1st International KEYSTONE Conference
IKC 2015
Coimbra Portugal
8-9 September 2015

Recommending Web Pages using Item-based Collaborative Filtering Approaches

Sara Cadegnani[1], Francesco Guerra[1], Sergio Ilarri[2], Maria del Carmen Rodrıguez-Hernandez[2], Raquel Trillo-Lado[2], and Yannis Velegrakis[3]
[1] Università di Modena e Reggio Emilia, [2] University of Zaragoza, [3] University of Trento

Outline

- Motivation
  - Why a recommender system for web pages?
  - Where is the innovation?
- Three Approaches for Recommending Web Pages
  - No History Method – NoHi
  - My Own History Method – MOHi,
  - Collective History Method – CoHi
- Experimental Evaluation
  - The dataset provided by “Comune di Modena”
- Conclusion and future Work

Motivation

- A large number of web sites is composed of a large number of web pages with a lot of information.
  - Official web sites of Public Administrations and other Public Institution Bodies
- A huge amount of visitors is interested in exploring and analyzing the information published
  - The EC ec.europa.eu and europa.eu websites have been visited by more than 520M people in the last year

How to find the desired information?

- The information is organized in **thematic categories** and **nested sections** that generally form large trees with high height
- In other websites, users are explicitly asked to declare their roles with respect to the website
  - Only the information relevant for the specific role is shown
- Nevertheless, conceptualizations and perspectives of users and publishers can differ
  - Visitors can spend a long time looking for information in which they are interested
  - In some cases there are thousands of pages
    - The "long tail" phenomenon affects also the task of searching information

Improving the users' experience

- Search form in the header of the web pages
  - Updating a complex indexed structure which must change when the web pages are modified
  - Keyword query disambiguation
- Special web page with a list of "useful links"
  - The same content to all the users visiting the website at a specific moment
  - This type of web sites are oriented to a wide heterogeneous public
    - What is interesting for a visitor can be useless for another
    - Useful -- > useless
- User profile-based suggested links require
  - Maintaining profiles of users
    - Users should be registered
    - Need to take into account complex procedures to maintain personal information while respecting their privacy and legal issues

Idea

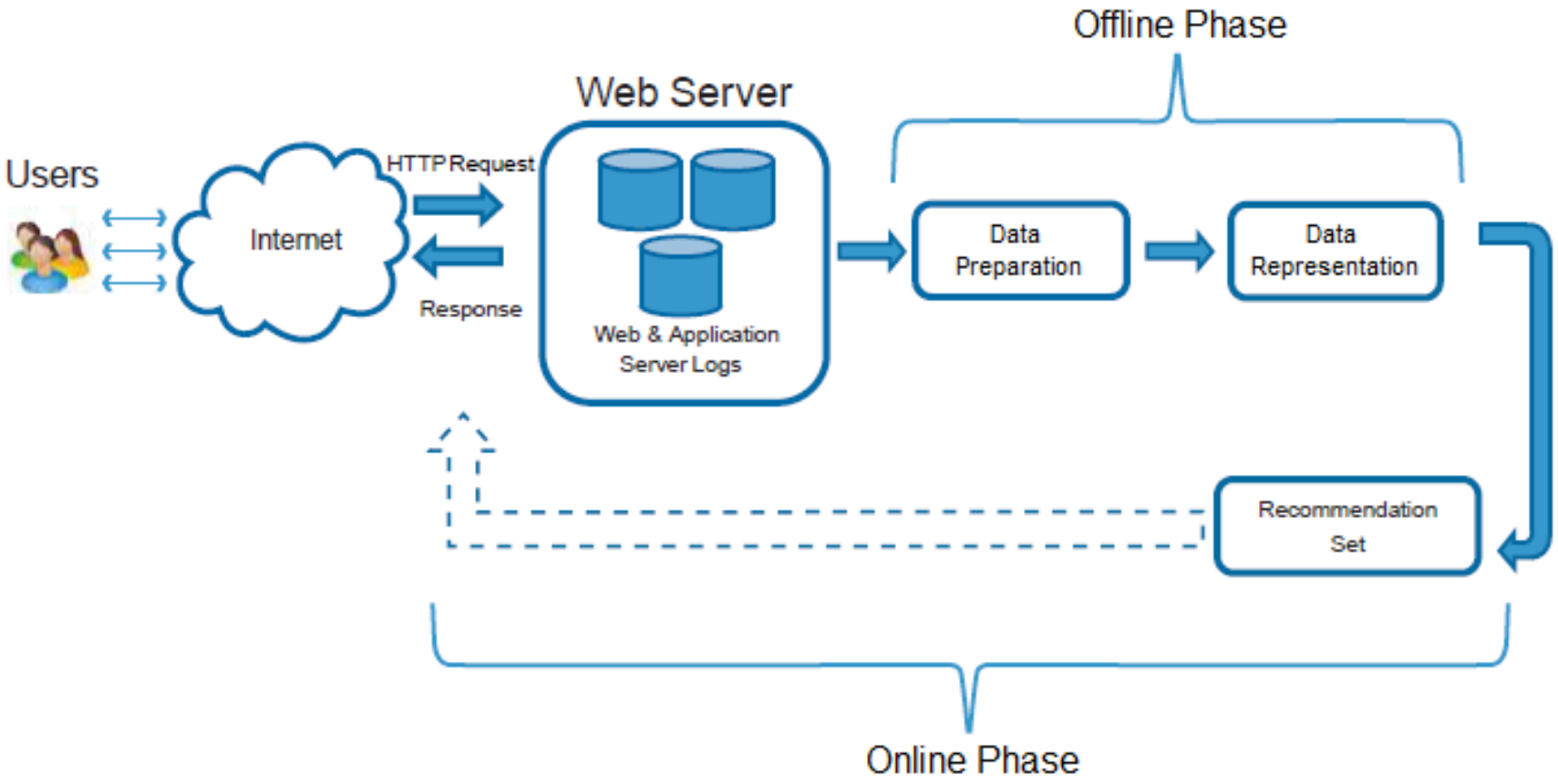
- Recommender system for web pages
  - to create a dynamic “suggested links” page
  - customized for the user who is navigating

