HIST4XAC: Una Eina d'Anàlisi de Documents Pel Manteniment del Patrimoni Històric

Adrià Molina Rodríguez
CVC - UAB - Escola d'enginyeria
amolina@cvc.uab.cat

Resum– Heritage Intelligent Support Tool for Xarxa d'Arxius Comarcals (HIST4XAC) és una eina web per visualització explorativa, *data profiler*, que intenta proveïr una col·lecció de serveis i utilitats basades en aprenentatge profund i visió per computador a arxivers i científics de l'àmbit social. Concretament les tecnologies es basen en eines de cerca d'informació a grans volums de dades en vers l'estimació de dates dels documents i imatges analitzades.
Paraules clau– Anàlisi de documents, documents històric, humanitats digitals, aprenentatge profund, visió per computador, ciències socials, archivística, gestió de la informació, perfilat de dades, estimació de dates.

Abstract– The Heritage Intelligent Support Tool for the Comarcal Archives Network (HIST4XAC) is a web visualization tool, *data profiler*, that aims to provide a collection of services and utilities based on deep learning and computer vision to archivists and social scientists. Specifically, the technologies are based on information search tools for large volumes of data towards the estimation of dates of the documents and images analyzed.

Keywords– Document analysis, historical documents, digital humanities, deep learning, computer vision, social science, archival science, information management, data profiling, date estimation.

1 Introducció

Amb l'arribada de la digitalització i el creixent flux de dades a les xarxes i infraestructures; també s'ha produit un creixement en l'interés per digitalització de la documentació histórica [1], [2]. Sols a catalunya, durant el 2020 es van digitalitzar 27TB de documents histporics per part de la Xarxa d'Arxius Comarcals [3]. Això produeix una sèrie d'avantatges en vers l'anàlisi d'aquesta documentació, ja que algorismes com és el cas de l'aprenentatge profund poden donar una perspectiva robusta i eficient d'un volum de docuentació que una sola persona no podria ser capaç d'analitzar.

No és extrany que, amb una certa periodicitat, les institucions encarregades de mantenir i enriquir el llinatge histórico rebin voluminoses col·lecions de dades pendents de catalogar. El que acostuma a ser més comú és que l'arxiver o encarregat de gestionar aquesta nova injestió de documentació, no necesiti més que una primera ordenació de les dades sota una sèrie d'etiquetes que pretenen catalogar aquests conjunts de la forma més efectiva possible.

Per altre banda a la investigació social i, sobretot, històrica; es necesiten moltes vegades eines comparatives o de cerca per contingut. Això implica la necesitat de cercar documents que puguin estar relacionats amb el camp o els interesos puntuals de l'investigador. Per tant, es necesita una eina no sols capaç de carregar aquest volum d'informació i etiquetar-ho, si no que aquesta eina a més a de ser capaç de donar informació general sobre la col·lecció, i retornar informació relacionada que pugui ser d'interés donada una certa *query*.

En aquest contexte s'ha desenvolupat l'HIST4XAC; un conglomerat de funcionalitats basades en l'anàlisi de documents que haurien de facilitar algunes de les tasques exploratives dels documents aportats per les institucions encarregades de mantenir el patrimoni històric. Així podem veure l'aplicatiu web com un *data profiler* per a documentació histórica, sobretot centrada en la datació de documents.

2 Eines Relacionades

Com s'exposarà a la Secció 3, aquesta eina s'engloba en el marc de la gestió de dades (*data management*) i la visualització. Es pot veure com una eina de *historical data warehousing*. Les eines de data warehousing han guanyat popularitat els últims anys, destacant algunes per la seva versa-

• E-mail de contacte: amolina@cvc.uab.cat
• Treball tutoritzat per: Oriol Ramos Terrades, Josep Lladós Canet i Lluis Gomez Bigordà (Ciències de la Computació).
• Curs 2021/2022

tilitat i bona adaptabilitat a les noves industries basades en dades, en particular aquelles basades en cloud, gràcies a la seva facilitat per escalar horitzontalment a través del desplegament d'instàncies i verticalment a través de l'adquisició de millors prestacions [4, 5].

Aquesta eina s'engloba també dins del marc de les humanitats digitals, on els avenços més recents giren entorn dos pilars fonamental; la divulgació i ludificació del patrimoni històric [6] i el suport a institucions i investigadors [7–11]; de fet a la XAC es van dur a terme dues activitats majoritaries catalogades com "Treballs d'investigació i recercaï "Participació en mitjans de comunicació"; amb altres marginalment frequents com "Visites guiades"o Çonferències, seminaris, etc." [3]. Ambdues tendències giren entorn la idea de la democratització de les tecnologies i sistemes inteligents en vers les humanitats digitals. En aquest sentit destaquen projectes com l'Europe Time Machine[1], que intenta asolir una digitalització del passat equiparable a la digitalització actual; el projecte Xarxes[2] que pretèn reconstruir la xarxa social històrica de Catalunya a partir del registre de padrons i arxius en general o Matricula[3], un arxiu online que pretén estudiar comunitats basades en l'anàlisi de documents proveients de col·leccions heterogenies.

