Mapa de coneixement de la UAB. Modelat i anàlisi de la producció científica basat en grafs de coneixement

Ariadna De Vicente Viladesau

Resum– El Portal del Coneixement de la UAB (Knowledge Map UAB) és una eina que té per objectiu gestionar la informació sobre el personal investigador, els grups de recerca i professorat del centre per tal de proveir una col·lecció de serveis i utilitats, que permet a les empreses fer consultes específiques i rebre resultats precisos en funció de les seves necessitats de recerca. Es tracta d'un nucli de gestió de dades amb estructures de grafs de coneixement, sobre les quals es poden aplicar algorismes d'intel·ligència artificial i de dades.

Paraules clau– Intel·ligència Artificial · Algoritmes de Detecció de Comunitats · Similitud de Documents · Sistema de Recomanació · Grafs de Coneixement.

abstract– The Knowledge Map UAB is a tool aimed at managing information about research staff, research groups, and faculty of the center to provide a collection of services and utilities that allow companies to make specific queries and receive precise results based on their research needs. It is a core data management system with knowledge graph structures, on which Artificial Intelligence and data algorithms can be applied.

Keywords– Artificial Intelligence · Community Detection Algorithms · Document Similarity · Recommender System · Knowledge Graphs.

1 INTRODUCCIÓ

En el context actual, les tecnologies de la informació i la comunicació tenen un paper fonamental en la societat, afectant múltiples àmbits de la vida quotidiana. És innegable que el seu impacte és especialment rellevant en l'àmbit acadèmic, on hi ha un enorme ventall de tècniques per a l'analítica de dades, gestió de la informació i el desenvolupament de noves formes de col·laboració i generació de coneixement.

Motivada per l'impacte que tenen les tecnologies en l'àmbit acadèmic i els desafiaments que aquest panorama en constant evolució presenta, he emprès aquest Treball de Fi de Grau (TFG) amb l'objectiu de contribuir al desenvolupament i l'avanç de les tecnologies a la Universitat Autònoma de Barcelona.

Considerant la necessitat d'eines eficients per gestionar la informació sobre les competències tècniques del seu personal investigador i grups de recerca, va néixer el Portal de Recerca de la UAB el qual és una eina per a fer visible la productivitat científica del personal investigador de la UAB. Aquest Portal és un repositori del coneixement generat a la universitat, que identifica els àmbits més productius i innovadors, les sinergies entre grups, i l'oferta científica i tecnològica que es pot compartir amb l'exterior. El Portal s'alimenta de múltiples fonts de dades, com ara les bases de dades internes de la universitat, portals científics externs i altres fonts d'informació rellevants.

El present TFG proposem el Portal de Coneixement, el qual és un sistema dinàmic i interoperable que emmagatzema informació sobre la producció de coneixement, les competències i la demanda externa per tal que serveixi com a suport.

Aquest sistema, no només permet la connexió interna entre el personal, sinó que també facilita la col·laboració entre diferents departaments i empreses externes, a través d'un sistema de recomanació i eines d'intel·ligència artificial. Aquestes eines han de ser capaces d'identificar interconnexions entre nodes d'informació, a partir de paraules clau específiques generant un nucli de gestió de dades basat en estructures de grafs de coneixement.

La font primària de dades en la que ens hem basat són les publicacions, projectant-les en un espai de "temes"a partir dels quals construïm el graf de coneixement. Per tant, es

• E-mail de contacte: aridevicente@gmail.com
• Treball tutoritzat per: Josep Lladós Canet (Computer Science)
• Curs 2022/23

necessita desenvolupar un sistema [1] que pugui determinar la similitud entre les publicacions [2][3] per unir persones de diferents departaments que tenen coneixements similars.

El Portal del Coneixement és una eina clau per a la gestió de dades de la Universitat, oferint una àmplia gamma de serveis i utilitats a empreses que busquen solucions innovadores i efectives. A més, pot millorar l'eficàcia de la recerca a la UAB i promoure el desenvolupament de solucions innovadores per a problemes socials i econòmics. El TFG té com a objectiu reforçar el Portal de Recerca de la UAB per millorar la col·laboració entre departaments i identificar competències similars en el personal investigador.

2 ESTAT DE L'ART

Atès que hem escollit representar les dades amb un graf de coneixement, les tècniques utilitzades que ens han permès construir l'analítica són la detecció de comunitats, els grafs de coneixement, els sistemes de recomanació i la similitud entre documents són àrees d'investigació molt rellevants en el camp de la informàtica i la intel·ligència artificial.

En aquesta secció presentem els algorismes més destacats, així com les tècniques més recents i avançades usades per abordar aquests problemes.

2.1 Grafs de coneixement

Els grafs de coneixement permeten representar el coneixement d'una forma estructurada, clara i organitzada, la qual cosa fa que la informació resulti més entenedora. En definitiva, són una eina que serveix de suport a l'hora de resoldre problemes, prendre decisions, trobar grups o conjunts de nodes dins les dades entre altres coses.

Un dels treballs que ha tingut més influència sobre la representació del coneixement és el treball conegut com "DBpedia: A Nucleus for a Web of Open Data" [4]. Resulta molt influent en el desenvolupament de grafs de coneixement, ja que descriu la creació i evolució de DBpedia, una base de coneixements basada en la Wikipedia. L'artícle va ser publicat l'any 2007.

També hi ha llibres que han marcat una diferència en l'actualitat com per exemple el llibre "Knowledge Graphs and Big Data Processing" [5]. Aquesta obra tracta de la teoria i la pràctica dels grafs de coneixement, que són útils per processar grans quantitats de dades. Es fa especial èmfasi en la representació del coneixement relacionat amb persones, documents i publicacions, així com en altres temes d'interès.

2.2 Similitud de documents

Al llarg dels anys, s'han anat implementant diferents algoritmes i tècniques per a determinar la similitud de documents, entre elles, s'han descobert mètriques per trobar similitud entre texts com *K-means*, *Word2Vec*, LSI, LDA, Bert embeddings entre d'altres [6] les quals són tècniques fonamentals en el processament del llenguatge natural, ja que, permet identificar la proximitat semàntica entre textos i trobar patrons rellevants en grans volums d'informació. També s'han desenvolupat classificadors de textos amb xarxes neuronals i Tensorflow. La classificació de textos amb xarxes neuronals i Tensorflow [7] implica entrenar el model amb un conjunt de dades d'entrenament etiquetades, i després utilitzar aquest model per predir la categoria o etiqueta dels nous textos que es volen classificar.

2.3 Algoritmes de detecció de comunitats

L'any 2012, Lei Tang i Huan Liu van escirure un llibre que parla sobre algoritmes de detecció de comunitats conegut com 'Community Detection and Mining in Social Media' [8]. Aquest llibre explora les tècniques més avançades per detectar comunitats en xarxes socials i altres tipus de grafs. També inclou una revisió exhaustiva de les metodologies i aplicacions actuals, així com les últimes tendències en la detecció de comunitats.

