

CS6322: Information Retrieval

Sanda Harabagiu

Projects Spring 2023



Project # 1

✦ Search Engine for **Movie Awards : 5 students**

– **STUDENT 1: Crawl 100 000 pages from Oscars.org, Wikipedia, News Corp. Web pages**

✦ **E.g. Start from:**



Winners: Best Picture, Best Supporting Actor, etc ...

✦ **Green Book**

✦ Director: Peter Farrelly

✦ Awards: Academy Award for Best Picture, [MORE](#)

✦ Screenplay: Peter Farrelly, Nick Vallelonga, Brian Currie

Project # 1 - continued

✦ Search Engine for **Movie Awards : 5 students**

– **Crawl 100 000 pages from imdb.com**

✦ **Continue crawling:**

Best Director:

Alfonso Cuarón

ROMA:

✦ **Director:**

✦ [Alfonso Cuarón](#)

✦ **Writer:**

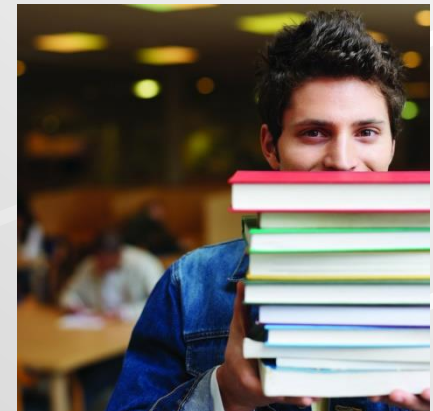
✦ [Alfonso Cuarón](#)

✦ **Stars:**

✦ [Yalitza Aparicio](#), [Marina de Tavira](#),
[Diego Cortina Autrey](#) |



1 Student responsible for
crawling



Project # 1 - continued

- ✦ Search Engine for **Movie Awards : 5 students**
 - **STUDENT 2: Index 100 000 pages & Create the WEB graph + use index and the graph to develop 2 **relevance models** + Topic-specific Page Ranking + HITS to rank the results**



1 Student responsible for
Indexing and relevance



Project # 1 - continued

✦ Search Engine for **Academy Awards**

– **STUDENT 3: Prepare the Graphical User Interface for:**

- ✦ **Introducing the query**
- ✦ **Presenting the results (including the clusters and query expansions)**
- ✦ **Showcasing the results from Google and Bing for the same query**

- ✦ **In 2 additional frames on the same web page**
- ✦ **The results should consist of three frames on the same page:**

- 1. Results of your search engine**
- 2. Results from Google**
- 3. Results from Bing**

Student responsible for :
User interfaces and
Comparisons with Google
And Bing



Project # 1 - continued

- ✦ Search Engine for **Movie Awards** – *Additional Improvements*

- **STUDENT 4: Cluster the web pages to improve results:**

- ✦ Use flat clustering
- ✦ Use 4 methods of agglomerative clustering
- ✦ Provide experimental results for 50 queries
 - ✦ with and without clustering



1 Student responsible for :
Clustering & experiments



Project # 1 - continued

✦ Search Engine for **Movie Awards – further improvements**

– **STUDENT 5: Query expansion through pseudo-relevance feedback:**

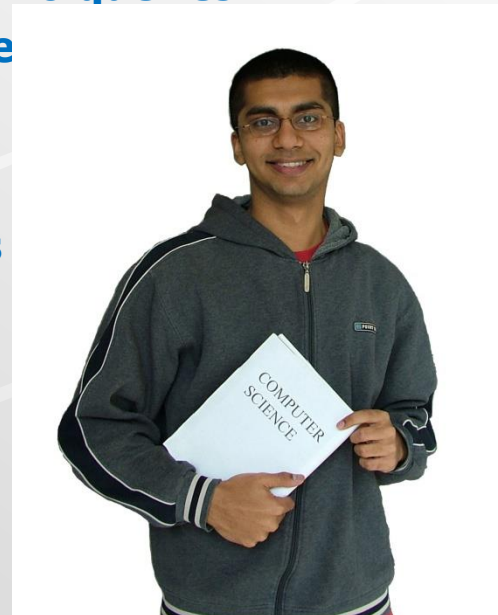
✦ **Implement the Rocchio algorithm and test it on 20 queries**

