Journal of Scientometric Research, 2025; 14(1):351-364.
https://www.jscires.org

Research Article

Thesis Reviewer's Committee Recommendation System

Omar Juárez Gambino[1,*], Consuelo-Varinia García-Mendoza[1], José-Manuel Suárez-Bautista[1], Hannah Gu[1], Hiram Calvo[2]

[1]Departamento de Ciencias e Ingeniería de la Computación, Instituto Politécnico Nacional, ESCOM, Av. Juan de Dios Bátiz/n esq. Av. Miguel Othón de Mendizabal, Mexico City, MEXICO.
[2]Laboratorio de Ciencias Cognitivas Computacionales, Instituto Politécnico Nacional, CIC, Av. Juan de Dios Bátiz/n esq. Av. Miguel Othón de Mendizabal, Mexico City, MEXICO.

ABSTRACT
This paper presents a recommender system for the assignment of thesis reviewers at ESCOM-IPN Mexico. The system analyzes the thesis proposals, extracting relevant information and comparing it against the profiles of professors and courses given by the academic departments. The system determines the most suitable profiles and departments for the thesis review through a similarity evaluation. The results show that the system can generate accurate recommendations with a high degree of coincidence with the current process carried out by the committee in charge. In addition, the system performs an alternative process by suggesting reviewers with profiles more closely related to the thesis proposals considering additional information such as research papers and areas of expertise.

Keywords: Review Assignment Problem, Recommendation System, Semantic Similarity, Thesis Reviewer Committee.

Correspondence:
Omar Juárez Gambino
Departamento de Ciencias e Ingeniería de la Computación, Instituto Politécnico Nacional, ESCOM, Av. Juan de Dios Bátiz/n esq. Av. Miguel Othón de Mendizabal, Mexico City-07738, MEXICO.
Email: jjuarezg@ipn.mx
ORCID: 0000-0002-4019-6584

Received: 25-07-2024;
Revised: 23-09-2024;
Accepted: 06-11-2024.

INTRODUCTION

An essential process in the educational trajectory of university students is the writing of the thesis. A panel of experts thoroughly reviews the thesis to determine whether the student's work is good enough to be awarded the degree. The selection of panel members should consider their academic profiles and their similarity with the topic of the thesis that will be reviewed.

The assignment of reviewers to any task, known as the Reviewer Assignment Problem (RAP),[1] is challenging and time consuming. Manual completion of this process is often biased and some relevant aspects are often overlooked, given the limited time and human resources available.[2] Automation of this process is highly desirable as it would reduce these problems. RAP has a wide field of applications, such as identifying reviewers for research projects,[3] reviewers for journal articles[4] and reviewers for grant assignments.[5] Given the relevance of this task, several automation approaches have been explored, ranging from rule-based systems[6] to those based on semantic similarity.[7]

The assignment of thesis reviewers is a particular case of RAP. Each university defines its processes for selecting and forming review committees. These processes consider factors such as the number of reviewers on the committee, the maximum number of theses that can be assigned to each reviewer and the profiles that reviewers must have. Systems that automate this process must select profiles most similar to the work proposed in the thesis.[8]

In this work, we propose a recommendation system for thesis reviewer assignments. The process at ESCOM-IPN[1] in Mexico will be reviewed as a case study. In ESCOM-IPN, there is no direct assignment of thesis reviewers by a committee. The committee reviews the title and abstract of the theses and, based on their content, selects the academic department (hereafter referred to as the academy) where professors with the appropriate profiles to review them are affiliated. The professors then select the theses assigned to their academy that interest them. This process has many problems, such as bias, limited choice of reviewers and unbalanced workload.

The proposed system performs a more detailed review of the proposal, considering additional sections such as keywords and objectives. In addition, it generates more complete profiles of the professors by taking into account not only the academies they belong to but also the articles published and theses supervised, among other information regarding their academic trajectory. Using all this information, the system can generate more accurate recommendations per academy, as is currently done and recommend the best-qualified professors to be thesis reviewers. The recommendations are based on the semantic similarity of the

[1]https://www.escom.ipn.mx/

OPEN ACCESS
Manuscript TECHNOMEDIA

DOI: 10.5530/jscires.20251458

content of the proposals concerning the profiles of the academies and the professors. The system was evaluated by comparing the system's recommendations with the assignments made by the committee.

The rest of the paper is organized as follows. First, the related work is presented in Section Related Works. Then Section thesis reviewer assignment process at escom-IPN explains the process followed at ESCOM-IPN for the assignment of thesis reviewers. Section Proposed Recommender System describes our recommendation system. Next, in Section Experiments and Results, experiments and results are shown, along with a brief discussion of the system's performance. Finally, the conclusions and future work are given in Section Conclusion And Future Work.

RELATED WORKS

Since 1992,[9] several studies have tried to solve the problem of assigning reviewers to submitted proposals from two perspectives: (1) as a Natural Language Processing (NLP) task that uses text mining methods to solve it; in this context, the problem is called Reviewer Assignment Problem (RAP) and (2) as an Operations Research task that models it as a Multiple Criteria Decision Making (MCDM) problem. In our work we follow the first perspective.

RAP has two stages: (1) reviewer representation and proposal information and (2) reviewer assignment. The second stage has been approached mainly from two perspectives: (1) the Information Retrieval-Based Reviewer Assignment Problem (IRRAP) and (2) the Optimization-Based Reviewer Assignment Problem (ORAP). IRRAP focuses on calculating the proposal-reviewer similarity score, while ORAP turns the IRRAP into an optimization problem and tries to solve this problem from the operations research perspective. Some studies focus only on one stage and others on both. The proposed recommender system tackles the IRRAP. Some similar works are described below.

According to,[1] three application areas have been the most studied: the assignment of reviewers to conference papers[10-16] to journal papers[2,7,17] and to project proposals[5,18] In addition, solutions have been proposed for the assignment of thesis reviewers and the allocation of grants.

In,[8] an undergraduate thesis reviewer recommender system is proposed. Two types of documents were used for this work: completed theses and thesis proposals from the Department of Informatics, Universitas Sebelas Maret (UNS). The information extracted from the completed theses was the abstract and the basic theory section, while the information from the thesis proposals was only the basic theory section. TF-IDF was used to represent the documents and they were subsequently clustered with the K-means algorithm. Finally, the authors used Euclidean distance to determine the closest completed theses to the thesis proposals. Supervisors of the three closest documents are assigned as reviewers of the thesis proposal.

In,[19] the Thesis Reviewers's Personalized Recommender Model (TPRPRM) is presented. The researchers fed the model with information from master's and doctoral theses from the University of Chinese Academy of Sciences (UCAS). The data set comprises the reviewer ID, the thesis ID, the matching label, the reviewed times and 30 topics covered. The model proposes topic-topic cross and feature cross methods to find the similarity between the topics of the proposed theses and the topics of the theses reviewed by the reviewers; this similarity is calculated with cosine similarity. The TPRPRM recommender model consists of 4 layers: (1) an embedding layer to compress a high-dimensional feature space, (2) an Attentive Factorization Machine layer to build topic relevance in the utmost probability, (3) a Neural Network layer to mine implicit topics similarity and (4) Output Layer to provide a more reliable and accurate recommendation list. Finally, the performance of TPRPRM is measured with the precision and novelty metrics, trying to achieve a balance between them.