2.4 Sistemes de recomanació

Un llibre que resulta molt interessar és 'Item-based collaborative filtering recommendation algorithms' [9] de Badrul M. Sarwar, George Karypis, Joseph A. Konstan i John Riedl. Aquest llibre cobreix diferents tècniques de recomanació així com les seves aplicacions en diferents àmbits. També ofereix una visió general de les dades que es poden utilitzar per a la recomanació i les mètriques per avaluar els sistemes de recomanació.

3 OBJECTIUS

Els objectius d'aquest Treball de Fi de Grau són:

• Objectius generals: Realitzar millores i afegir una capa augmentant la intel·ligència del Portal de Recerca de la UAB per tal que sigui una eina per a millorar la col·laboració entre departaments i la identificació de competències similars en el personal investigador.

• Objectius específics:

- Desenvolupar l'eina per a la gestió de les competències tècniques del seu personal investigador i grups de recerca.

- Que l'eina permeti l'emmagatzematge i la gestió de dades de manera eficient i escalable.

- Emprar una estructura de grafs que permeti la representació i emmagatzematge de tota la informació.

- Facilitar la connexió entre el personal investigador i grups de recerca, tant internament com externament, per fomentar la col·laboració i la creació de solucions innovadores.

- Identificar els grups de recerca dins de l'estructura del graf de coneixement per detectar comunitats i fomentar la col·laboració entre els diferents departaments.

- Incorporar eines d'intel·ligència artificial per a la creació d'un sistema de recomanació, que permeti identificar les persones més adients per a una empresa que busqui solucions a un problema determinat.

- Creació d'un sistema de cerca avançada per accedir ràpidament a la informació emmagatzemada en el Graf de Coneixement.

- Realització d'una eina de visualització que permeti la representació gràfica dels nodes d'informació i les seves interconnexions.

4 METODOLOGIA

La metodologia que s'ha seguit ha estat la següent:

• Estudi de l'estat de l'art: es realitzarà una revisió bibliogràfica i anàlisi de les eines similars al Portal de Recerca i el seu ús en la gestió de dades en entitats acadèmiques.

• Requisits: S'identificaran les principals característiques i funcionalitats del Portal de Recerca de la UAB.

• Identificació de les fonts de dades: es determinaran les bases de dades internes (EGRETA) i externes, per obtenir una visió global i actualitzada de la informació per a la creació del sistema.

• Disseny del model: S'elaborarà un model que permeti l'eficient i escalable emmagatzematge i gestió de les dades, així com la categorització dels nodes de la xarxa de coneixement.

• Categorització i connexió de nodes: S'utilitzaran algorismes de data mining per categoritzar els nodes del graf i detectar connexions entre persones. I es faran servir paraules clau per identificar les interconnexions entre els nodes d'informació.

• Implementació i entrenament del model: Es desenvoluparà i entrenarà el model usant les dades recopilades, així com s'avaluarà la seva eficàcia i precisió.

• Integració amb altres sistemes: Es realitzarà la integració del sistema de coneixement amb altres sistemes de la universitat, com ara bases de dades internes i externes per aconseguir una visió global i actualitzada de la informació.

• Verificació i validació: S'avaluarà el sistema de coneixement i es verificarà que compleix amb els objectius establerts, així com es validarà el seu funcionament en un entorn real.

• Visualització: Es durà a terme un entorn per l'usuari per tal que pugui veure els resultats de forma clara.

Pel que fa a la planificació del treball que és un aspecte clau per aconseguir amb èxit els objectius plantejats. És important dedicar el temps i esforç necessaris, definir clarament l'objectiu i abast del treball, establir les tasques i les dates límit tenint en compte els possibles problemes que puguin sorgir.

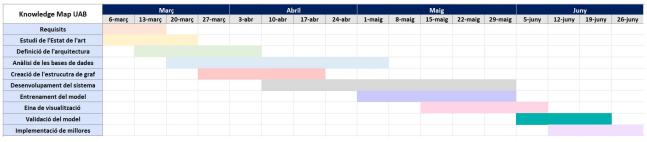


Fig. 1: Diagrama de Gantt per gestionar el temps i les tasques del projecte.

En relació a l'estat de l'art, hem fet una cerca exhaustiva sobre el context històric. Pel que fa a l'arquitectura del sistema, s'han definit dissenys i discutit quina estructura és la més adient per una bona implementació del model. S'ha dedicat gran part del temps a fer una anàlisi exhaustiu dels diferents conjunts de dades amb els que treballarem i això ens ha permès trobar la informació més redundant. Tal com podem observar en el Diagrama de Gantt a la figura 1, també hem realitzat la creació de l'estructura del graf, creant els nodes i arestes necessaris. Paral·lelament a la creació del graf de coneixement, també hem començat el desenvolupament del sistema, aplicant tècniques de Machine Learning per trobar similituds entre documents. També ens hem centrat en el desenvolupament i entrenament del sistema per a obtenir millors resultats. Per acabar, s'ha executat la tasca de visualització prevista juntament amb la creació d'un entorn per a l'usuari.

En termes generals, s'han acomplert totes les tasques proposades inicialment tot i que cal destacar, que la part de desenvolupament ha suposat més hores de les previstes.

5 DESENVOLUPAMENT

El desenvolupament d'aquest Treball de Fi de Grau (TFG), es basa en els objectius i la metodologia establerts prèviament. A continuació, s'explica com s'ha abordat cada aspecte per aconseguir els resultats esperats.

En primer lloc, s'ha dut a terme un estudi de l'estat de l'art en el camp de la gestió de dades i les eines similars al Portal. Això ha permès comprendre el context històric i conèixer les diferents tècniques. Aquest coneixement ha servit com a base per a l'elaboració de l'estratègia de desenvolupament del sistema.

Un pas fonamental ha estat identificar els requisits i les funcionalitats principals del Portal. Aquesta etapa ha implicat una anàlisi exhaustiu de la base de dades de la UAB coneguda com a EGRETA[1].

Elaborant aquesta anàlisi del Portal hem vist que aquest conté la informació de perfils d'investigadors, unitats de recerca, publicacions, projectes, fingerprints i xarxa de relacions. El Portal mostra tota aquesta informació continguda a EGRETA de forma que es faci visible el treball de la comunitat investigadora de la UAB. A continuació es detalla el desenvolupament d'un sistema de cerca avançada.

5.1 Anàlisi del conjunt de dades

L'anàlisi de les dades és una part molt important a considerar, ja que, ens permet detectar les característiques principals que descriuen les dades amb les quals treballarem. Realitzant aquest procés d'inspecció, neteja i transformació de les dades, podrem destacar la informació essencial i més útil.

Hem treballat amb vàries bases de dades diferents. Amb la base de dades del Centre de Visió per Computador (CVC) la qual ha servit per a fer un prototip i primeres versions del sistema per validar el seu bon funcionament. Aquest està compost per taules de persones, publicacions i projectes.

