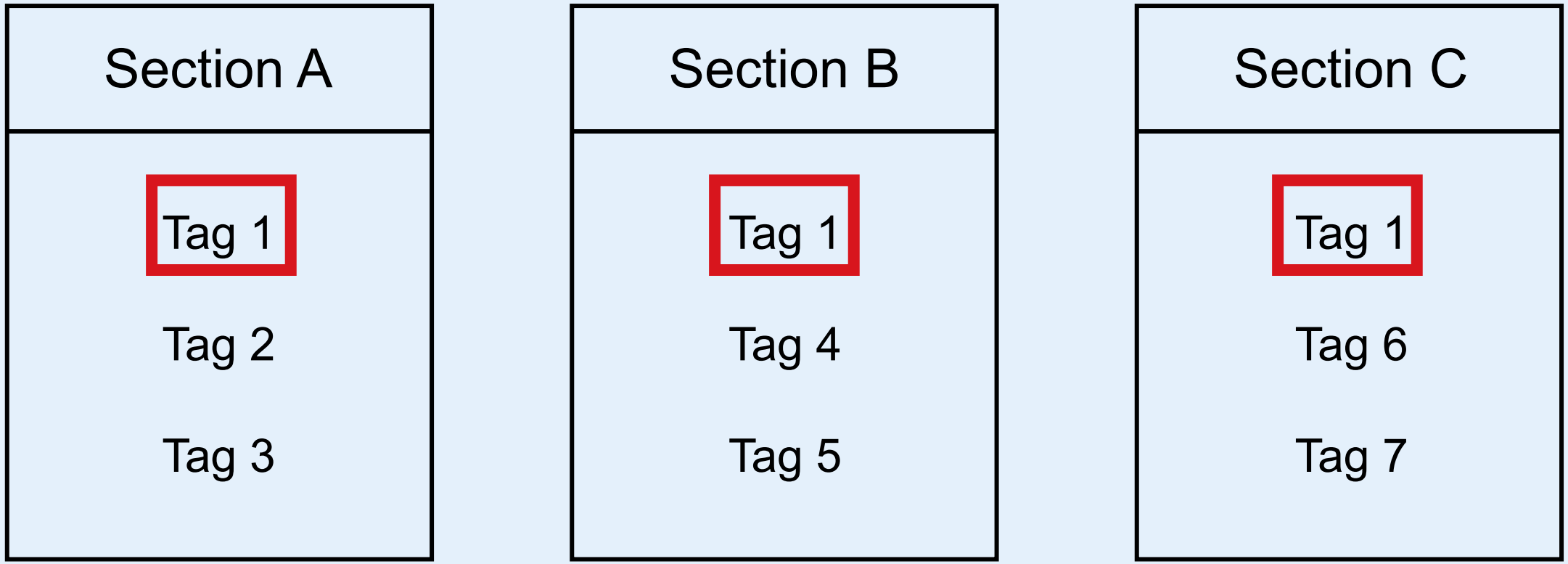


Find Decisive Tags



How to Rank Related Section?