In,[20] the reviewer selection process for an undergraduate thesis is modeled as a Multiple Criteria Decision Making (MCDM) problem. A thesis committee, composed of three academics, defined seven evaluation criteria (i.e., position, competencies, research group, publications, reviewer experience, academic background and period of employment) to select the two most suitable reviewers from a pool of four candidates. The selection was performed using the IF-TOPSIS method, which combines the Intuitionistic Fuzzy (IF) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) methods. First the linguistic terms were characterized as numbers to evaluate the alternatives and assign a weight to each criterion, then with the Euclidean distance, the relative proximity coefficients of the alternatives were calculated and ranked. Additionally, an Aggregated Intuitionistic Fuzzy Decision Matrix (AIFDM) is proposed to include the opinions of the thesis committee.

After reviewing the state of the art, we can observe that the task of RAP is very relevant given its numerous applications. Different approaches have been proposed to solve this problem, particularly for the assignment of article reviewers. On the other hand, work related to the assignment of thesis reviewers, which we consider equally important, is scarce, so our work contributes to this direction.

THESIS REVIEWER ASSIGNMENT PROCESS AT ESCOM-IPN

In this work, the process of assigning thesis reviewers at ESCOM-IPN will be examined as a case study. The approval of a thesis at ESCOM-IPN is a crucial evaluation of a student's knowledge acquisition in their major, representing a significant

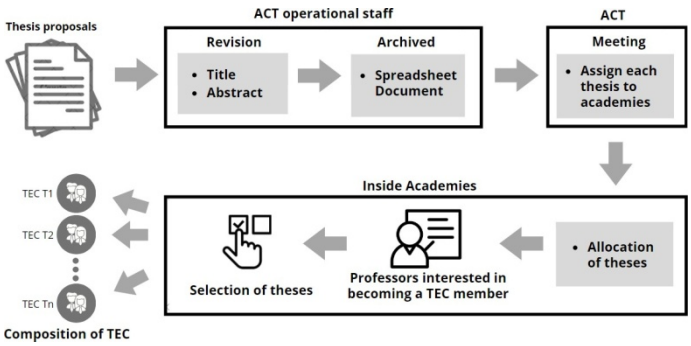


Figure 1: Current process of thesis reviewer assignment.

milestone toward graduation. In the context of ESCOM-IPN, a thesis represents the culmination of a student's academic journey, demonstrating the development of a project that validates the knowledge and skills acquired during their academic program.

According to the regulations, each thesis is entrusted to a Thesis Evaluation Committee (TEC) of four to five professors. Three of these professors take the role of reviewers, while the remaining TEC members are thesis supervisors. The supervisors are responsible for proposing the thesis topics and guiding the students, while reviewers are responsible for monitoring, recommending and providing suggestions in their respective areas of expertise.

Currently, an Academic Thesis Committee (ACT) is in charge of carrying out the administrative processes involved in this matter. One of its functions is to assign theses to the different academies of ESCOM-IPN, where the professors considered qualified to be thesis reviewers are affiliated. In other words, the committee does not directly assign theses to reviewers. Instead, it defines which academies each thesis should be addressed to so that the professors of these academies can select the ones that interest them and thus join the evaluation committee as reviewers. ACT comprises key stakeholders, including the director and deputy director of ESCOM-IPN, the head of the Comprehensive and Institutional Training Department, directors of academic departments and the heads of the academies.

Decisions about the assignment of a specific thesis to a particular academy are based on the review of the title and abstract provided in the proposal. Each thesis proposal is assigned to two or three specific academies (avoiding assigning it to only one) and then the professors of these academies select the theses according to their interests. Figure 1 shows the thesis reviewer assignment process. This important process involves several stages that demand considerable time and effort from ACT members, particularly since each semester witnesses approximately 120 proposals for new theses. The manual nature of this process becomes laborious given the high volume of thesis submissions.

Furthermore, due to a lack of time and limited human resources, the ACT relies only on the title and abstract for decision-making. This, coupled with the student's ability to describe the thesis through these two sections, may lead to erroneous academy assignments and consequently to less qualified reviewers. Even if the selected academies are the appropriate, there is a risk that more experienced professors may not select the thesis because there is no direct assignment. There is also the possibility of qualified professors being unaware of the existence of a thesis if they belong to an academy excluded by the ACT during the assignment process.

Given these challenges, there exists a compelling need for a more streamlined and targeted approach in assigning thesis reviewers. This requires considering both the expertise of individual professors and the specific requirements of each thesis.

Therefore, the proposed solution consists of developing a system that automatically extracts information from thesis proposals, categorizes them according to academic areas and suggests suitable professors. The automation of this task will significantly simplify the current process, offering a more efficient and effective means of assigning reviewers. This not only speed up the process but also has the potential to improve the overall quality of thesis evaluations at ESCOM-IPN. The following section describes the stages of development of the recommender system.

PROPOSED RECOMMENDER SYSTEM

Given the problems with the current reviewer assignment process, we propose a system that automates the analysis of thesis proposals, considers more detailed information and determines the academies with the most qualified reviewers. The objective is to streamline the reviewer selection process and find the most relevant profiles while respecting the operating standards established by the ACT.

The implemented approach uses a probabilistic model that identifies the relevant academies for a thesis proposal, considering the courses taught in these academies. In addition, the system identifies the professors whose profiles are the most similar to that proposal.

Two different approaches are used to tackle the problem of reviewer assignment. The first follows a process similar to that