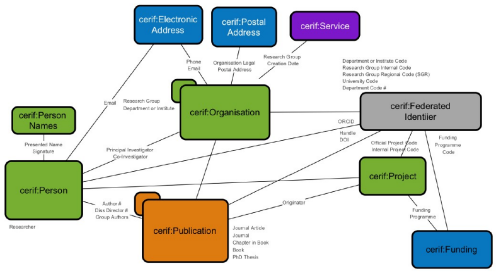


Fig. 2: Esquema de la Base de Dades del CVC.

[1] https://www.uab.cat/web/investigar/itineraris/eines-i-suport-a-la-recerca/egreta-entorn-per-a-la-gestio-de-la-recerca-i-la-transferencia-1345808184219.html

Com podem veure en la figura 2, la base de dades del CVC està composta per la informació de persones en la taula cerif: Person que té un total de 86 registres. Les persones tenen atributs com cfPersId, cfPersEAddr, cfFirstNames, cfFamilyNames, unitatOrg. També té informació sobre les publicacions emmagatzemades a la taula cerif: Publication amb un total de 2.546 registres. Les publicacions en canvi, tenen atributs com url, title, *abstract*, *keywords*, autors... Els registres de la taula cerif: Organisation donen la informació sobre els autors de les publicacions.

D'altra banda, la segona base de dades amb la que hem treballat és la que ens proporciona la UAB, coneguda com l'Entorn per a la Gestió de la Recerca i la Transferència (EGRETA). EGRETA, és un programari versàtil, interoperable i configurable que permet gestionar la investigació des de l'inici fins al final del seu cicle de vida, mantenint actualitzades les dades de publicacions i altres activitats de recerca realitzades pel personal investigador.

Aquesta base de dades està composta per taules com Persons, ResearchOutputs, Projects, Organisations, ExternalOrganisations, ExternalPersons, AuthorCollaborations, StudentTheses, Publishers, Courses, Datasets, Equipments, Activities, Journals... A continuació, a la figura 3, mostrem una part de l'esquema de les taules de la UAB; a l'annex s'ha adjuntat l'esquema complet.

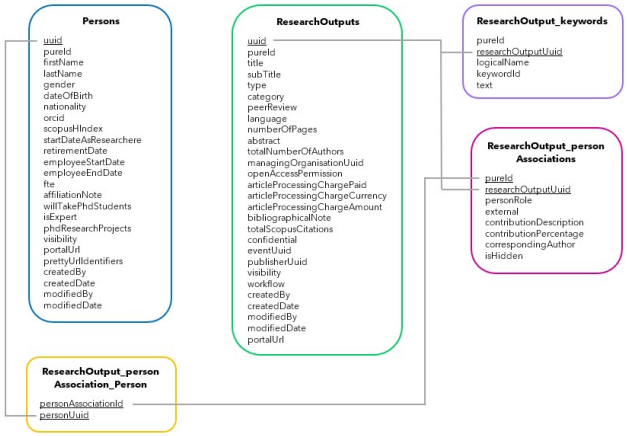


Fig. 3: Esquema d'EGRETA, la Base de Dades de la UAB.

Com podem veure en la figura 3, les dades que utilitzarem per al desenvolupament són els registres de persones de la taula Persons que té un total de 22.819 registres i 26 camps. Aquests camps de la taula persones són atributs com uuid, pureId, firstName, lastName, gender, dateOfBirth, nationality, orcid, scopusHIndex, retirementDate, portalUrl entre d'altres.

Pel que fa a la taula ResearchOutputs, fa referència a les publicacions entre els anys 2019 i 2023, la qual té un total de 181.791 registres i 30 camps. Les publicacions tenen atributs com uuid, pureId, title, subTitle, type, category, Language, numberOfPages, *abstract*, totalNumberOfAutors, createdDate entre d'altres.

La informació sobre les Paraules Clau de les publicacions està a la taula ResearchKeywords la qual té un total de 99.421 i 6 camps. Aquesta taula està relacionada amb la de les publicacions a través del uuid de la publicació.

Per acabar, per tal de lligar els autors amb les publicacions són necessàries la taula ResearchOutputPersonAssociation i ResearchOutputPersonAssociationPerson. La taula ResearchOutputPersonAssociation amb 1.009.736 registres i 10 camps conté el reserachOutputUuid i el pureID el qual s'associa al camp personAssociationId de la taula ResearchOutputPersonAssociationPerson en la qual també hi ha el personUuid la qual té 244.296 registres i 3 camps.

Un cop analitzades les dades, pel fet que hi ha dades Nan, el que fem és netejar el conjunt de dades amb les dades que trobem realment necessàries.

5.2 Creació dels nodes del graf

Un cop elaborada l'anàlisi de les dades, després d'haver seleccionat els atributs necessaris, ens disposem a fer la creació dels nodes que formaran part del graf. Un graf és una representació visual d'informació que consisteix en nodes i arestes. Els nodes són punts o elements individuals, mentre que les arestes són les connexions o vincles que s'estableixen entre els nodes. Mitjançant aquesta estructura, els grafs permeten representar i visualitzar les relacions i interaccions entre els diferents elements de manera clara i concisa.

Per fer-ho, hem creat un graf a l'entorn de Neo4j[2]. Neo4j és una base de dades orientada a grafs, dissenyada específicament per emmagatzemar, gestionar i consultar dades estructurades en forma de grafs. A diferència de les bases de dades relacionals tradicionals, que es basen en taules i relacions fixes, Neo4j utilitza la teoria de grafs per representar les dades. Això, permet que les relacions es puguin consultar de manera eficient. Per tal de crear els nodes hem fet servir el llenguatge de consulta Cypher desenvolupat per Neo4j per interactuar amb les bases de dades de grafs.

Els nodes que hem creat han estat els de tipus Persona i de tipus Publicació. Cada node i relació poden tenir propietats associades, que descriuen característiques específiques dels elements.

Tenint en compte que les publicacions estan compostes per un títol i un *abstract*, a l'hora d'afegir aquests elements com a atributs dels nodes, els hem filtrat per eliminar stopwords i quedar-nos únicament amb l'arrel de les paraules.

Les publicacions tenen associades unes Paraules Clau que defineixen l'àrea sobre la qual parla. Per tal d'emmagatzemar aquesta informació, hem creat nodes de tipus Word que estan connectats amb els nodes Publicació de la qual formen part. Considerant la rellevància dels títols, hem creat també nodes de tipus Paraula que corresponen a aquelles arrels de paraules que componen els títols de la publicació.

5.3 Creació de les arestes del graf

Un cop creats els nodes del graf, ara el següent pas és la creació de les connexions entre nodes. En primer lloc, creem les arestes que es poden realitzar de forma lògica a partir del conjunt de dades. Aquestes connexions es fan entre els autors que han realitzat una publicació, les paraules clau i les paraules del títol que componen una publicació.