- Our recommender system takes into account:
  - The web pages that the user is visiting in the current session
  - Navigational paths (routes)
  - The website structure
  - Lexical and semantic knowledge about the pages
    - The extraction of keywords/topics representing the content can be a huge and complex task for some websites
    - We exploit the URL as a means for approximating the content of the pages

Three approaches for recommending web pages

- No History Method – NoHi
  - Only the current user context is considered
    - This method takes into account the information of the web page that the user is currently visualizing to make the recommendation
- My Own History Method – MOHi
  - The user navigation history is considered
    - The last K web pages visited in the current session
- Collective History Method – CoHi
  - The previous sessions of other users are considered

Simple functional architecture of our recommender systems



Offline –Phase
Web page representation

Page-Feature Matrix

	F_1	F_2	F_3	…	F_m
P_1	W_{11}	W_{12}	W_{13}	…	W_{1m}
P_2	W_{21}	W_{22}	W_{23}	…	W_{2m}
P_3	W_{31}	W_{32}	W_{33}	…	W_{3m}
…	…	…	…	…	…
P_n	W_{n1}	W_{n2}	W_{n3}		W_{nm}

- Features
  - Token found in the URLs (**28 992** terms in our experiments – unigram and bi-gram), after stop-word removal and stemming
  - Weights
    - Binary matrix
    - Frequency matrix
    - tf-idf matrix

Offline –Phase
Webpage representation (2)

Session-Feature Matrix

	F_1	F_2	F_3	…	F_m
S_1	W_{11}	W_{12}	W_{13}	…	W_{1m}
S_2	W_{21}	W_{22}	W_{23}	…	W_{2m}
S_3	W_{31}	W_{32}	W_{33}	…	W_{3m}
…	…	…	…	…	…
S_n	W_{n1}	W_{n2}	W_{n3}		W_{nm}

