Recommender system for science: A basic Taxonomy

Ali Ghannadrad, Morteza Arezoumandan, Leonardo Candela and Donatella Castelli

Twitter: @AliGhannadrad
Email: ali.ghannadrad@isti.cnr.it
Outline

- Introduction
- Methodology
- Analysis
- Conclusion and Future works
Introduction

● The volume of science doubles every 10 to 15 years [1].

● It becomes difficult for researchers to discover relevant scientific artefacts.

● Recommender systems are software systems devised to recommend items to users based on their observed interests.

● No systematic literature survey has been performed to document the state of the art of recommender systems in science settings.

● We provided a taxonomy regarding the scientific artefacts recommender systems stemming from a systematic mapping study of the current literature.

Methodology

- This research was carried out as a **Systematic Mapping Study** (SMS).
- The goal of the study is reflected in these **research questions**:
  1. How are users (and their interests) represented?
  2. What are the items of interest, and how are these items characterised?
  3. Which recommender algorithms have been used?
  4. Which evaluation methods have been used?
## Methodology

### Conducting search:

- Selecting keywords and creating query

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Synonym and related concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommender</td>
<td>Recommendation</td>
</tr>
<tr>
<td>Scientific products and Science</td>
<td>Scientific - Researcher - Science - Articles - Papers - Datasets</td>
</tr>
</tbody>
</table>
Methodology

- Conduct search on scientific repositories
  - 1. ACM
  - 2. IEEEXplore
  - 3. ScienceDirect
  - 4. Springer
  - 5. Scopus

- We identified **3787** primary papers.

- **Papers Screening:**
  - After removing the duplications, we explore the papers in terms of publication type and year to find the inclusion and exclusion criteria.
Methodology

- **Journal articles** and **Conferences proceedings** are considered as inclusion criteria.
Methodology

- Published papers between 2015 and 2022 are considered as inclusion criteria.
Methodology

- After reviewing the papers we reached the final dataset which contains 209 papers.

<table>
<thead>
<tr>
<th>Repository;</th>
<th>ACM</th>
<th>IEEEXplore</th>
<th>ScienceDirect</th>
<th>Springer</th>
<th>Scopus</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>After removing duplicates:</td>
<td>114</td>
<td>64</td>
<td>152</td>
<td>40</td>
<td>2205</td>
<td>2575</td>
</tr>
<tr>
<td>After applying criteria:</td>
<td>64</td>
<td>6</td>
<td>53</td>
<td>11</td>
<td>853</td>
<td>987</td>
</tr>
<tr>
<td>After reviewing:</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>192</td>
<td>209</td>
</tr>
</tbody>
</table>
Analysis - Classification scheme

Recommender system for science

Evaluation
- Evaluation methods
  - Decision support-based
  - Error-based
  - Ranking-Based

- Evaluation metrics
  - Online
  - Offline

Algorithms
- Collaborative-Based
- Content-Based
- Hybrid-Based
- Graph-Based

Content-Based + Collaborative-Based
Graph-Based + Content-Based
Graph-Based + Collaborative-Based

Users
- User Representation
  - Implicit
  - Explicit
- User Types
  - Individual
  - Group

Items
- Type of scientific artefacts
  - TF-IDF
  - Topic Modelling
  - Word Embedding
  - Graph Embedding
  - Others
- Item Representation
  - Paper-Workflow
  - Dataset-Others
Analysis - User types and representation

- Only 4 papers are identified where the target is a group of researchers.
- 205 papers out of 209 papers are focused on individual users.
Analysis - User types and representation

- Implicit representation: The system captures users’ interests **indirectly**.
- Explicit representation: The system relies on the **user's input** which could be a query, paper, dataset, etc.

<table>
<thead>
<tr>
<th>Implicit user representation</th>
<th>Graph representation</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Profile</td>
<td>The system constructs a representation of the user preferences.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The system relies on connections or links between nodes which are representing the user.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The system merges different representation methods.</td>
<td></td>
</tr>
</tbody>
</table>
Analysis - Item types and representation

- **16** heterogeneous typologies of **artefacts** are identified.
- **Paper recommender system** are proposed in 134 of the 209 papers reviewed.
- **Software recommender system** is unprecedented.
- Others: Keyword, Tag, Research area, Paper submission, etc.
Analysis - Item types and representation

- For almost all of the scientific artefacts it is possible to have a **text-based characterization**.

- We analysed and classified text-based representations methods.

```
Text-based item representation

- TF-IDF
- LDA
- Topic Modeling
- LSA
- Word Embedding
- Word2Vec
- Doc2Vec
- Glove
- Graph Embedding
- Mixed
```
Analysis - Item types and representation

- TF-IDF, Topic Modeling and Word embedding are applied in the case of Content-based Filtering.
- The goal of word embedding method is to capture semantic and syntactic regularities.
- Graph embedding can be used in Graph-based algorithms like citation network.
Analysis - Algorithms

- 56 of the 134 paper recommender systems used **hybrid approaches**, while 37 of them used **content-based algorithm**.

- The most used combination in Hybrid-based is **Graph-based + Content-based**.

```
Paper Recommender system algorithms

- Content-based (37)
- Collaborative-based (15)
- Graph-based (26)
- Bibliographic coupling
- Co-citation
- Hybrid-based (56)
  - Graph-based + Content-based (34)
  - Graph-based + Collaborative-based (2)
  - Content-based + Collaborative based (15)
```
Analysis - Evaluation

- **Online evaluation method**: observe the user interactions regarding the given recommendations.
- **Offline evaluation method**: test the effectiveness of recommender system algorithms on a certain dataset.
Conclusion and Future works

- We had a **Systematic Mapping Approach** on the **recommender system for science**.
- **209** papers of interest have been published between 2015 and 2022 are reviewed.
- A **taxonomy** of recommender system for science is presented.
- The **paper recommender system** is the predominant recommendation class and there is a huge gap in recommending other scientific artifacts like **datasets** and **softwares**.
- Lack of recommending scientific artefacts to the **group of researchers**.
- Most of the scientific artefacts recommendation systems relied on the **offline evaluation**.
- **Diversity** and **serendipity** of the recommended items can be taken into account.
Next Step..

- Exploiting OpenAIRE graph to compare effectiveness of different Dataset Recommender system approaches.
Thank you for your attention.

Twitter: @AliGhannadrad
Email: ali.ghannadrad@isti.cnr.it