En segon lloc, ens proposem crear connexions entre nodes que no connectaríem a simple vista amb les dades que tenim. Tenint en compte que el nostre objectiu és determinar com són de semblants dos investigadors, el que hem

[2]https://neo4j.com/docs/

de fer primer és crear arestes entre publicacions on es vegi reflectida la semblança entre elles. Aquestes arestes tindran un pes el qual correspondrà a la similitud entre publicacions sent el valor zero quan són completament idèntiques.

Fig. 4: Diagrama de nodes i relacions del graf.

5.4 Distància entre documents

El següent pas després d'haver generat el graf de coneixement és determinar la distància entre tots els documents. Per tal de fer això, veiem que de cada publicació volem recalcar tres tipus d'informació diferents; el títol, l'*abstract* i les paraules clau. Per tant, el que volem aconseguir és fer una comparativa entre aquests tres elements i elaborar una fórmula que determini com són de semblants dues publicacions de forma global.

5.4.1 Distància entre paraules clau

De cada publicació tenim un seguit de paraules clau que determinen en termes generals de què tracta la publicació. El primer que hem de fer, com hem esmentat anteriorment, és filtrar aquestes paraules clau per eliminar els stopwords i quedar-nos amb l'arrel de les paraules per tal de determinar que tant el plural, singular, femení o masculí d'una mateixa paraula tenen el mateix significat.

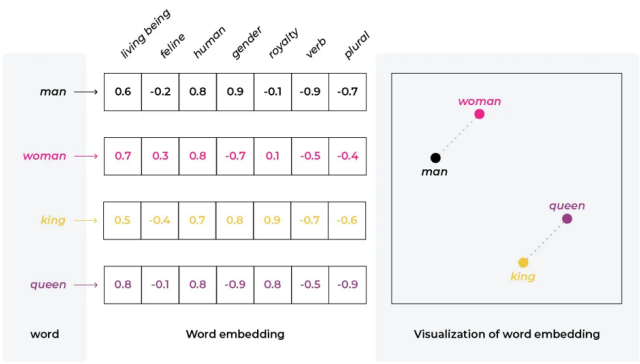
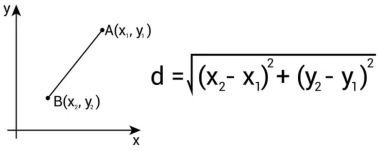


Fig. 5: Esquema d'un *embedding* de *Word2Vec*.

El que cal fer a continuació és una comparativa entre dues paraules per tal de determinar com són de similars. Per fer això, existeixen diferents algoritmes com el *Word2Vec* de la figura 5, el qual és un algoritme d'aprenentatge profund utilitzat en processament del llenguatge natural per a la representació vectorial de paraules. Al representar totes les paraules en forma de vector es poden efectuar diferents operacions matemàtiques les quals ens determinen la distància entre elles. Així doncs, donada una paraula clau, la representem en forma de vector dins l'*embedding* i calculem la distància que hi ha entre paraules de dues publicacions diferents.

Algunes distàncies que podríem calcular són la distància euclidiana, la similitud del cosinus, el coeficient de Jaccard entre d'altres. Així doncs, implementem el càlcul de les tres distàncies per a testejar més endavant quina s'adapta millor al problema plantejat.



(1)

La distància euclidiana és una mètrica que s'utilitza per calcular la distància entre dos punts en un espai n-dimensional. Aquesta distància es basa en el teorema de Pitàgores. La d de l'equació (1) representa la distància euclidiana entre els dos punts. On (x1, y1) són les coordenades del primer punt i (x2, y2) les del segon.

$$\text{Similarity}(p,q) = \cos\theta = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}} \sqrt{\sum_{i=1}^{n} q_i^{\,2}}} \tag{2}$$

La similitud del cosinus és una mesura utilitzada per determinar la similitud entre dos vectors en un espai vectorial. És àmpliament feta servir en àmbits com la recuperació de la informació, la recomanació de contingut i l'anàlisi de textos. L'equació (2) de la similitud del cosinus entre dos vectors p i q es compon pel producte escalar entre els dos vectors. On $\|p\|$ i $\|q\|$ són la norma euclidiana del vector p i q.

$$JaccardIndex = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$
$$JaccardDistance = 1 - JaccardIndex \tag{3}$$

La distància de Jaccard és una mesura que s'utilitza per avaluar la similitud entre conjunts. És especialment útil en l'àmbit de l'anàlisi de dades. La distància de Jaccard es basa en el principi de la intersecció i la unió de conjunts. La idea principal és comparar la quantitat d'elements comuns entre dos conjunts en relació amb la quantitat total d'elements presents en ambdós conjunts. Per calcular la distància de Jaccard com s'indica a l'equació (3), primer s'obté la intersecció entre els conjunts A i B i després es calcula la unió dels conjunts. La distància de Jaccard s'obté dividint la cardinalitat de la intersecció entre la cardinalitat de la unió i finalment, restant 1 al valor calculat.

Un cop calculades les distàncies entre les paraules, és necessari determinar una mesura de distància global entre les dues publicacions.

Inicialment, vam seleccionar la distància més petita entre una paraula clau d'una publicació i totes les paraules clau de l'altra publicació, i vam sumar aquestes distàncies per aconseguir una distància global. No obstant això, aquest enfocament presentava un problema: si més d'una paraula clau d'una publicació s'assemblava a una de l'altra publicació, considerèvem que aquestes publicacions eren molt semblants, tot i que en realitat només s'assemblaven a una de les paraules clau.

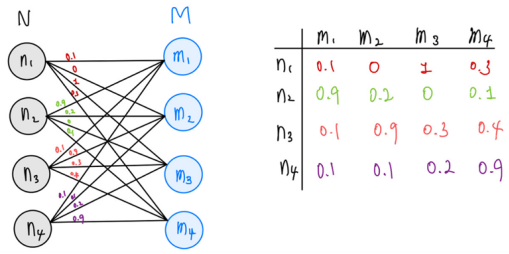


Fig. 6: Esquema de l'algoritme hongarès.

Per superar aquest problema, vam utilitzar l'*Hungarian Algorithm* de la figura 6, que permet trobar l'assignació òptima entre dos conjunts d'elements tenint en compte els costos associats.

Vam implementar aquest algorisme emprant una matriu de costos que pren en consideració que les publicacions poden tenir un nombre diferent de paraules clau entre elles. A partir d'aquesta matriu, es redueixen els costos per garantir que cada fila i cada columna contingui com a mínim un zero. A continuació, es marquen les files i les columnes de manera que cap element marcat sigui seleccionat per a l'assignació. Aquests passos es repeteixen fins que s'obtingui una solució òptima.

Un cop realitzats aquests càlculs, sumem totes les assignacions per aconseguir el valor de distància entre les publicacions en funció de les seves paraules clau. D'aquesta manera, assolim una mesura més precisa i completa de la similitud entre les dues publicacions segons les paraules clau.