Decisive tag: Tag 1

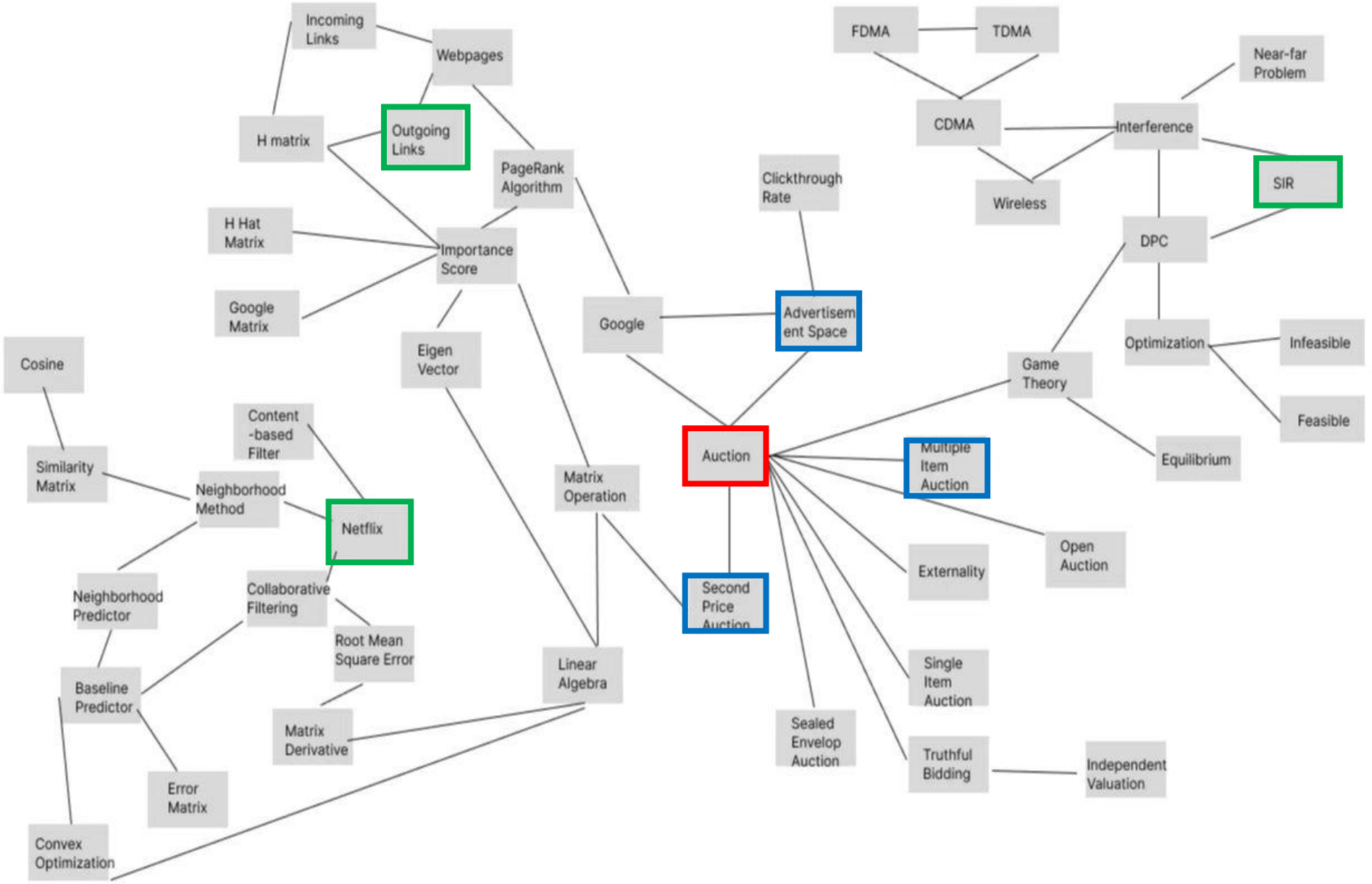


Distance Based Relevance Score (RS)

Decisive tag: Auction

Section A

Section B



Distance Based Relevance Score (RS)

For a single decisive tag — Dijkstra algorithm:

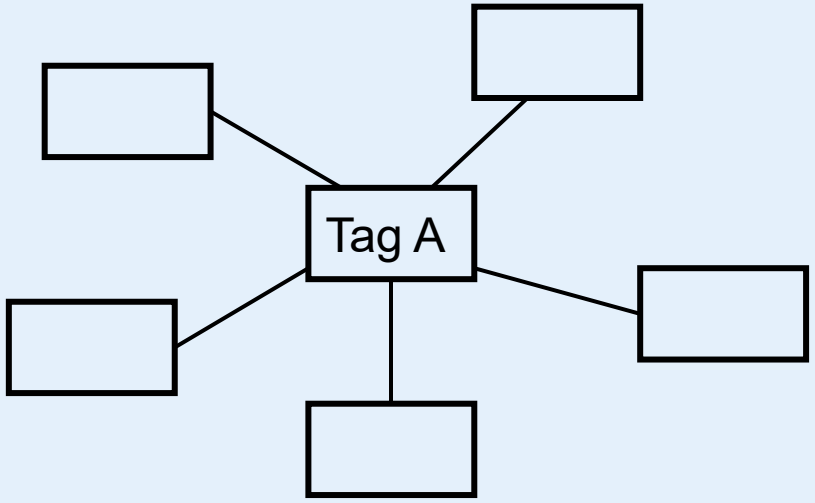
$$RS(x, y) = \frac{\alpha}{D(x, y) + \beta}$$

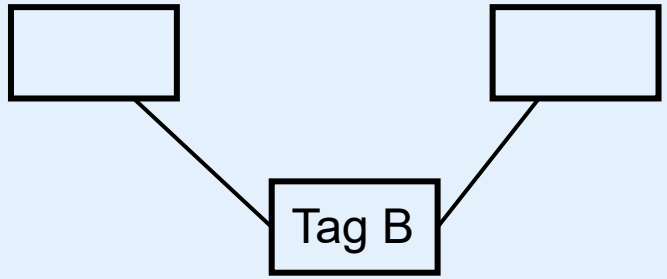
For a section containing N decisive tags:

$$SRS(s, d) = \frac{1}{N} \sum_{i \in T} RS(i, d)$$

Importance Score (IS)

Does the number of neighbor tags matter?





