Towards Intelligent Content Discovery in Digital Libraries

Emanuela Mitreva, Maxim Goynov and Desislava Paneva-Marinova

Institute of Mathematics and Informatics at Bulgarian Academy of Sciences

Problem

- In recent years, a large number of items have been digitized:

✔ Bulgaria’s cultural heritage is being preserved.

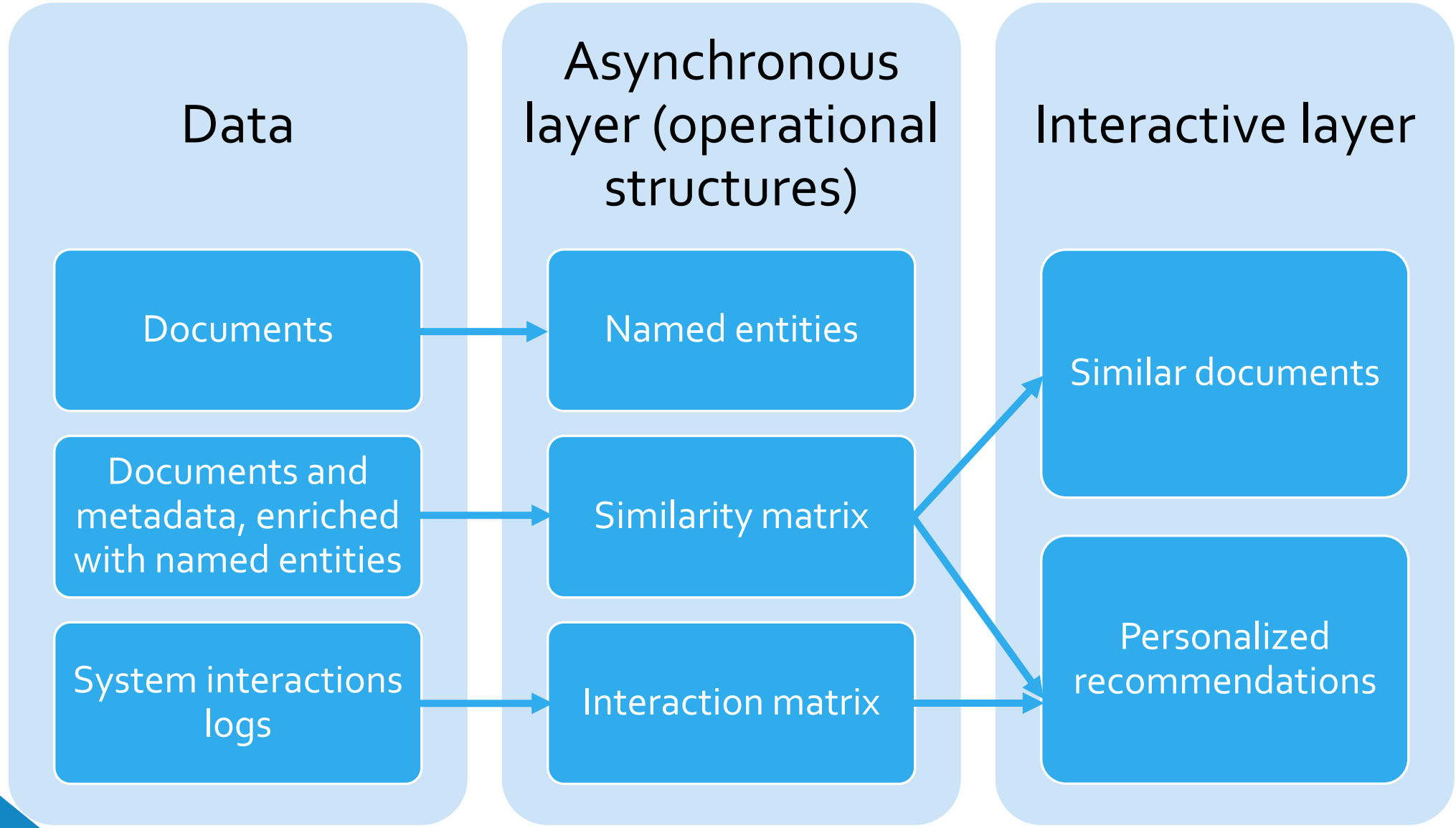
✔ More people with diverse needs and interests have access to these resources.