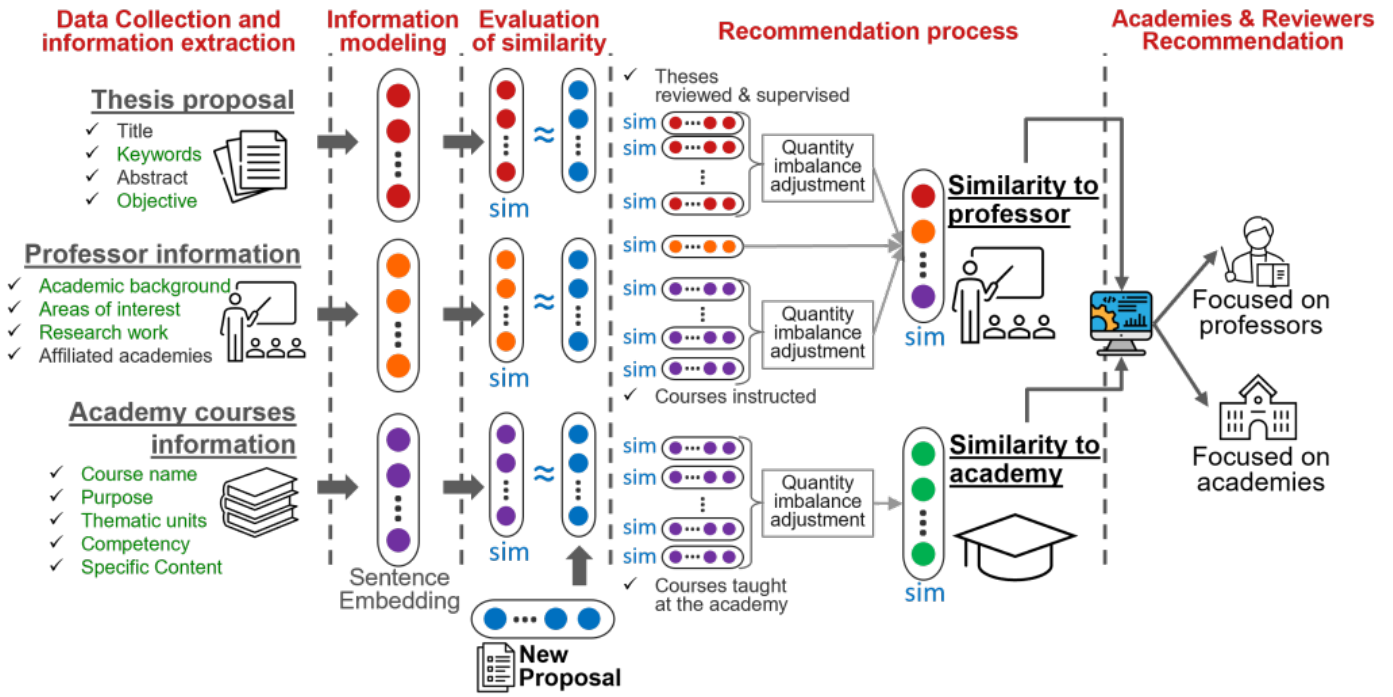
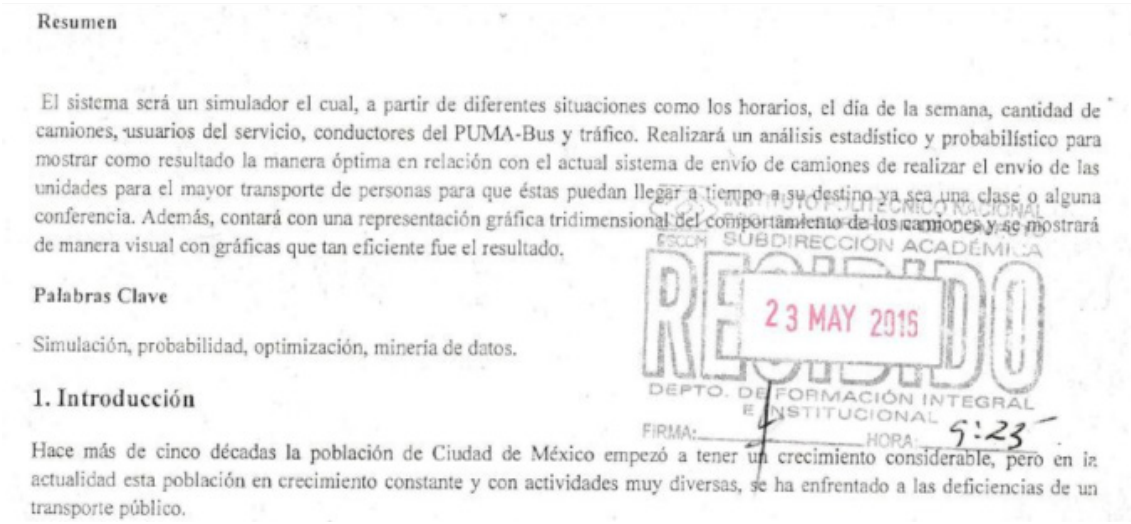


Figure 2: Overview of the proposed recommender system.



Resumen

El sistema será un simulador el cual, a partir de diferentes situaciones como los horarios, el día de la semana, cantidad de camiones, usuarios del servicio, conductores del PUMA-Bus y tráfico. Realizará un análisis estadístico y probabilístico para mostrar como resultado la manera óptima en relación con el actual sistema de envío de camiones de realizar el envío de las unidades para el mayor transporte de personas para que éstas puedan llegar a tiempo a su destino ya sea una clase o alguna conferencia. Además, contará con una representación gráfica tridimensional del comportamiento de los camiones y se mostrará de manera visual con gráficas que tan eficiente fue el resultado.

Palabras Clave

Simulación, probabilidad, optimización, minería de datos.

1. Introducción

Hace más de cinco décadas la población de Ciudad de México empezó a tener un crecimiento considerable, pero en la actualidad esta población en crecimiento constante y con actividades muy diversas, se ha enfrentado a las deficiencias de un transporte público.

Figure 3: Information overlaid on a scanned thesis proposal document.

currently used by the ACT. In this approach, the content of the courses taught in the academies is reviewed to find similarities between the course content and the thesis proposals. The academies with courses with contents similar to the proposal are recommended. In contrast, the second approach directly proposes the three most qualified professors to be reviewers. This assignment is carried out by generating profiles of the professors, considering relevant aspects of their academic trajectory. These profiles are then compared with the thesis proposals and the professors whose profiles show the greatest similarity are recommended as reviewers.

A good source of information is essential for the system to work properly, so the first step was to collect the necessary data. During this phase, data from previously completed thesis proposals, professors and academies were collected, extracted and pre-processed to model the relevant information. Subsequently, the system compares this information with the new thesis proposals and finds, according to the selected approach, the relevant academies for thesis assignments or the most suitable professors to act as reviewers. Figure 2 shows an overview of the recommender system stages and data used in them. In the Data Collection and Information Processing phase, the components currently used by the ACT are shown in black. At the same time,

Table 1: Academies of ESCOM-IPN.

Academy	Acronym
Artificial Intelligence	AI
Data Science	DS
Software Engineering	SE
Mathematics and Physics	MTPH
Social Sciences	SS
Distributed Systems	DISTSys
Fundamentals of Electrical Circuits	FEC
Strategic Projects and Decision Making	SPDM
Digital Systems	DSys
Computer Science	CS

Table 2: Thesis proposal sections.

Number	Section	Description
1	Title	Specific name of the proposal that reflects the focus of the work.
2	ID	A unique identifier for the proposal.
3	Abstract	Summary of the content through a concise description.
4	Keywords	Topics involved in the work.
5	Introduction	Context and definition the problem to be solved.
6	Objective	The intended goals of the thesis.
7	Justification	Importance of the topic addressed.
8	Products or expected results	Specific achievements to be reached at the end of the thesis.
9	Methodology	Steps followed to reach the objective.
10	Schedule	Planning of activities to be carried out.
11	References	List of works that support the research and argument.
12	Students and supervisors	Name and a brief description of the thesis authors' areas of interest.

the additional aspects that are being evaluated by the system in order to improve the analysis are indicated in green. This illustrates the system's consideration of a broader range of data and information sources. The subsequent subsections detail the operation of the system stages.

Data collection

One of the first difficulties in developing the system was the lack of a repository to host the thesis proposals previously submitted. The current form of storage is through CDs/DVDs, which contain the proposals in PDF documents. These documents were extracted and stored in a cloud repository. During this process, 1,000 proposals submitted between 2015 and 2023 were collected.

The information about the professors and the academies was collected from different sources. The Administrative Department provided the information about the academies, while the information about the professors comes from the ACT, Administrative Department, previous thesis proposals and academic research websites such as Google Scholar and DBLP.

At the end of this process, data was obtained on 260 professors and 116 courses taught in the 3 careers offered by ESCOM-IPN: Computer Systems Engineering[2,3], Data Science Degree[4] and Artificial Intelligence Engineering[5]. The courses taught in the 3 majors are organised in 10 academies shown in Table 1.

Information extraction and pre-processing

Three datasets were generated with information from thesis proposals, professors and courses taught in the academies. A general cleaning phase was carried out in which only alphanumeric characters were kept and unwanted characters, such as punctuation marks (e.g., periods and commas), special characters and multiple spaces, were removed. A detailed description of the dataset creation process is provided in the following sections.