5.4.2 Distància entre *abstracts*

Pel que fa a la comparativa dels *abstract*s de les publicacions hem fet una anàlisi diferent. En primer lloc, i de la mateixa manera que amb les paraules clau, hem eliminat les stopwords i ens hem quedat amb l'arrel de les paraules. Però a continuació, el que hem fet ha estat implementar una tècnica anomenada *Term Frequency-Inverse Document Frequency (TF-IDF)* la qual avalua la importància de les paraules en un document a partir del model *Bag-of-Words*. Aquest algorisme assigna un pes a cada paraula en funció de la freqüència d'aparició i freqüència inversa d'aparició en el document com indica l'equació (4).

$$
\begin{aligned}
TF(t,d) &= \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \\
IDF(t) &= log\frac{N}{1+df} \\
TF-IDF(t,d) &= TF(t,d) * IDF(t)
\end{aligned}
\tag{4}
$$

Com el seu nom indica, l'algorisme té dues parts diferenciades. El càlcul del TF el qual es basa en quantes vegades apareix un terme dins un document el qual es calcula dividint el nombre d'aparicions del terme en el document pel nombre total de termes del mateix document. I, d'altra banda, l'IDF que és la mesura de rellevància d'un terme en el conjunt de documents el qual es calcula dividint el nombre total de documents pel nombre de documents que contenen el terme.

Un cop hem assignat un TF-IDF a cada publicació, calculem les tres distàncies, la distància euclidiana, la similitud del cosinus i el coeficient de Jaccard per a veure quina té millors resultats segons el problema plantejat i el conjunt de publicacions amb el que estem treballant.

5.4.3 Distància entre títols

Pel que fa al càlcul de la distància entre títols el que hem fet ha estat comparar-los amb els dos algorismes explicats anteriorment, primer hem creat els nodes de tipus Paraula que corresponen a les arrels de les paraules que apareixen en el títol sense tenir en compte preposicions, connectors i d'altres, per a després implementar de nou el *word2vec* per aquestes paraules i desenvolupar l'algoritme *Hungarian Algorithm* el qual optimitza l'assignació entre paraules.

D'altra banda, també hem implementat l'algorisme TF-IDF que enlloc de comparar els *abstract*s compara els títols. D'aquesta manera, tenim dos valors diferents que ens determinen com són de similars els títols de les dues publicacions. Les dues implementacions mesuren el mateix, així que a base de proves veurem quina implementació és la més acurada.

$$
\begin{aligned}
final\ distance = total\_distance\_kw * weight\_keywords + \\
total\_distance\_title * weight\_title + \\
total\_distance\_abstract * weight\_abstract
\end{aligned}
\tag{5}
$$

Ara que ja tenim les distàncies de les paraules clau, els títols i els *abstract*s entre totes les publicacions cal determinar amb un valor la distància entre elles. Per a fer-ho, és necessari calcular amb una fórmula com es mostra en l'equació (5), la influència que té cadascuna per determinar la distància final. És per això que el que hem fet ha estat definir un pes per a cada valor el qual hem entrenat per tal de veure el que dona millors resultats. Els millors resultats dels pesos han estat de 0.35 pel weightKeywords, 0.24 pel weightTitle i 0.4 pel weight *abstract* de l'equació. Per acabar, el que hem fet és crear una aresta entre dos nodes de tipus publicació en cas siguin molt similars, és a dir, que la distància sigui relativament petita. Per a fer això, hem determinat un *threshold* el qual també hem ajustat a 0.7, ja que, és el que dona millors resultats.

5.5 Distància entre persones

Per tal d'establir una distància entre persones, es consideraran les publicacions que han fet conjuntament. I no només es tindran en compte aquelles publicacions que hagin fet juntes, sinó que també es consideraran aquelles d'autors diferents, siguin similars entre elles. Recordem que per a saber com són de similars dues publicacions, tenim una aresta entre nodes de tipus publicació que es diu PUBLSIML la qual guarda el pes de com són de similars.

Així doncs, hem de buscar tots els camins possibles que vagin del node Person1 al node Person2. Per tal de fer això, hem realitzat la següent consulta Cypher.

Aquesta comanda primerament realitza un MATCH per retornar tots els camins que van des del node Person1 fins al node Person2. S'estableix una longitud màxima de 3 per evitar camins massa llargs que perdrien la similitud de les publicacions inicials.

```
query = ("MATCH path = (p1: Person {pureID: $id_p1})-[*..3]->
                       (p2: Person {pureID: $id_p2})"
         "WHERE all(n in nodes(path)
           WHERE n:Publication OR n=p1 OR n=p2) "
         "RETURN nodes(path), relationships(path)"
         "ORDER BY length(path), relationships(path)[1].pes")
```

Un problema que ens vam trobar va ser que la consulta retornava camins on els nodes del mig no eren de tipus publicació, així doncs, per evitar això, fem el WHERE que revisa que tots els nodes del camí siguin de tipus publicació o que siguin de tipus persona en cas que siguin el node inicial o final. Després fem el RETURN per a retornar tots els nodes del camí i totes les relacions.

Un altre problema a considerar és que sovint totes les publicacions d'una mateixa persona estan connectades entre elles, ja que, solen ser articles similars. Per tant, s'aplica una funció per filtrar els camins que no són rellevants, considerant només el camí més curt entre dues persones.

És important que la llista de camins estigui ordenada de més curt a més llarg. Per això, s'aplica un ORDER BY a la consulta Cypher per ordenar els resultats segons la longitud del camí i el pes entre les publicacions similars.

En la funció, s'utilitza un diccionari de nodes visitats per evitar camins repetits. Es calcula una variable 'sumaacumulada' per indicar la similitud entre dues persones segons les seves publicacions. Per a les persones que són autors de la mateixa publicació, s'incrementa amb un valor fix establert com a 1. En canvi, per a les persones amb publicacions diferents però molt semblants, s'incrementa la variable 'sumaacumulada' considerant el pes de l'aresta entre les publicacions, que indica la seva distància.

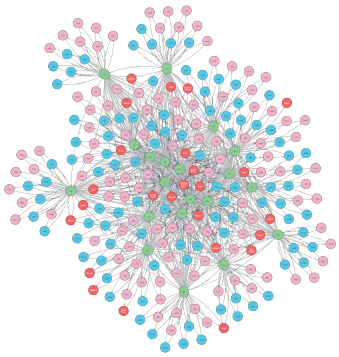


Fig. 7: Graf amb connexions de similitud de publicacions.

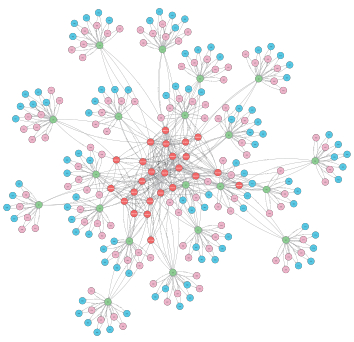


Fig. 8: Graf amb connexions de similitud de persones.