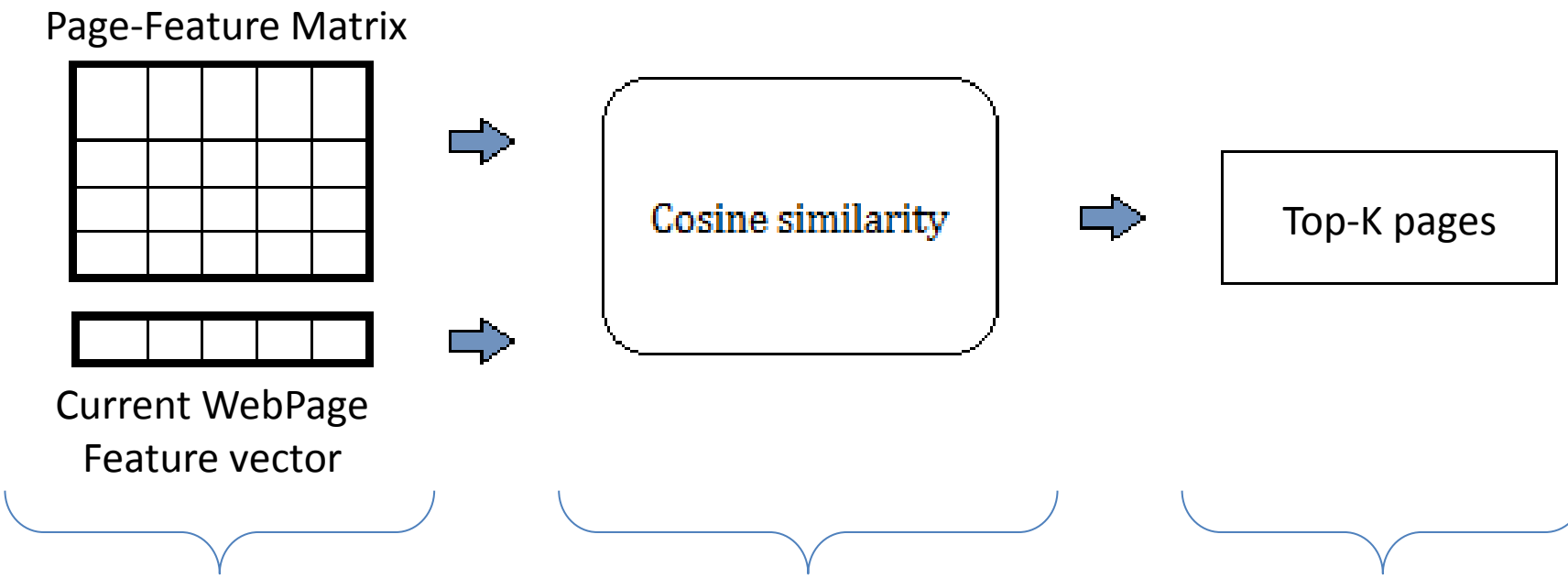
- A session is represented as

$$S_{k_{history}} = P_1 [w_{11} \; w_{12} \; w_{13} \; \dots \; w_{1m}] \oplus P_2 [w_{21} \; w_{22} \; w_{23} \; \dots \; w_{2m}] \oplus \dots \oplus P_k [w_{k1} \; w_{k2} \; w_{k3} \; \dots \; w_{km}]$$

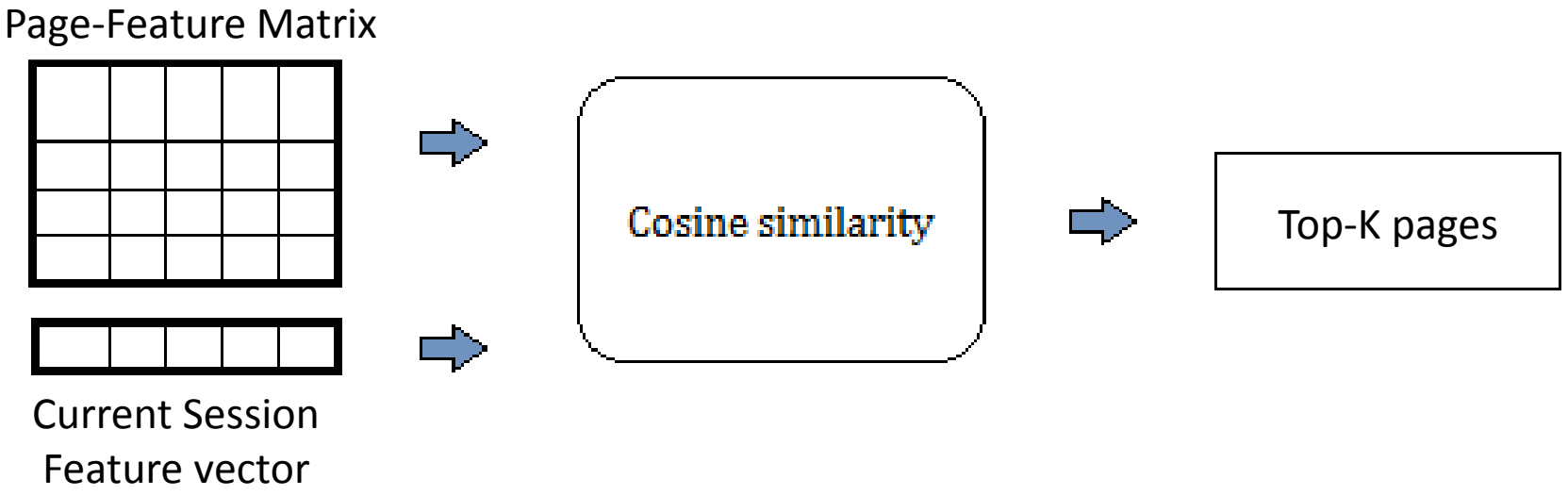
What is a session?