Thesis proposals dataset

Thesis proposals are stored in PDF documents. According to the regulations established by the ACT, the document must have twelve sections described in Table 2.

After reviewing the content of the documents, it was decided to use sections 1 to 4 and 6 to model the thesis proposals. The content of the selected sections accurately describes what is proposed in the thesis and differentiates it from other proposals.

To extract text from PDF documents we use PyPDF2[6] library. It was expected that all documents were text-based; however, it was found that a large number of these documents were scanned in image format. Due to this, Optical Character Recognition (OCR) techniques were used through the Google Cloud Platform[7] to extract text. Some situations, such as poor image quality and overlapping elements (signatures and stamps), made OCR not to work properly. Figure 3 shows an example of this problem where a stamp overlaps the abstract (Resumen) and introduction (1. Introducci´on) sections. The use of OCR techniques made it possible to increase the number of proposals in the dataset.

[2]https://www.escom.ipn.mx/htmls/oferta/mapaCurrISC2020.php
[3]https://www.escom.ipn.mx/htmls/oferta/mapaCurrISC2009.php
[4]https://www.escom.ipn.mx/htmls/oferta/mapaCurrIIA2020.php
[5]https://www.escom.ipn.mx/htmls/oferta/mapaCurrLCD2020.php
[6]https://pypi.org/project/PyPDF2/
[7]https://cloud.google.com/use-cases/ocr

However, given the difficulties encountered, they were not integrated into the proposed recommender system. The ACT will establish guidelines to ensure that only text-based documents should be used in the recommender system.

Since the thesis proposal follows a defined structure, regular expressions were used to extract the selected sections. Although students adhere to the format defined by the ACT, slight variations were found in the content of the sections. Flexible regular expressions were designed to handle these variations. For example, some proposals only define one general objective, while others specify several specific objectives.

Out of the 1,000 proposals collected, 40 had to be excluded due to incomplete information. The final dataset included 960 proposals.

Professors profile dataset

Professors' information allows us to define profiles that consider their experience in different areas of knowledge. These profiles can be used to establish the degree of similarity with respect to the topics addressed in the thesis proposals. We created a dataset of professor profiles with the information detailed below:

- Professor Name. Full name of the professor.
- Participation of professors as member of TEC. The list of theses in which the professors have participated as supervisor or reviewer provides firsthand knowledge of their experience in thesis projects. Participation as supervisor was obtained from section 12 *Students and supervisors* of thesis proposal (See Table 2) while participation as reviewer was obtained from a file provided by the ACT.
- Academic background: The professors' educational qualifications, institutions attended and any specialized training. This information was also retrieved from section 12 *Students and supervisors* of thesis proposal.
- Areas of interest: This data is valuable for understanding the academic background and interests of the professors. This information was obtained from section 12. *Students and supervisors* of thesis proposal.
- Research works: A list of articles in which the professors have participated. This information helps to understand the professors' interests in research tasks. This data is beneficial for con- sidering professors who have not taught courses related to the most recurrent themes in thesis projects or have not participated in a TEC before. This information was manually collected from Google Scholar (and its redirected links) and DBLP. However, it is important to emphasize that the manual information collection process will no longer be necessary with the URL collected from each academic website. Therefore, the recommender system will automatically update the information on subsequent research work using the URLs.
- Courses instructed: This data was used to know the professors' teaching experience, as it is useful because most of the thesis topics are related to the courses taught at ESCOM-IPN. The information was extracted from a file provided by the Administrative Department.
- Affiliated academy: The academy to which the professors belong is useful, since it allows us to know which theses should be assigned to them. This information was extracted from a file provided by the Academic Department.
- The construction of this dataset was based on identifying each professor and attaching each of the mentioned fields with which they were related. The full name was used to identify the professors uniquely. The problem of variants used in the name was faced during this process. Some examples are described below:
- Different order between surnames and names: In some documents, the order in which a professor's names and surnames were written varied, for example, Juan Pérez, Pérez Juan.
- Names included academic degrees: Some proposals included honorific titles alongside a pro- fessor's name. Example: Dr. Juan Pérez, Lic. Juan Pérez, Mtro. Juan Pérez, Ing. Juan Pérez.
- Spelling mistakes: Some names had spelling mistakes, for example, Juan Peres, Juán Pérez. These details prevented the identification of the same professor, resulting in duplicated rows.
- Therefore, it was decided to clean each name entry following the above process:

1. Convert the entire string to lowercase.

2. Remove accents and mentions of academic degrees using regular expressions.

3. Correct spelling errors manually.

4. Arrange the full names alphabetically. (Example: Juan Pérez and Pérez Juan to Juan Pérez.)

5. After applying the described treatments, better results were obtained, ensuring a unique record per professor.

Academy courses dataset

As was explained in Section Thesis Reviewer Assignment Process At Escom-IPN, the ACT assigns theses to different academies of ESCOM-IPN. Academies constitute the academic structure through which courses are grouped. The dataset com- prises 116 courses across the 10 academies (refer to Table 1) of ESCOM-IPN. These courses are part of the curricular plans for three distinct majors: Computer Systems Engineering, Data Science Degree and Artificial Intelligence Engineering.

The following content was considered for each course:

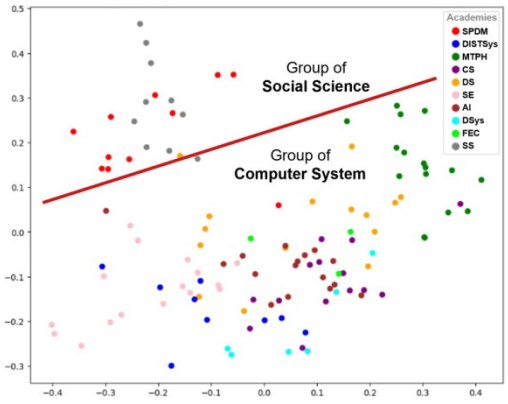


Figure 4: Vector representation of courses with dimensionality reduction from 768 to 2 using the PCA algorithm.

- Course name. Title of the unit of learning.
- Purpose. The overall objective of the course.
- Thematic units. Divides the course into distinct thematic sections.
- Competency of each unit. Specifies the skills, knowledge, or abilities students should gain from each thematic unit.
- Specific content of each unit. Details the topics, concepts, or materials covered within each thematic unit.

The course information was obtained from the PDF documents containing the synthetic programs. The PyPDF2 library and regular expressions were used to extract the aforementioned content.

Information modeling

Word embedding,[21] a technique within natural language processing, entails the numerical vector representation of words or phrases in a continuous numeric space. This method depicts words as vectors of real numbers in a multidimensional space, with the vectors acquired from extensive lin- guistic data. These representations capture both semantic and syntactic relationships among words, ensuring that words with similar meanings possess closely situated vector representations. Prominent techniques for creating word embeddings include Word2Vec, developed by Mikolov *et al.*[22] and others like GloVe[23] and FastText,[24] each with distinct characteristics. The outcome is a set of high-dimensional vectors where each dimension represents a specific aspect of word semantics.