✦ **Use for automatic query expansion for 50 queries**

- association clusters,
- metric clusters
- Scalar clusters

– **Provide experimental results for 50 queries**

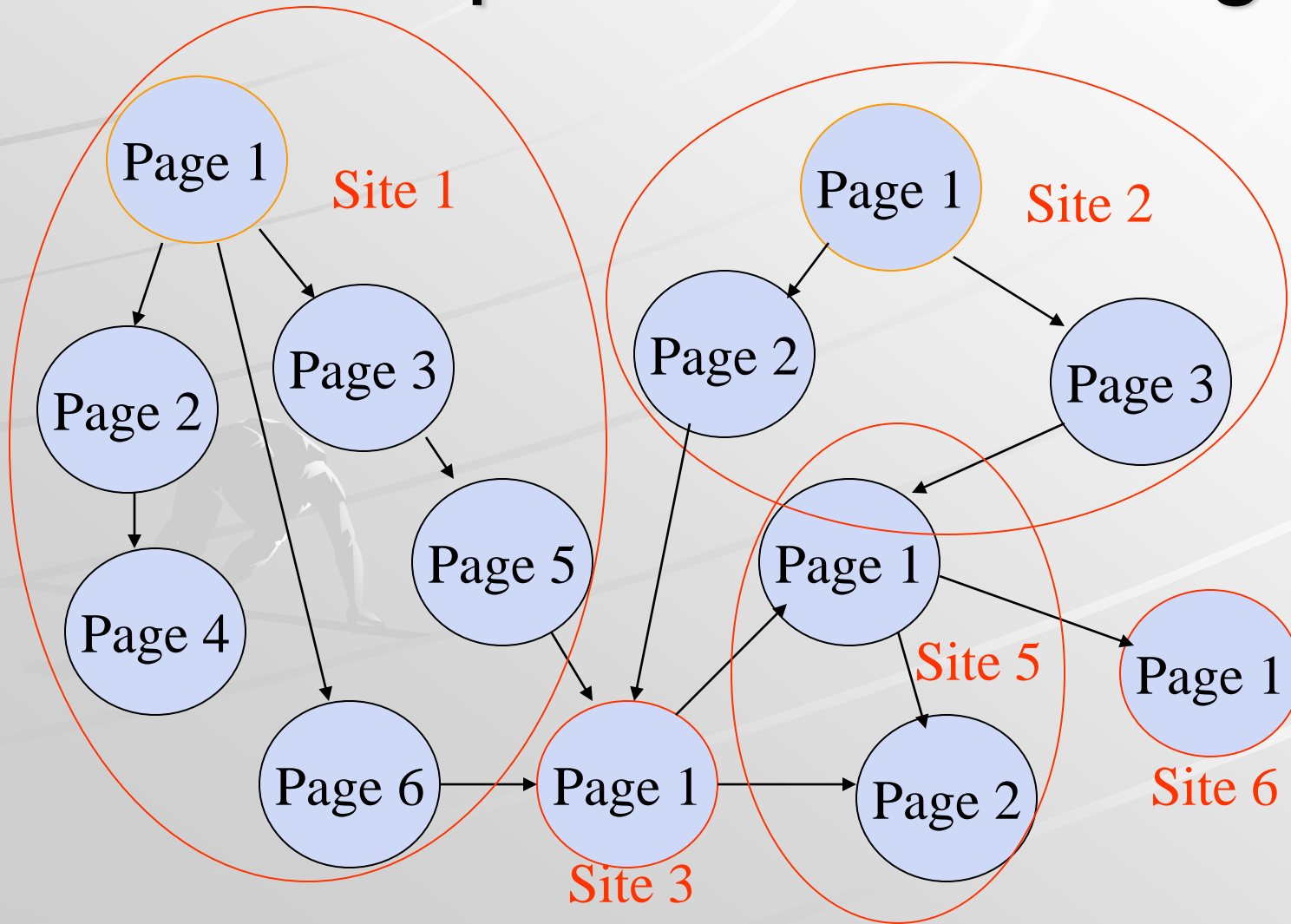
1 Student responsible for :
Relevance feedback and
Query expansion