5.6 Mètriques de centralitat de nodes

Un cop tenim el nostre graf amb les connexions entre publicacions segons com són de similars i les connexions entre persones segons les publicacions que han fet, podem dir que tenim en aquesta estructura de graf totes les dades emmagatzemades.

Ara, el nou objectiu és realitzar un seguit de mètriques de centralitat de nodes sobre el graf per tal de mesurar la influència dels nodes dins aquesta xarxa, en aquest cas, dins el graf de persones de la UAB. Això ens permetrà tenir la informació necessària per donar suport a les consultes que un usuari faci del graf. Les mètriques de centralitat de nodes que hem mesurat són les següents:

Primer de tot la **Betweenness centrality** que és una manera de detectar la quantitat d'influència que un node té sobre el flux d'informació en un graf. S'utilitza sovint per trobar nodes que serveixen com a pont entre una part del graf i una altra. La funció que l'implementa és la següent: gds.betweenness.stream().

En segon lloc, el **Degree centrality** el qual mesura el nombre de relacions entrants i sortints d'un node. La funció que l'implementa és la següent: gds.degree.stream().

A continuació la **Closeness centrality** la qual és una manera de detectar nodes que poden difondre informació de manera molt eficient a través d'un graf. La funció que l'implementa és la següent: gds.beta.closeness.stream().

Tot següit, l'**Eigenvector centrality** que és un algoritme que mesura la influència transitiva o la connectivitat dels nodes. La funció que l'implementa és la següent: gds.eigenvector.stream().

Seguidament el **Page Rank** el qual és un algoritme que mesura la influència transitiva o la connectivitat dels nodes. La funció que l'implementa és la següent: gds.pageRank.stream().

Un cop calculades totes les mesures de centralitat de nodes, les guardem en diferents atributs per a cada node del graf.

5.7 Algorismes de detecció de comunitats

Un cop tenim calculades les mesures de centralitat de nodes, ara el que volem és implementar algoritmes de detecció de comunitats en el nostre graf. Els algoritmes de detecció de comunitats que hem mesurat són les següents:

Calculem l'algoritme **Louvain**, ja que, és un algoritme per detectar comunitats en xarxes. La funció que l'implementa és la següent: gds.gds.louvain.stream().

Tot seguit, realitzem el **Label Propagation** el qual és un algoritme ràpid per trobar comunitats en un graf. La funció que l'implementa és la següent: gds.labelPropagation.stream().

Un cop calculats tots els algorismes de detecció de comunitats també els guardem en diferents atributs per a cada node.

6 VALIDACIÓ EXPERIMENTAL

En aquesta secció tenim com a objectiu fer les validacions del sistema per a comprovar el seu bon funcionament.

Per tal de fer-ho, en lloc d'executar el programa amb totes les dades de la UAB, trobem que és una bona opció obtenir resultats a partir de conjunts petits de dades.

El primer conjunt consta de publicacions fictícies. S'han generat títols, *abstract*s i paraules clau que representen una publicació amb una intel·ligència artificial. Aquest conjunt de dades té quatre àmbits diferents: la tecnologia, l'art, l'esport i l'alimentació. En aquest conjunt de dades tenim en total 20 publicacions diferents, 5 per a cada categoria.

La primera de les tasques a realitzar ha estat la creació del graf a partir d'aquest primer conjunt de dades. Hem creat els nodes Publicació que són els de color blau a la figura 10, els nodes Paraula que són els de color vermell i els nodes Word que representen les Paraules Clau de color rosa.

Després el que hem fet ha estat calcular la similitud entre publicacions a partir de la distància entre títol, *abstract* i paraules clau. En aquesta primera execució, s'han establert els pesos a 0.45 de paraules clau, 0.3 de títol i 0.25 d'*abstract*, ja que, pensem que l'*abstract* dona informació irrellevant i que, en canvi, les paraules clau i el títol són més concretes. En cas que la distància final sigui menor al *threshold* de valor 0.7 es creen les arestes de similitud entre publicacions.

L'eina que hem utilitzat per tal de realitzar la visualització de les dades ha estat l'entorn Gephi[3]. Aquest, és un entorn preparat per a la visualització de grafs i anàlisi de xarxes. S'utilitza per explorar i comprendre l'estructura i les interaccions en conjunts de dades de xarxes complexes. És per això que hem connectat la base de dades de Neo4j des de Gephi i carregat tota l'estructura de graf.

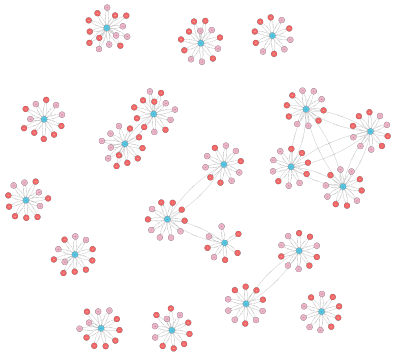


Fig. 9: Prototip del graf de coneixement.

Com podem veure a la figura 9, algunes de les publicacions estan connectades entre elles com és l'exemple de les quatre publicacions de la part superior dreta de la figura. Aquestes publicacions formen part de la mateixa categoria; tecnologia. Això ens dona una primera evidència que el sistema funciona. D'altra banda, es veuen a la figura dues publicacions que són semblants a una i això ve donat perquè totes tres parlen de l'àmbit esportiu. Com podem analitzar, hi ha moltes publicacions que no estan connectades amb cap altra, és per això que modificarem el valor del *threshold* que connecta publicacions i els valors dels pesos per a veure com afecten la distància entre publicacions.

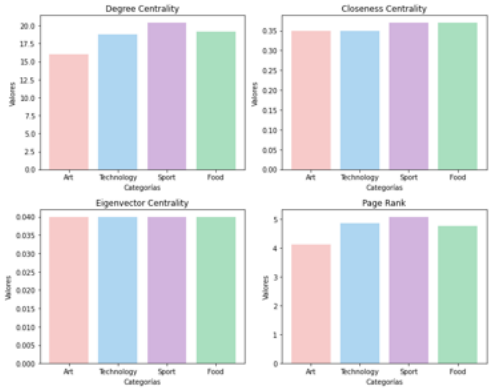


Fig. 10: Gràfica que representa les scores de les mètriques d'algunes categories.

En aquesta gràfica de la figura 10, podem analitzar per a les categories d'art, tecnologia, esport i aliments quatre de les mètriques que hem implementat; Degree Centrality, Closseness Centrality, Eigenvector Centrality i Page Rank. Podem observar com la categoria de més centralitat en els nodes és la d'esports, ja que, en ella s'acostuma a parlar sobre una bona alimentació i tecnologies en l'esport.

Per tal de fer més comprovacions, hem agafat una publicació de les dades reals d'EGRETA i hem fet una còpia d'aquesta. Hem realitzat petites modificacions i les hem comparat. Hem modificat part del títol, alguna paraula clau i hem vist que aquestes no tenen tanta rellevància com l'*abstract*, que dona una informació més global de la publicació, mentre que tant el títol com les paraules clau són massa específiques per a trobar coincidències. Això ens ha fet modificar els valors dels pesos a 0.28 pel que fa a les paraules clau, 0.34 pel títol i 0,38 a l'*abstract*.