✖ An environment of **information overload** is created, making it increasingly difficult for users to find the resources they need.

Towards Intelligent Content Discovery

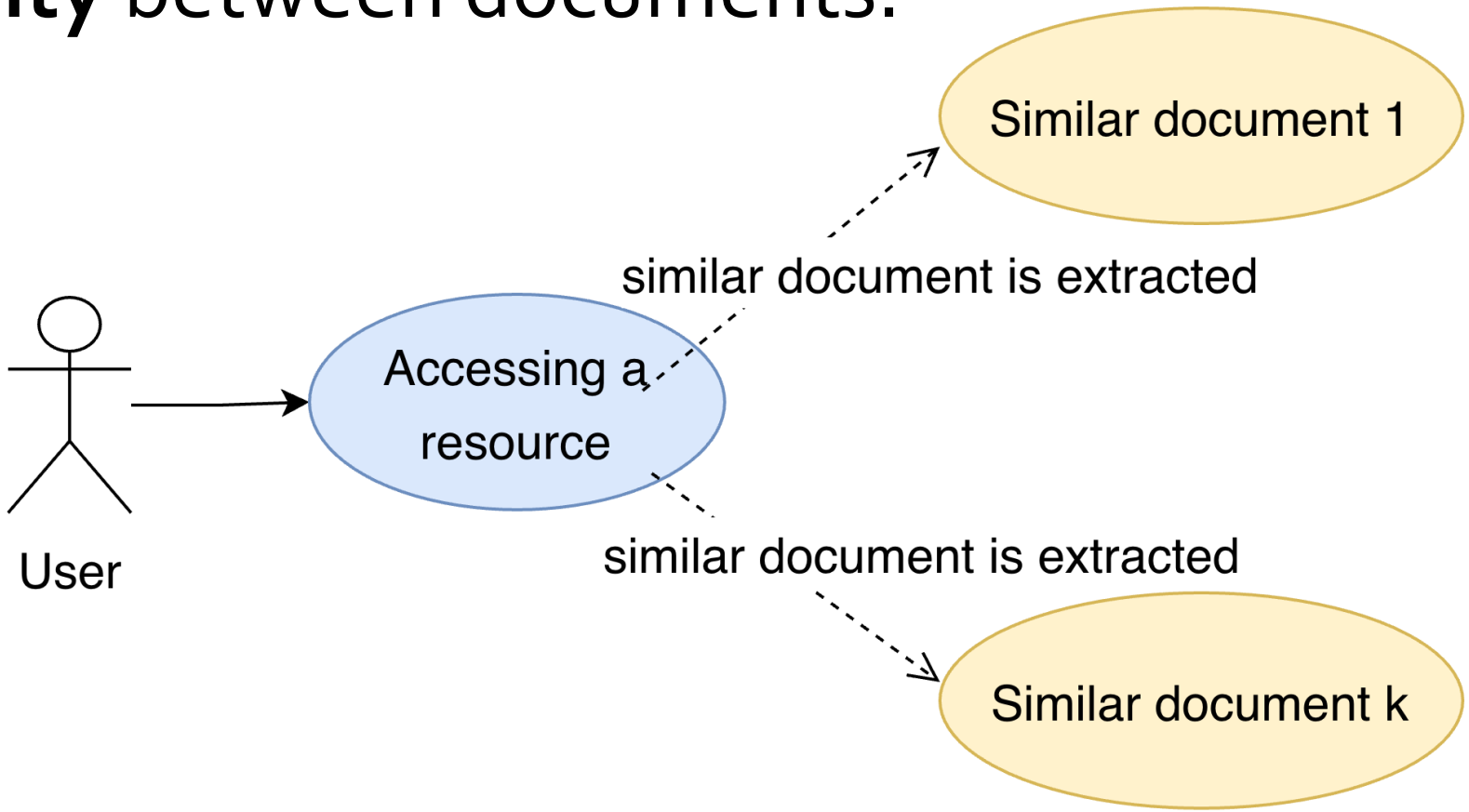
- The data used in this study consist of digital resources from the **periodicals** of the National Library of Plovdiv "Ivan Vazov".
- **Textual** resources, **user interactions**, and **enriched metadata** (named entities) were combined.
- The architecture was designed to provide two types of recommendations: "**similar documents**" and "**personalized recommendations**".