Basic theorem?

Example/exercise?

Inverse Page Frequency (IPF)

Two similar search phrases:

Second price auction

What is second price auction

More frequent occurrence

Less important to count

$$ipf_i = log\left(\frac{P}{C_i}\right)$$

User Feedback Adjustment

$$FS(s) = \begin{cases} \frac{\min\{p-q,\sigma\}}{\delta} & \text{if } p-q>0 \\ \frac{\max\{p-q,-\sigma\}}{\delta} & \text{if } p-q<0 \end{cases}$$

P: positive feedback

Q: negative feedback

σ，δ：constant – important but not dominant feedback

Experiment Result & Discussion

·**Select testing searching string:**

“ vote,
neighborhood
borda count
root mean square error
page rank
movie rating ”

·**Select the best target section ID:**

603, 604
402
603
401, 503
302, 303, 304
401, 402

·**Use different algorithms to get results:**

1.IS 2.RS 3.IS + RS
4.IS + RS + IPF(Inverse Page Frequency)

MRR testing algorithm

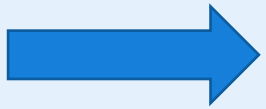
We use a modified Mean Reciprocal Rank (MRR) as our evaluation criteria. The equation is as follows:

Q : number of query test cases we have in our test set.

ni: the number of correct answers we have

Ranki: rank of a specific correct answer in our actual experiment output result.

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}$$



$$\frac{1}{Q}\sum_{i=1}^{Q}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{rank_i}\right)$$

Experiment Results

page rank/302, 303, 304			
IS	RS	IS+RS	IS+RS+IPF
501	306	501	306
301	301	301	301
201	501	306	303
504	303	201	305
305	502	303	803
503	304	504	501
303	302	502	304
306	504	503	302
502	201	305	504
302	305	304	502
304	503	302	

Apply modified MRR method...

Algorithm Evaluation

MRR	score
IS	0.4518
RS	0.4963
IS+RS	0.6119
IS+RS+IPF	0.5889

IS: importance score
RS: relevance score
IPF: inverence page frequency

Limitation

- Tags Graph
- Test case
- Section

Reference

[1] Dijkstra, E.W. A note on two problems in connexion with graphs. Numer. Math. 1, 269–271 (1959). https://doi.org/10.1007/BF01386390
[2] L. Page, S. Brin, R. Motwani, and T. Winograd, ″The PageRank Citation Ranking: Bringing Order to the Web,″ 1999.
[3] Rajaraman, A.; Ullman, J.D. (2011). ″ Data Mining″ (PDF). Mining of Massive Datasets. pp. 1–17. doi:10.1017/CBO9781139058452.002.ISBN 978-1-139-05845-2.
[4] Amardeilh, F. (2007). Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle (Doctoral dissertation, Université de Nanterre-Paris X).
[5] Kamal EL GUEMMAT, Sara OUAHABI. “Understandings of Exploratory Pedagogy for Educationnal search engine”. The Turkish Online Journal of Distance Education, 20 (2), 2019a. ISSN 1302-6488.
[6] Buckley C and Voorhees E (2000) Evaluating evaluation measure stability. In: Proceedings of SIGIR’00, New York, 2000, pp. 33–40. ACM Press.

Contribution

Task	Contributor
System Design	Chen Xuanwen Peng Yiwei Zeng Zhuoru Zhuang Yan
Ranking Algorithm Design	Chen Xuanwen Peng Yiwei Zeng Zhuoru Zhuang Yan
Database Building	Chen Xuanwen Peng Yiwei Zeng Zhuoru Zhuang Yan
Coding	Chen Xuanwen Peng Yiwei Zeng Zhuoru Zhuang Yan

Contributor	Contribution (%)
Chen Xuanwen	22%
Peng Yiwei	22%
Zeng Zhuoru	34%
Zhuang Yan	22%

Q&A