Web Crawlers

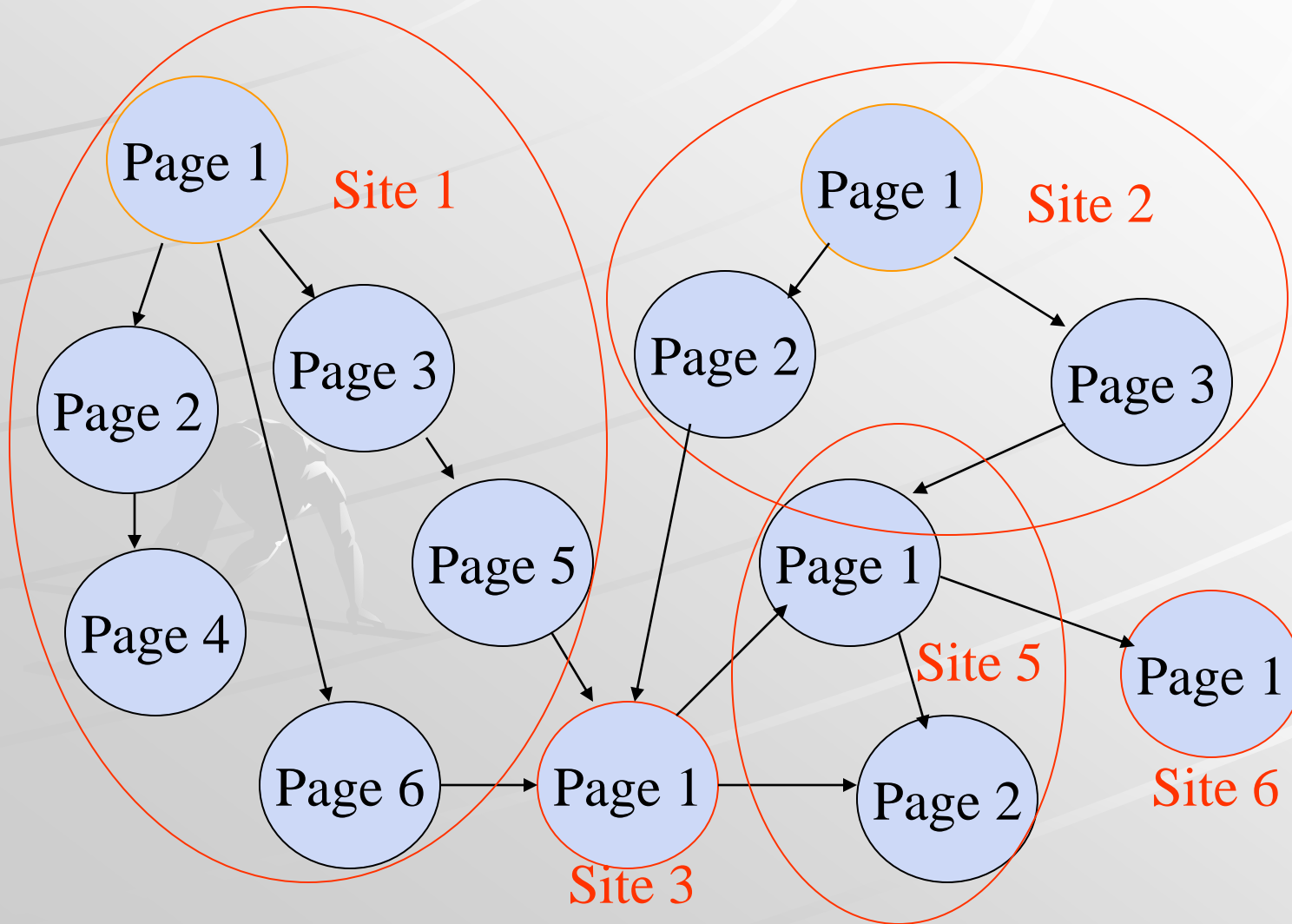
- ◆ How do the web search engines get all of the items they index?
- ◆ How much stuff is out there?
- ◆ How do you store millions of words from hundreds of sites so that you can find them quickly (and efficiently)?

Depth-First Crawling



Site	Page
1	1
1	2
1	4
1	6
1	3
1	5
3	1
5	1
6	1
5	2
2	1
2	2
2	3

Breadth First



Site	Page
1	1
2	1
1	2
1	6
1	3
2	2
2	3
1	4
3	1
1	5
5	1
5	2
6	1

Additional Search Engines

✦ 5 students

- ✦ Search Engine for **Countries/ Wikipedia**
- ✦ Search Engine for **Travel/ TripAdvisor**
- ✦ Search Engine for **Books**
- ✦ Search Engine for **Startups/ TechCrunch**
- ✦ Search Engine for **Painting/ Museums**
- ✦ Search Engine for **Tennis/ Wimbledon+Roland Garros+..**
- ✦ Search Engine for **Soccer/ World Cup+..**
- ✦ Search Engine for **Olympics/ Sports web pages**
- ✦ Search Engine for **Politics/ Election Web sites**