Towards Intelligent Content Discovery (2)

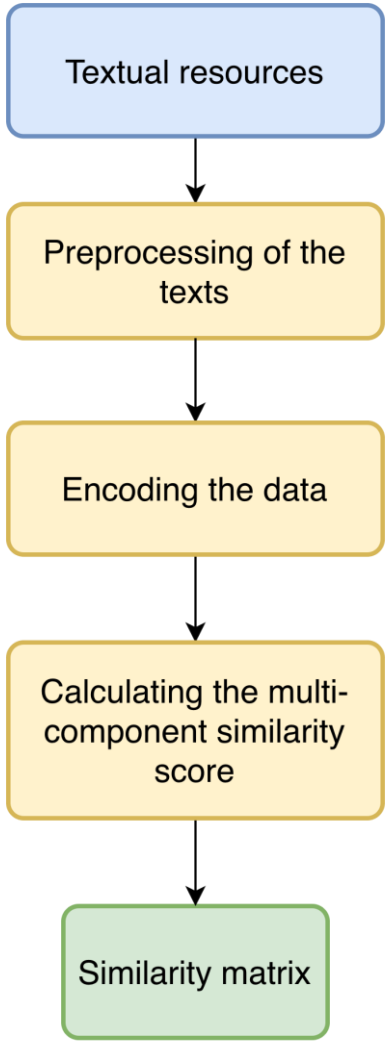


Similar Documents

A **user-independent** recommendation module for suggesting similar documents based on **semantic similarity** between documents.



Similar Documents – Similarity Matrix



- The preprocessing stage includes the removal of special characters, OCR errors, and stop words.
- The text is encoded using a **transformer-based** model into semantic vector representations (embeddings), which preserve semantic information and enable the measurement of similarity between texts.
- Document similarity is computed using a **multi-component scoring method**, and the resulting values are stored in a similarity matrix.

Similar Documents - Multi-component Score

- The final similarity score between documents is calculated using a multi-component scoring method that combines:
  - **global** semantic similarity;
  - **local text fragment** matches;
  - similarity of **thematic profiles**;
  - overlap of **named entities**.

Similar Documents - Multi-component Score (2)

$$S_{sematic}(i,j) = \alpha * S_{mean}(i,j) + \beta * S_{best-match}(i,j) + \gamma * S_{topic}(i,j)$$

Captures the overall style and dominant topic of the articles.

The maximum cosine similarity between document segments is calculated to identify similar articles across different documents.

Similarity is determined through thematic profiles obtained by fuzzy clustering, enabling the association of documents with closely related topics.