Beyond word-centric embeddings, there exists a category of models known as Sentence Embeddings,[25,26] exemplified by models such as Sentence Transformers,[27] designed to generate vector representations for entire phrases or sentences, capturing the holistic semantic information. Utilizing advanced model architectures like BERT,[28] Sentence Transformers produce embeddings reflecting sentence semantics. These pre-trained models can be fine-tuned or directly applied to specific tasks, such as information retrieval, text summarization, or text classification. Like Word2Vec, embeddings generated by Sentence Transformers allow measuring semantic similarity between sentences, which is very useful for comparing and retrieving documents based on their content.

For our system, the *intfloat/multilingual-e5-base model*[8] from Hugging Face is used, employing the Sentence Transformer library in Python. This model provides a vector representation of dimension 768. An embedding was generated for each element of the proposal, professor and academy courses datasets. The embeddings were stored in a database in binary format.

The embedding generation process occurs in two phases: offline and online. In the offline phase, representations are generated outside the recommendation process, involving the entire training dataset. In the online phase, real-time representations are generated during the reviewer recommendation flow for new thesis proposals. The generation of these representations follows the pre-processing steps explained in Subsection *Information extraction and pre-processing*.

Figure 4 shows the visual representation of the resulting embeddings of the academy courses reduced to 2 dimensions using the PCA algorithm. It is evident that the vectors corresponding to courses in the academies of Social Sciences (SS) and Strategic Projects and Decision Making (SPDM) are distinctly separated from those of computer system courses. This shows that the embeddings are useful for quantifying similarity in terms of the content and objectives of each course.

Evaluation of similarity

This section explains how the similarity between thesis proposals and the academy courses is evaluated, as well as with respect to the professor profiles. Although the goal is to measure the similarity between a proposal and a reviewer profile, we propose two similarity approaches: direct and composite.

Direct similarity

We define a direct similarity as the direct comparison between vector representations of attributes. In order to assess the similarity between the attributes, we employed the cosine similarity metric. By calculating the cosine similarity between the vectors representing the attributes, we can quantify the degree of similarity between two sets of attributes.
The cosine similarity between V and Q is defined in (1)

$$sim_{cos}\left(V, Q\right) = \frac{V \cdot Q}{|V||Q|} \quad (1)$$

where:

- V and Q are vectors of equal dimension
- $\cdot$ denotes the dot product

[8]https://huggingface.co/intfloat/multilingual-e5-base

- || denotes the norm (magnitude) of a vector

We created 4 embeddings, one for each of the following sections extracted from the thesis protocol: Title, Abstract, Keywords and Objective (for more details on these sections see Subsection Thesis proposals dataset). The cosine similarity is calculated for each of the following elements to obtain the direct similarity.

Academy courses: Given the set of courses $C = \{C_1, C_2, ..., C_{116}\}$, each course C_k is represented by five embeddings described in Subsection *Academy courses dataset* while a proposal P by four embeddings. The similarity between a proposal P and a course C_k is given by (2).

$$Sim(P, C_k) = \frac{1}{20}\sum_{i=1}^{4}\sum_{j=1}^{5} sim_{cos}(p_i, c_{kj}) \quad (2)$$

Previous thesis proposals: Since a set of previous proposals $T=\{T_1, T_2, ..., T_{960}\}$, where each previous proposal T_w is represented by four embeddings just like a proposal P therefore, their embeddings were added to integrate a single vector $p\hat{}$ for a proposal and t for a previous proposal w. The similarity between a proposal P and a previous proposal T_w is given by (3)

$$Sim(P, T_w) = sim_{cos}(\hat{p}, t_w) \quad (3)$$

Where

$$\hat{p} = p_1 + p_2 + p_3 + p_4 \quad (4)$$

$$t_w = t_{w1} + t_{w2} + t_{w3} + t_{w4} \quad (5)$$

Academic background and areas of interest of a professor: For convenience, the academic background and areas of interest of a professor are represented by one vector b. The degree of similarity between a proposal P and the academic background and areas of interest of a professor b is given by (6).

$$Sim(P, b) = \frac{1}{4}\sum_{i=1}^{4} sim_{cos}(p_i, b) \quad (6)$$

Professor Research: The research of a professor W is represented by n embeddings. The similarity between P and W is given by (7).

$$Sim(P, W) = \frac{1}{4n}\sum_{i=1}^{4}\sum_{j=1}^{n} sim_{cos}(p_i, w_j) \quad (7)$$

Composite similarity

A professor profile comprises several elements that describe his or her academic career. A composite of these elements will be used to form this profile. Something similar happens with the academies since they comprise various courses. We define composite similarity as the set of direct similarities that make up the similarity between a proposal and a professor profile and between a proposal and an academy, respectively.

Similarity with the professor profile

We obtain the similarity of a thesis proposal to the profile of a professor with the following composition.

Courses instructed: Given the set of direct similarities of academy courses $S\ C = \{Sim(P, C_1), Sim(P, C_2), ..., Sim(P, C_{116})\}$, only the courses instructed by the profe- sor with similarity above or equal to a threshold γ are added to the set $F=\{Sim(P, C1), Sim(P, C2), ..., Sim(P, Ca)\}$ and F is the average of F . The threshold γ is τ -percentile over S_C. The composite similarity of courses instructed by a profesor is described in (8).

$$W_{SimP-CI} = \frac{1}{1+e^{-\gamma \cdot a}} \cdot \overline{F} \quad (8)$$

Participations in TEC as supervisor: Given the set of direct similarities of previous thesis proposals $S_T=\{Sim(P, T_1), Sim(P, T_2), ..., Sim(P, T_{960})\}$, only the previous thesis in which the professor participated as a supervisor with similarity above or equal to a threshold β are added to the set $G = \{Sim(P, T_1), Sim(P, T_2), ..., Sim(P, T_k)\}$, G is the average of G. The threshold β is τ -percentile over S_T . The composite similarity of participations in TEC as supervisor is described in (9)

$$W_{SimP-Supervisor} = \frac{1}{1+e^{-\beta \cdot k}} \cdot \overline{G} \cdot \frac{k}{c} \quad (9)$$

where:

- c: total number of participations of the professor as a supervisor
- β: factor between 0 and 1 that determines how quickly $W_{SimP-Supervisor}$ adjusts in response to changes in k

Participations in TEC as reviewer: Given the set S_T, only the previous thesis in which the professor participated as a reviewer with similarity above or equal to a threshold β are added to the set $H = \{Sim(P, T_1), Sim(P, T_2), ..., Sim(P, T_d)\}$, H is the average of H. The composite similarity of participations in TEC as reviewer is described in (10).

$$W_{SimP-Reviewer} = \frac{1}{1+e^{-\lambda \cdot d}} \cdot \overline{H} \cdot \frac{d}{l} \quad (10)$$

where:

- l: total number of participations as reviewer
- λ: factor between 0 and 1 that determines how quickly $W_{SimP-Reviewer}$ adjusts in response to changes in d

Finally, the composite similarity between a proposal and a professor profile is calculated by the following equation:

Algorithm 1 Professor similarity ranking$(p, S_{prof}, C_{sim}, PT_{sim})$

Require: thesis proposal p, set of professors S_{prof}, set of courses similarities C_{sim}, set of previous theses similarities PT_{sim}
Ensure: ranked set of reviewer candidates R_p
$p_e \leftarrow$ get_embeddings(p)
$R_p \leftarrow \varnothing$
for each s_i in S_{prof} **do**
    #Composite proposal-proffesor profile similarity:
    $Sim_{p-s_i} \leftarrow$ calculate $Sim_{\text{P-Profile}}(p_e, s_i, C_{sim}, PT_{sim})$ **Add** (s_i, Sim_{p-s_i}) into R_p
end for
sort R_p descending by $Sim_{\text{P-Profile}}$
return R_p

$$Sim_{P-Profile} = \frac{1}{5} \cdot (W_{SimP-CI} + W_{SimP-Supervisor} + W_{SimP-Reviewer} + Sim(P,b) + Sim(P,W)) \quad (11)$$