- A session lasts at most 30 minutes
- A session contains at least 5 items (i.e., the user has to have visited at least 5 web pages)

No History Method – NoHi

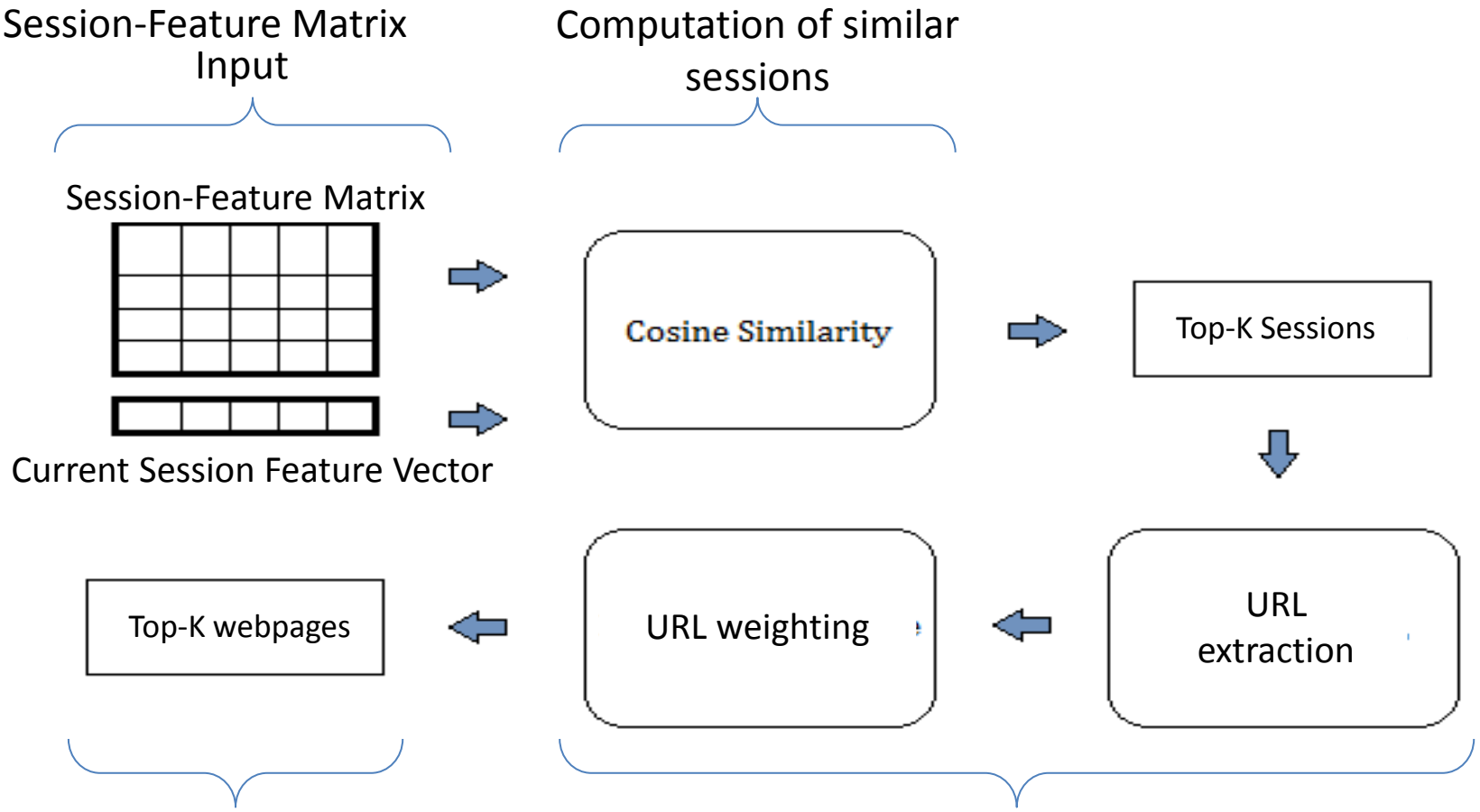


My Own History Method – MOHi



$$\mathbf{S} = \mathbf{P_1}\,[w_{11}\ w_{12}\ w_{13}\ \dots\ w_{1m}] \oplus \mathbf{P_2}\,[w_{21}\ w_{22}\ w_{23}\ \dots\ w_{2m}] \oplus \dots \oplus \mathbf{P_k}\,[w_{k1}\ w_{k2}\ w_{k3}\ \dots\ w_{km}]$$