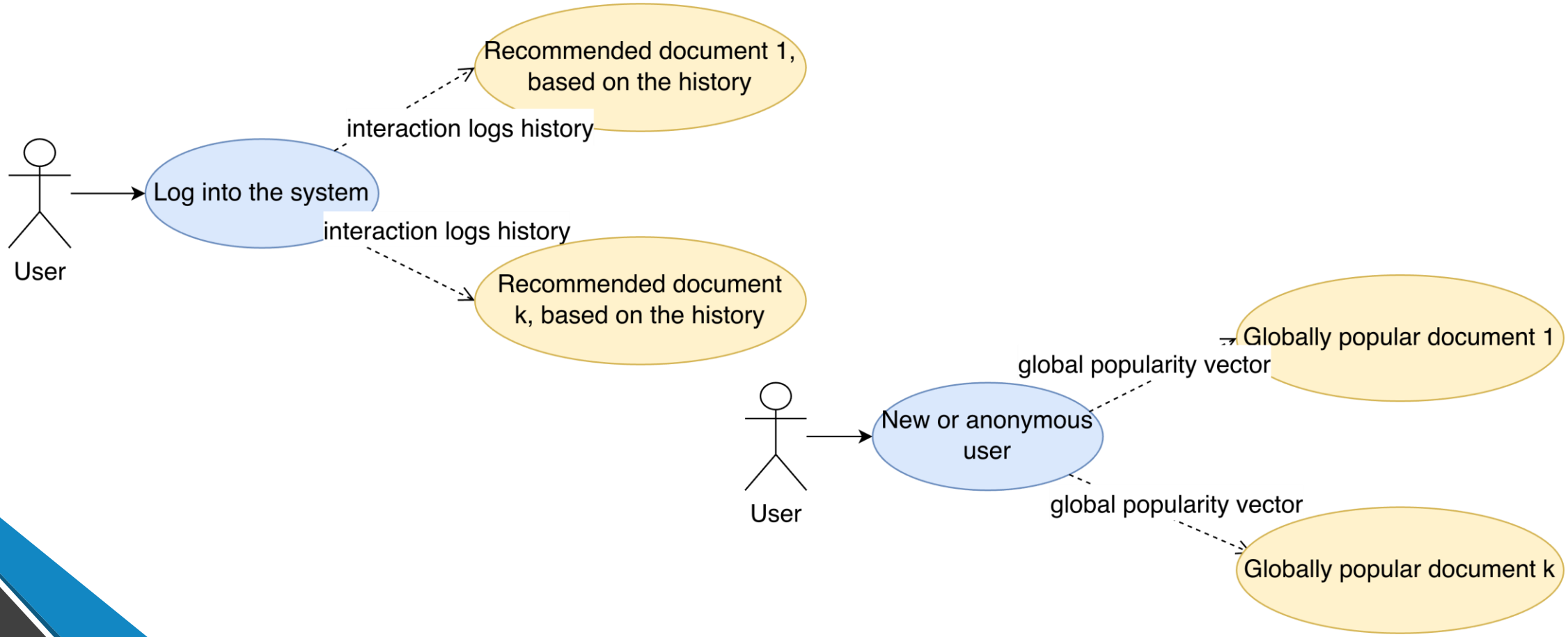
$$S_{final}(i,j) = (1 - \lambda) * S_{semantic}(i,j) + \lambda S_{named\ entities}(i,j)$$

Similar Documents - Multi-component Score (3)