Similarity to academy courses

To obtain the similarity between a thesis proposal and an academy A, we rely on the similarity to its courses. The content of these courses allows us to identify the areas of knowledge of the academies. Given the set S_c, only the courses that belong to academy A with similarity above or equal to a threshold δ are added to the set $I = \{Sim(P, C_1), Sim(P, C_2), ..., Sim(P, C_g)\}$. The similarity is given by (12)

$$W_{\text{Sim P-A}} = \frac{1}{1 + e^{-\delta \cdot g}} \cdot \bar{I} \quad (12)$$

Recommendation process

The recommendation process determines the set of professors most closely related to a thesis proposal and selects three to form the review committee.

Algorithm 2 Academy similarity ranking(p, A, C_{sim})

Require: proposal p, set of Academies A, set of Courses similarities C_{sim}
Ensure: ranked set of academies similarities A_p
$A_p \leftarrow \varnothing$
$p_e \leftarrow$ get_embeddings(p)
for each a_i in A **do**
    #Composite proposal-academy similarity
    $Sim_{p-a_i} \leftarrow$ calculate $Sim_{\text{P-A}}(p_e, a_i, C_{sim})$
    Add (a_i, Sim_{p-a_i}) into A_p
end for
sort A_p descending by $Sim_{\text{P-A}}$
return A_p

Reviewer candidates

The first step in the recommendation process is to obtain the similarity ranking between the thesis proposals and the set of professors. The top-ranked professors will be considered as reviewer candidates. The process is described in Algorithm 1.

Committee selection

The next step in the recommendation process is to select the members of TEC. The previously ordered reviewer candidates must comply with the following guidelines established by the ACT:

- TEC must be made up of 3 reviewers.

Algorithm 3 TEC recommendation process

Require: proposal p, set of professor S_{prof}, set of Academies A, set of Courses C_{set}, set of previous theses T_{set}

Ensure: committee recommended TEC_1 (approach academies), committee recommended TEC_2 (approach professors)

$TEC_1 \leftarrow \varnothing$
$TEC_2 \leftarrow \varnothing$
$A_1 \leftarrow \varnothing$
$A_2 \leftarrow \varnothing$
$p_e \leftarrow$ get_embeddings(p)
#Direct proposal-course similarity for the whole set of courses
$C_{sim} \leftarrow$ calculate Sim(p_e, get_embeddings(c)) for all c in C_{set}
#Direct proposal-previous similarity for the whole set of previous proposals
$PT_{sim} \leftarrow$ calculate Sim(p_e, get_embeddings(t)) for all t in T_{set}
$R_p \leftarrow$ Professor similarity ranking($p, S_{prof}, C_{sim}, PT_{sim}$)
#Recommendation focused on the professor profile:
for each s, sim in R_p **do**
    $sA \leftarrow$ get_professor_academy(s)
    if s **Not** Supervisor on p **then**
        if $len(TEC_1) < 3$ **then**
            if $A_1[sA] < 2$ **then**
                $A_1[sA] = A_1[sA] + 1$
                Add (s, s_A) into TEC_1
            end if
        end if
    end if
end for
#Recommendation focused on academies:
$TEC_2 \leftarrow$ Top3(Academy similarity ranking(p, A, C_{sim}))
return TEC_1, TEC_2

- There must be a maximum of 2 reviewers from the same academy.
- Supervisors may not participate as a reviewer on the same thesis.

As described in Section 3, the current ACT process does not directly assign reviewers. Instead, proposals are assigned to academies. Considering the above, two approaches are proposed for the selection of TEC members.

Recommendation focused on academies: For this approach we rely on the current ACT allocation process, in which a proposal is distributed through the academies. The process of obtaining academy similarities is described in Algorithm 2.

Recommendation focused on the professor profile: This approach considers the similarity ranking of the candidates' profiles and the constraints of the ACT. The best-ranked profiles are selected and then the professors' academies are obtained to form the TEC. Algorithm 3 shows the process for making recommendations, following the academy-based and professor profile-based approaches.

EXPERIMENTS AND RESULTS

This section describes the experiments carried out with the proposed recommendation system. The experimentation was divided into two stages. In the first stage, called training, the embeddings of the elements extracted from the datasets described in Subsection *Information extraction and pre-processing* were obtained. The second phase, called testing, consisted of applying the similarity metrics described in Subsection *Evaluation of similarity* to a set of thesis proposals not used in the training phase and generating the recommendations of the academies. Table 3

Table 3: Datasets used for evaluation of the recommender system.

Stage	Dataset	Total
Training	Thesis proposals	960
	Professor profiles	260
	Academy courses	116
Testing	Thesis proposals	93

Table 4: Results obtained in the test set.

τ -percentile values	50	60	70	80	90	95
Academy approach	0.458	0.458	0.461	0.463	0.491	0.465
Professor profile approach	0.476	0.473	0.462	0.458	0.551	0.501

Table 5: Count by numbers of academy matches.

Number of academy matches for a proposal	0	1	2	3
Count by academy approach	8	41	36	8
Count by professor profile approach	6	32	43	12

shows the data used for training and testing and the number of items in each dataset.

Evaluation method

As explained above, the ACT merely decides to which academies thesis proposals should be sent so that academy professors select the ones they are interested in to act as reviewers. The proposed system recommends a TEC, but given that the ACT has no influence on the final confirmation of the TEC, nor is there anyone responsible for verifying and validating its conformation, the comparison between the recommended TECs and those that were actually conformed does not seem to us to be feasible. Therefore, the evaluation of the system was carried out by comparing the academies selected by the ACT and those recommended by the system. Despite this consideration, the recommendations made by the system in the academy-based approach make for more in-depth reviews of the information collected. Even in the professor profile approach, which considers many more aspects, the academies selected are those to which the highest-ranked professor belongs.

The proposed evaluation metric measures the matches between the academies selected by the ACT and those recommended by the system. We call this metric Similarity Metric for Academy Recommendation (SMAR) and is defined in Equation 13.

$$\text{SMAR} = \frac{|R \cap A|}{3} \qquad (13)$$

- R is the set of academies recommended by the system.
- A is the set of academies selected by the ACT.

RESULTS

For the experiments, different values of the parameter τ -percentile were tested. This parameter helps to regulate the level of similarity between the proposals and the profiles of the professors or academies, depending on the approach used. In Table 4 shows the average SMAR obtained by the recommender system in both approaches. We observed that with a variation in the percentile parameter, the average SMAR varies. Also, it is noted that with the 90 – *percentile* a maximum SMAR average is reached for both cases, which tells us that this value is a sufficiently strict threshold to filter out the most similar items but not enough to omit similar of them.