Collective History Method – CoHi



Experimental Evaluation – the dataset

http://www.comune.modena.it

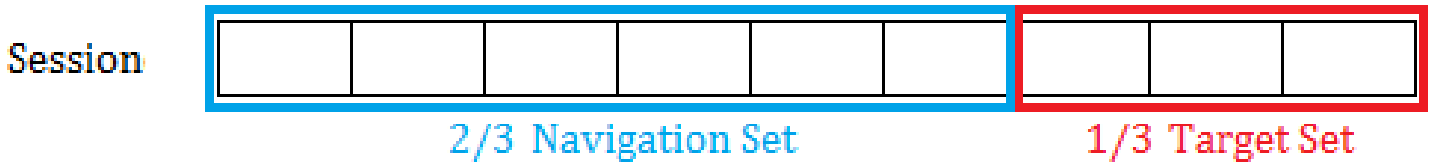
- Crawler for the structure analysis
  - Graph representation
  - More than 13 000 pages
  - More than 30 thematic areas
  - Average in-degree and out-degree of the pages around 13
  - Diameter is 8
  - Average path length is 4.57
- Analysis of the logs (2014)
  - More than 2.5 million sessions
  - Average length 2.95
  - Around 10 000 pages (72.29% of the overall number) visited by at least 1 visitor
  - 2 809 sessions (0.11% of the overall number of sessions) include in their page the “search engine page”

Experimental Evaluation – Experimental Settings

- Sessions composed of at least 5 pages:
  - 303 693, 11% of the overall amount
  - Average length 7.5 pages
  - 5 437 unigrams
  - 23 555 bigrams

- For evaluating the predictions, we divided the sessions in:
  - Test Set: 1/3 sessions (100 062)
  - Training Set: 2/3 sessions (203 631)

- In each session



Experimental Evaluation – Configurations

- No Exclusion - NE
  - URLs that the user has already visited in the current session can also be suggested
- Exclusion - E
  - URLs that the user has already visited in the current session cannot be suggested
- Sub No Exclusion - SNE
  - No Exclusion where we consider only the sessions with no repeated web pages in the set of navigation history
  - For comparing the performance with the one of a typical recommending system
    - These systems usually do not to recommend items already known/owned by the users
    - In the context of websites it is normal that people navigate the same pages multiple times
    - For this reason in this configuration we consider only cases where in the navigation history there are no pages visited several times in the same sessions
    - The same constraint is not applied in the set of correct results
- Sub With Exclusion - SE
  - Sub No Exclusion removing sessions containing repeated web pages independently of their position in the session
  - In this case, we are simulating the behavior of a typical recommending system

Results (accuracy)

	Bin	Freq	TF-IDF
NE	0,209	0,204	**0,218**
E	0,129	0,125	**0,133**
SNE	0,243	0,235	**0,256**
SE	0,252	0,242	**0,264**

Method *NoHi*

	Bin	Freq	TF-IDF
NE	0.416	0.397	**0.466**
E	0.101	0.095	**0.101**
SNE	0.186	0.178	**0.194**
SE	0.186	0.173	**0.188**

Method *MOH*

	Bin	Freq	TF-IDF
NE	0.587	0.584	**0.595**
E	0.194	0.192	**0.203**
SNE	0.314	0.31	**0.332**
SE	0.363	0.36	**0.384**

Method *CoHi*

Conclusion and future work

- Results are consistent with the state of the art
  - Similar to the ones obtained with recommender systems adopting more complex/complete knowledge
- The proposed evaluation analyzes the worst case
  - It has been done following the log analysis – what the user did
- Future work
  - Improving the session analysis we assumed that if a user visits a page, he/she is interested in the content of that page in the web site
    - However, it is possible that a user visits a web page for other reasons (mistake)
  - Analysis of time spent in the web pages to filter the logs
  - Analysis post implementation