			%keywords	%title	%abstract	
dist_kw	dist_title	dist_abst	0,28	0,34	0,38	DistTotal
0,5	0	0	0,14	0,00	0,00	**0,14**
0	0,5	0	0,00	0,17	0,00	**0,17**
0	0	0,5	0,00	0,00	0,19	**0,19**
1	0	0	0,28	0,00	0,00	**0,28**
0	1	0	0,00	0,34	0,00	**0,34**
0	0	1	0,00	0,00	0,38	**0,38**
0,5	0,5	0	0,14	0,17	0,00	**0,31**
0	0,5	0,5	0,00	0,17	0,19	**0,36**
0,5	0	0,5	0,14	0,00	0,19	**0,33**
0,5	0,5	0,5	0,14	0,17	0,19	**0,50**
1	1	0	0,28	0,34	0,00	**0,62**
0	1	1	0,00	0,34	0,38	**0,72**
1	0	1	0,28	0,00	0,38	**0,66**
1	1	1	0,28	0,34	0,38	**1,00**
0,5	1	1	0,14	0,34	0,38	**0,86**
0,5	0,5	1	0,14	0,17	0,38	**0,69**
1	1	0,5	0,28	0,34	0,19	**0,81**

Fig. 11: Influència dels pesos a la similitud de publicacions.

En aquesta figura 11 podem observar els resultats de la distància entre les dues publicacions amb els pesos establerts. Les columnes de color verd representen la distància dels elements, les blaves el pes i els resultats es veuen a la columna DistTotal. Podem veure que en donar-li més importància l'*abstract*, per molt que les paraules clau o el títol no tingui gaires coincidències, el valor final encara és baix.

Hem fet el mateix, però amb altres valors pels pesos com és el cas de la figura 12. On hem establert els pesos a 0.35 de paraules clau, 0.25 al títol i 0.4 per l'*abstract*. D'aquesta manera li donem més importància a les paraules clau.

Per tal de veure quina de les combinacions s'adapta més a les nostres necessitats, hem agafat 20 publicacions de les

[3]https://gephi.org/users/documentation/

			%keywords	%title	%abstract	
dist_kw	dist_title	dist_abst	0,35	0,25	0,4	DistTotal
0,5	0	0	0,18	0,00	0,00	**0,18**
0	0,5	0	0,00	0,13	0,00	**0,13**
0	0	0,5	0,00	0,00	0,20	**0,20**
1	0	0	0,35	0,00	0,00	**0,35**
0	1	0	0,00	0,25	0,00	**0,25**
0	0	1	0,00	0,00	0,40	**0,40**
0,5	0,5	0	0,18	0,13	0,00	**0,30**
0	0,5	0,5	0,00	0,13	0,20	**0,33**
0,5	0	0,5	0,18	0,00	0,20	**0,38**
0,5	0,5	0,5	0,18	0,13	0,20	**0,50**
1	1	0	0,35	0,25	0,00	**0,60**
0	1	1	0,00	0,25	0,40	**0,65**
1	0	1	0,35	0,00	0,40	**0,75**
1	1	1	0,35	0,25	0,40	**1,00**
0,5	1	1	0,18	0,25	0,40	**0,83**
0,5	0,5	1	0,18	0,13	0,40	**0,70**
1	1	0,5	0,35	0,25	0,20	**0,80**

Fig. 12: Influència dels pesos a la similitud de publicacions.

dades d'EGRETA i les hem analitzat al detall per a veure si les connexions entre publicacions s'adapten a la realitat.

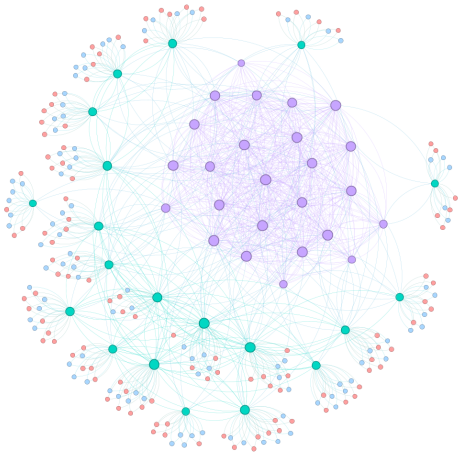


Fig. 13: Graf de persones i publicacions més similars.

El resultat que obtenim amb aquests pesos és el que es mostra en la figura 13. Podem veure com les publicacions que són els nodes de color verd estan ben connectades amb els nodes persona de color lila, però veiem que no estan altament connectades les publicacions entre elles. Això ve donat perquè el valor 0.53 del *threshold* és massa baix. Per tant, modifiquem aquest valor a 0.7.

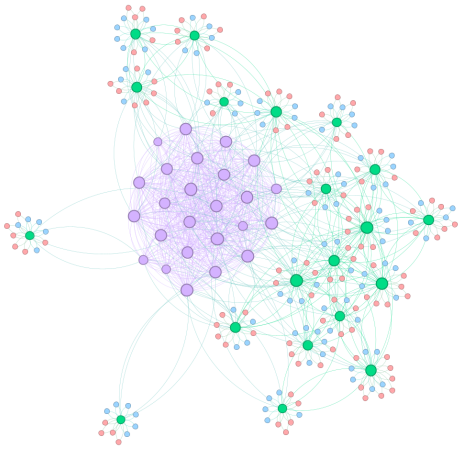


Fig. 14: Graf de persones i publicacions més similars.

Al modificar aquest valor del *threshold* a 0.7, podem veure amb la figura 14 com ara sí que hi ha publicacions en la part inferior dreta altament connectades entre elles. Això ve donat perquè en aquest segon conjunt de dades d'EGRETA, hem seleccionat algunes publicacions del departament de Computer Science i d'altres departaments que són els que no estan tant connectats amb altres publicacions, ja que, acostumen a ser de temàtiques diferents.

Després de vàries proves, hem vist que la millor combinació de pesos és 0.35 per les *keywords*, 0.24 pels títols i 0.4 pels *abstract*s, ja que, és la que satisfà el sistema de recomanació que desitgem. El resultat d'executar tot el sistema que determina la similitud d'investigadors a partir de la similitud de publicacions amb aquests valors de pesos establerts és el que es mostra a la figura 15.

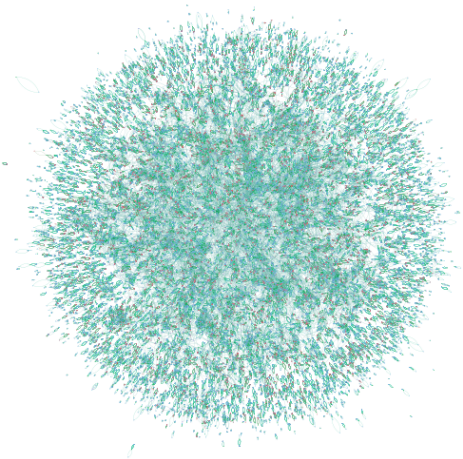


Fig. 15: Graf que representa la similitud d'investigadors i publicacions amb les dades d'EGRETA.