To further examine the performance of the recommender, we calculated the count of proposals in which the system had none, one, two or three academy assignments equal to those made by the ACT in each proposal. Table 5 shows the results of the counting. As can be seen, the recommendations made by the system coincide in at least 80% of the proposals, in both approaches, with 1 or 2 of the academies selected by the ACT. Figure 5 clearly shows the trend of overlap from one to two academies with respect to the assignments made by the ACT. Here, it can also be seen that the extremes 0 and 3 are the least frequent. This result is something that we expected since the information used by the ACT to select the academies is a subset of the information that the proposed system considers. Falling mostly into the count 0 would indicate that the information used by the system is not related to that used by the ACT while obtaining a count of 3 in most cases would

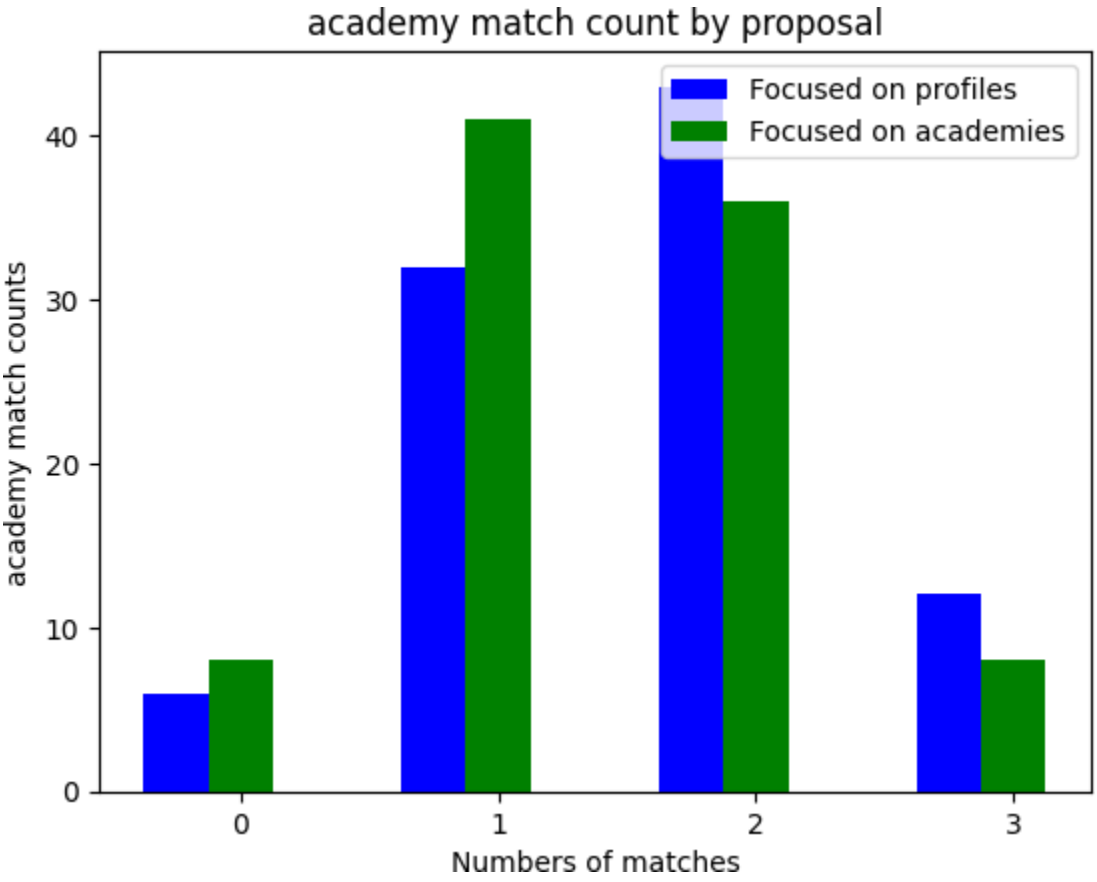


Figure 5: Comparison of academy match counts between approaches.

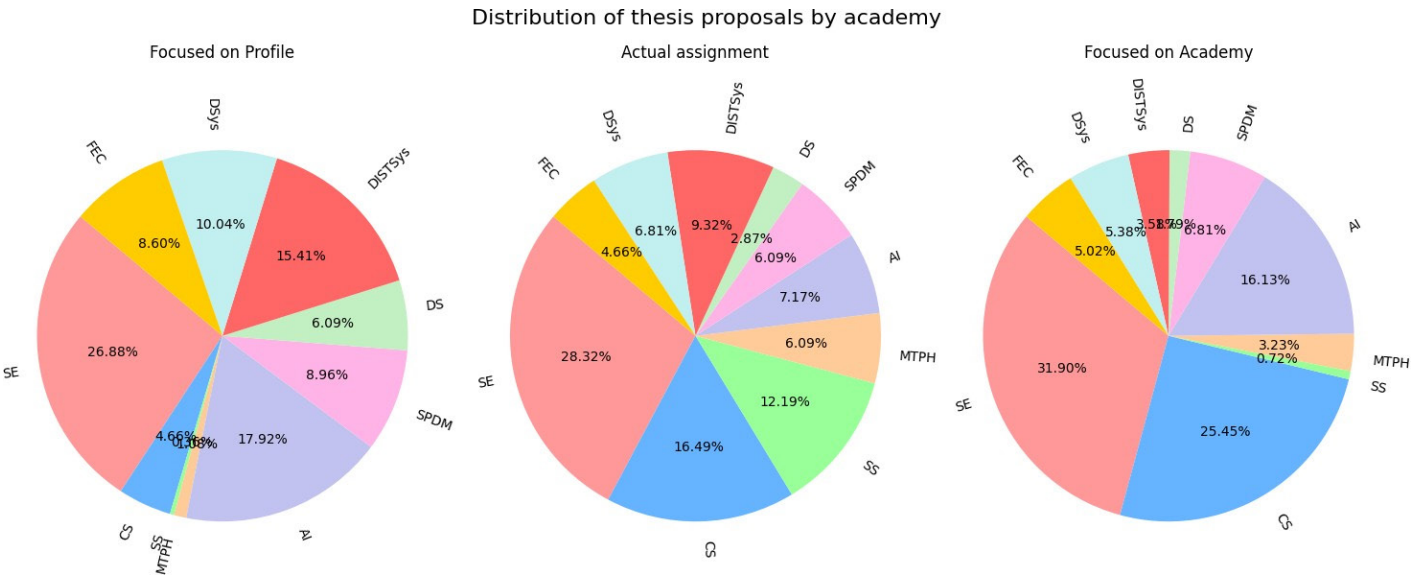


Figure 6: Distribution of thesis proposals by Academy.

indicate that the additional information used by the system does not provide elements to make a selection different from that made by the ACT.

Another revealing aspect to consider in the results is the system's distribution of thesis proposals by academy. Currently the ACT takes some considerations to distribute the thesis proposals in an equitable way among the different academies. This is so that all ESCOM-IPN professors have the opportunity to participate as reviewers. This has as a consequence that some assignments made by the ACT are not based on adequate profiles but on administrative demands. In spite of this effort the assignments still show a bias towards academies with courses related to computer science areas. This can be seen in Figure 6, in the Actual

assignment graph. Since our recommendation system does not consider these administrative aspects, but only the similarities between the departments and the professors' profiles, it is to be expected that the academies less related to computer science (SS and MTPH) will be the least recommended by the system. In addition, these are the academies that have historically proposed the fewest thesis projects. This can be seen in Figure 6, Focused on Profile and Focused on academy graphs.