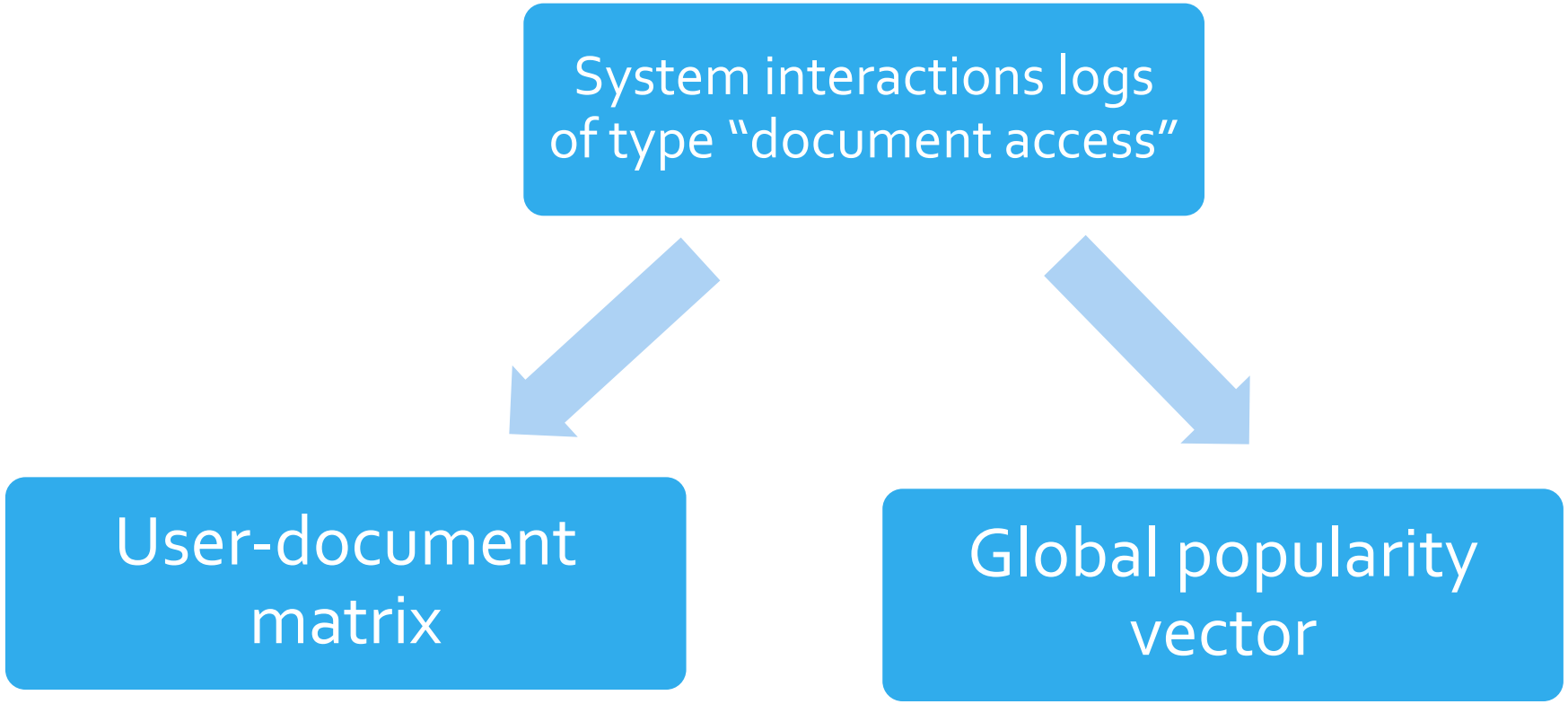
- Different-sized samples, a **synthetic** corpus, and an **ablation analysis** were used to determine the optimal weights:
  - $\alpha = 0.47$ – global semantics;
  - $\beta = 0.07$ – local matches;
  - $\gamma = 0.47$ – thematic profiles;
  - $\lambda = 0.5$ – named entities.
- **Named entities** play a key role in periodical publications covering multiple topics.
- **Local** matches contribute only marginally and help suppress noise from irrelevant passages.

Personalized Recommendations

- Combines **content similarity** and **behavioral data.**
- Generates recommendations based on the **user's interaction history**.



Personalized Recommendations (2)



Personalized Recommendations – Weighted Recommendation Confidence Score

- A **hybrid** algorithm for personalized recommendations is proposed, based on **item-based collaborative filtering**, a **content-based component**, and a **global popularity vector**.

$$score(u,d) = \sum_{i \in history(u)} w_{u,i} * S_{final}(i,d) + \partial(u) * popularity(d)$$

System interactions logs

Semantic similarity, based on the similarity matrix

Global popularity vector

Personalized Recommendations

- The **hybrid** algorithm demonstrates a **moderate** improvement over the baseline algorithm in the general case. Its primary **advantage** lies in its robustness and correct behavior in **edge-case scenarios**:
  - In **cold-start** situations, the system correctly falls back to **popularity-based** recommendations.
  - For **new documents** and sparse data, the **content-based** component compensates for the lack of user interactions.
  - The system detects **shifts in user interests** and adapts recommendations to emerging preferences.
  - When recommendation candidates are **exhausted**, the system switches to **popularity-based** recommendations.

Questions?