7 ENTORN PER A L'USUARI

S'ha desenvolupat un entorn per a l'usuari que actua com a sistema recomanador, proporcionant suggeriments i recomanacions personalitzades basades en les seves preferències i interessos.

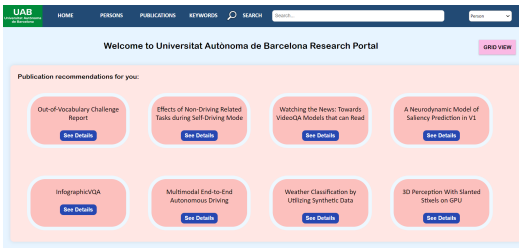


Fig. 16: Recomanacions de publicacions a l'usuari.

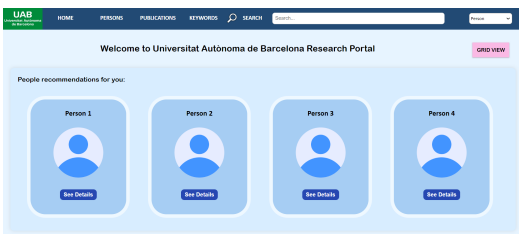


Fig. 17: Recomanacions d'investigadors a l'usuari.

L'entorn permet interactuar amb el sistema recomanador, donant la possibilitat a l'usuari d'escriure un resum i paraules clau d'interès. Mitjançant l'anàlisi dels interessos de l'usuari, se li recomanen investigadors i publicacions.

L'entorn creat ofereix una cerca segons publicacions, persones o paraules clau i proporciona un mode de visualització en forma de graf per tal d'enriquir la seva experiència.

8 Conclusions

En aquest treball hem aplicat amb èxit tècniques de processament del llenguatge natural i mineria de dades per a l'anàlisi de similituds en un context acadèmic. Mitjançant l'ús d'algoritmes com la distància euclidiana, la similitud del cosinus i el coeficient de Jaccard, hem calculat amb precisió la similitud entre les paraules clau, els títols i els resums de les publicacions. Hem millorat els resultats mitjançant l'algorisme hongarès per a l'assignació òptima entre les paraules clau i els títols. La representació de paraules com a vectors numèrics utilitzant *Word2Vec* ha demostrat ser efectiva en l'avaluació de similituds entre els títols de les publicacions. A més, l'ús de TF-IDF ha contribuït a assignar pesos als termes clau i resums, afegint més precisió a la mesura de similitud.

Els resultats obtinguts són prometedors i indiquen el potencial per a l'anàlisi i comparació de publicacions. Aquesta metodologia pot ser emprada en futurs estudis per aprofundir en l'agrupació de documents similars. A més, pot ser útil per als investigadors per identificar de manera eficient publicacions relacionades amb un tema específic.

Cal destacar que tot i els resultats positius, aquest estudi presenta algunes limitacions. En particular, la nostra anàlisi es va centrar en les paraules clau, títols i resums, deixant de banda altres elements textuals com el contingut complet dels documents. Per tant, hi ha una oportunitat per a futures investigacions per ampliar aquest treball i incloure altres aspectes textuals per a una anàlisi més completa i precisa.

En resum, aquest estudi ha proporcionat una visió de les tècniques usades per a l'anàlisi de similituds en un context acadèmic. El treball ha demostrat la els beneficis d'aquest enfocament, així com les seves limitacions. Les conclusions obtingudes són una base sòlida per a futurs treballs en aquest àmbit i ofereixen una comprensió més profunda de l'anàlisi de similituds en publicacions acadèmiques.

9 Treball futur

Malgrat els avenços realitzats en aquest Treball de Fi de Grau, hi ha diverses àrees que podrien ser explorades i aprofundides en treballs futurs.

Algunes propostes serien fer una investigació més profunda sobre les necessitats de l'usuari, es podiren realitzar tècniques de *tracking* per a veure els interessos i crear perfils d'usuari més acurats.

D'altra banda, la integració d'aquesta intel·ligència al Portal de Recerca de la UAB seria un punt important a desenvolupar per tal de fer-la una eina més eficient.

Es podria aprofundir en l'ús de les tècniques d'anàlisi de dades per tal d'aconseguir resultats més precisos i es podrien utilitzar altres fonts de dades externes com *Google Schoolar* per a la generació del graf de coneixement.

10 Agraïments

Vull expressar els meus sincers agraïments a totes les persones que han contribuït a la realització d'aquest Treball de Fi de Grau:

En primer lloc, vull agrair al meu tutor del projecte Josep Lladós, per la seva orientació, suport i valuosos consells durant tot el procés de desenvolupament del treball. La seva experiència i coneixement han estat essencials per a l'èxit d'aquest projecte.

També vull agrair als membres del departament de *Document Analyisis Group* per compartir els seus coneixements i proporcionar una base sòlida en el meu camp d'estudi.

A més, vull expressar el meu agraïment als meus amics i família per donar-me suport constant i motivació durant el desenvolupament del treball.

Referències

[1] Gavrilova, Y. (2020, de Desembre). Machine Learning and Text Analysis. https://serokell.io/blog/machinelearning- text-analysis

[2] Maklin, C. (2019, de Maig). TF IDF Algorithm in Python. https://towardsdatascience.com/naturallanguage-processing-feature-engineering-using-tf-idf-e8b9d00e7e76

[3] Vatsal. (2021, de Juliol). Word2Vec Explained. https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak Zachary Ives. (2007). DBpedia: A Nucleus for a Web of Open Data. https://link.springer.com/chapter/10.1007/978-3-540-76298-0-52

[5] Janev, Valentina, Graux, Damien, Jabeen, Hajira, Sallinger, Emanuel. (2020). Knowledge Graphs and Big Data Processing. https://library.oapen.org/handle/20.500.12657/41294

[6] Adrien Sieg. (5 Juny 2018). Text Similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894: :text=What

[7] Déborah Mesquita. (7 Maig 2017). Clasificando textos con Redes Neuronales y TensorFlow. https://medium.com/@dehhmesquita/classificando-textos-com-redes-neurais-e-tensorflow-5063784a1b31

[8] Lei Tang, Huan Liu. (Gener 2010). Community Detection and Mining in Social Media. https://www.researchgate.net/publication/220696280-Community-Detection-and-Mining-in-Social-Media

[9] Badrul Sarwar, George Karypis, Joseph A Konstan, John Riedl. (2001). Item-based collaborative filtering recommendation algorithms. https://experts.umn.edu/en/publications/item-based-collaborative-filtering-recommendation-algorithms