CONCLUSION AND FUTURE WORK

The reviewer assignment problem is interesting given its wide range of applications. Recommender systems have supported the solution of this problem and are an important approach to consider. In this work, the implementation of a recommender system for the selection of thesis proposal reviewers was proposed. The process carried out at ESCOM-IPN was taken as a case study. After collecting information from different sources, a recommender system was trained to suggest a thesis review committee made up of the professors most closely related to the thesis proposal to be evaluated. Given the particularities of the process currently carried out at ESCOM-IPN, the system generates recommendations of academies where the appropriate professors are found to review the work. The results show that the system takes advantage of the additional information provided to make recommendations for professors with profiles more related to the proposals. This can be seen in the distribution that it makes of the theses by academies, with those with themes more related to the thesis proposals prevailing. Finally, the coincidences of the recommendations concerning the ACT assignments allowed us to verify that the system uses part of the information considered by the ACT but can generate different recommendations. For future work, it is planned that the ACT will carry out an evaluation of the professor recommendations made by the system in order to have additional validation of the results obtained. It is also planned to test other information modeling techniques such as topic extraction and with these evaluate the similarity between the proposals and the professors.

ACKNOWLEDGEMENT

This research was funded by CONACyT-SNI and Instituto Politécnico Nacional (IPN), through grants SIP-20240979, SIP-20240974 and EDI.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

RAP: Reviewer Assigment Problem; **NLP:** Natural Language Processing; **TEC:** Thesis Evaluation Committee; **ACT:** Academic Thesis Committee; **OCR:** Optical Character Recognition; **SMAR:** Similarity Metric for Academy Recommendation.

REFERENCES

1. Aksoy M, Yanik S, Amasyali MF. Reviewer assignment problem: A systematic review of the literature. J Artif Intell Res. 2023;76:761-827. doi: 10.1613/jair.1.14318.
2. Hoang DT, Nguyen NT, Collins B, Hwang D. Decision support system for solving reviewer assignment problem. Cybern Syst. 2021;52(5):379-97. doi: 10.1080/01969722.2020.1871227.
3. Janak SL, Taylor MS, Floudas CA, Burka M, Mountziaris TJ. Novel and effective integer optimization approach for the nsf panel-assignment problem: A multiresource and preference-constrained generalized assignment problem. Ind Eng Chem Res. 2006;45(1):258-65. doi: 10.1021/ie050478h.
4. Chughtai GR, Lee J, Shahzadi M, Kabir A, Hassan MA. An efficient ontology-based topic-specific article recommendation model for best-fit reviewers. Scientometrics. 2020;122(1):249-65. doi: 10.1007/s11192-019-03261-2.
5. Cechĺarová K, Fleiner T, Potpinková E. Assigning evaluators to research grant applications: the case of Slovak research and development agency. Scientometrics. 2014;99(2):495-506. doi: 10.1007/s11192-013-1187-1.
6. Di Mauro N, Basile TM, Ferilli S. Grape: an expert review assignment component for scientific conference management systems. Innov Appl Artif Intell. 2005:789-98.
7. Tan S, Duan Z, Zhao S, Chen J, Zhang Y. Improved reviewer assignment based on both word and semantic features. Inf Retrieval J. 2021;24(3):175-204. doi: 10.1007/s10791-021-09390-8.
8. Saptono R, Setiadi H, Sulistyoningrum T, Suryani E. In: Examiners recommendation system at proposal seminar of undergraduate thesis by using content-based filtering International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE PUBLICATIONS; 2018. p. 295-9.
9. Dumais ST, Nielsen J. Automating the assignment of submitted manuscripts to reviewers. In: Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval; 1992. p. 233-44. doi: 10.1145/133160.133205.
10. Pradhan T, Sahoo S, Singh U, Pal S. A proactive decision support system for reviewer recommendation in academia. Expert Syst Appl. 2021;169:114331. doi: 10.1016/j.eswa.2020.114331.
11. Yang C, Liu T, Yi W, Chen X, Niu B. Identifying expertise through semantic modeling: A modified bbpso algorithm for the reviewer assignment problem. Appl Soft Comput. 2020;94. doi: 10.1016/j.asoc.2020.106483.
12. Kalmukov Y. An algorithm for automatic assignment of reviewers to papers. Scientometrics. 2020;124(3):1811-50. doi: 10.1007/s11192-020-03519-0.
13. Mirzaei M, Sander J, Stroulia E. Multi-aspect review-team assignment using latent research areas. Inf Process Manag. 2019;56(3):858-78. doi: 10.1016/j.ipm.2019.01.007.
14. Kobren A, Saha B, McCallum A. Paper matching with local fairness constraints. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery; 2019. p. 1247-57. doi: 10.1145/3292500.3330899.
15. Stelmakh I, Shah NB, Singh A. Peerreview4all: fair and accurate reviewer assignment in peer review. In: Proceedings of the 30th international conference on algorithmic learning theory; 2019. PMLR. p. 828-56.
16. Anjum O, Gong H, Bhat S, Hwu WM, Xiong J. PaRe: A paper-reviewer matching approach using a common topic space; 2019. doi: 10.18653/v1/D19-1049.
17. Jin J, Niu B, Ji P, Geng Q. An integer linear programming model of reviewer assignment with research interest considerations. Ann Oper Res. 2020;291(1-2):409-33. doi: 10.1007/s10479-018-2919-7.
18. Liu X, Wang X, Zhu D. Reviewer recommendation method for scientific research proposals: a case for nsfc. Scientometrics. 2022;127(6):3343-66. doi: 10.1007/s11192-022-04389-4.
19. Yong Y, Yao Z, Zhao Y. Beyond accuracy: a feature crossing method for Chinese thesis reviewer recommendation. In: IEEE International Conference on Systems, Man and Cybernetics (SMC). Vol. 2021. IEEE PUBLICATIONS; 2021. p. 1151-8. doi: 10.1109/SMC52423.2021.9658668.
20. Budiprasetyo G, Kirana AP, Mentari M, Pratama DC. Recommendation system for thesis examiner selection using intuitionistic fuzzy topsis method for effective multicriteria decision-making. In: International Conference on Electrical and Information Technology (IEIT). IEEE PUBLICATIONS; 2021. p. 172-8. doi: 10.1109/IEIT53149.2021.9587449.
21. Mars M. From word embeddings to pre-trained language models: A state-of-the-art walk-through. Appl Sci. 2022;12(17). doi: 10.3390/app12178805.
22. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 2013.
23. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
24. Bojanowski, Grave P, E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comp Linguist. 2017;5:135-46. doi: 10.1162/tacl a 00051.
25. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 conference on empirical methods in natural language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 670-80. doi: 10.18653/v1/D17-1070.

26. Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, *et al.* Universal sentence encoder; 2018. p. 1803.11175. doi: 10.48550/arXiv.1803.11175 . https://doi.org/10.48550/arXiv.
27. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint Conference on natural language processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3980-90. doi: 10.18653/v1/D19-1410.
28. Kenton JD, Toutanova LK. Bert: pretraining of deep bidirectional transformers for language understanding. In: Proceedings of the naacL-HLT. Vol. 1; 2019. p. 2. doi: 10.18653/v1/N19-1423.

Cite this article: Gambino OJ, García-Mendoza CV, Suárez-Bautista JM, Gu H, Calvo H. Thesis Reviewer's Committee Recommendation System. J Scientometric Res. 2025;14(1):351-64.