3 OBJECTIUS

Una visió més detallada dels objectius es pot trobar al primer document de proposta del projecte. En linies generals és plantegen els següents reptes, a nivell d'objectius generals diferenciem els següents:

- Millora de la qualitat de vida dels procesos de digitalització de documents històrics.
- Augment de la productivitat i la qualitat de les dades proveïdes als estudis socials i històrics [12].
- Baixa latència a l'injecció de noves dades de llinatge històric.
- Impacte positiu en el manteniment de la memória histórica a les institucions.

Per a aconseguir-los, es proposen eines de visualització i perfilat de dades recollides a la Secció 5.4; entre elles destaquen la capacitat de dividir un volum de dades segons el periòde històric al que pertanyen, així com la seva visualització en galeries; la projecció en dues dimensions de l'espai de dades retornat pels models de visió o la cerca d'informació a les bases de dades.

A nivell tècnic es proposen una sèrie de requeriments relacionats amb la infrastructura de procesament de dades:

- Les consultes han de ser ràpides
- Les institucions com l'Arxiu nacional està dividit en arxius comarcals; els quals poden tenir diversos treballadors interessats en l'ús simultani de l'aplicatiu. Ha de ser horitzontalment escalable i facilment desplegable en instàncies.
- Donat que és un aplicatiu amb accés restringit als usuaris membres de la institució, no cal considerar rols que restringeixin l'accés a determinades dades. Les dades han de ser compartides per tota la organització, hauria de servir com una mena de **data warehousing**.

Sintetitzant la secció, l'objectiu principal d'aquest projecte és la creació d'una eina de *data management* per a arxius històrics basada en el perfilat de dades i la visualització.

4 PLANIFICACIÓ

Com podem observar al diagrama Gantt, el desenvolupament de l'aplicatiu és pràcticament lineal, amb una branca vermella que indica els trams als quals, un cop desenvolupada una funcionalitat, s'ha de general el seu fitxer `CSS` per ajustar al disseny de l'aplicació. Podem observar a alta resolució la proposta de diagrama Gantt al repositori de GitHub.

Desglosant la planificació proposada a l'inici del projecte hem obseervat que en linies generals s'ha complert en els plaços establers, excepte per un desplaçament uniforme d'una setmana de retards. Això provoca que funcionalitats com el re-entrenament hagin sigut menys prioritzades per aquesta primera proposta d'aplicatiu per la gestió de dades.

A més, per tal de donar més pes al correcte funcionament de l'infrastructura no s'ha prioritzat la construcció d'arxius `CSS` per embellir el resultat. Donat que l'objectiu és proveïr un servei de *data management* [12] més que no pas un projecte de desenvolupar programari estrictament parlant, es creu que la decisió ha sigut la indicada, i que el resultat és un bon prototip del que hauria de convertir-se en una eina de perfilat de dades históriques construit en els plaços raonablement similars a la planificació inicial.

5 DESENVOLUPAMENT TÈCNIC

El desenvolupament de l'aplicatiu ha sigut bastant modular; en general s'ha intentat en tot moment establir primerament funcionalitats o `APIs` sobre les quals es construeixen les diferents eines. El fet d'haver restringit especialment la creació de funcions o crides "a mida"de les necesitats puntuals ha supossat en certs punts un repte, pero ha provocat un desenvolupament més àgil i robust a llarg termini.

5.1 Motor d'Aprenentatge Profund

Com s'ha esmentat a la Secció 1 un dels principals atractius de l'aplicació és l'ús de models de cerca d'informació amb aprenentatge profund.

Tot i que la flexibilitat de l'aplicació hagi permés incorporar diversos tipus de models ben coneguts al camp del deep learning [13–15] aquest apliacatiu neix com a resposta o continuació a models que s'han estat desenvolupant paralelament en vers al retorn d'informació. Aquestos han sigut ampliament estudiats a estudis previs [9, 10, 16].

Esencialment, aquests models es basen en l'optimització de funcions de ranking, com és el cas de l'*smooth-nDCG* Equació 1. A diferència de funcions de ranking derivables més conegudes com el *Smooth-mAP* [17, 18], el *Normalized Discounted Commulative Gain* permet evaluar rankings no

[1] https://www.timemachine.eu/
[2] http://dag.cvc.uab.es/xarxes/
[3] https://data.matricula-online.eu/en/

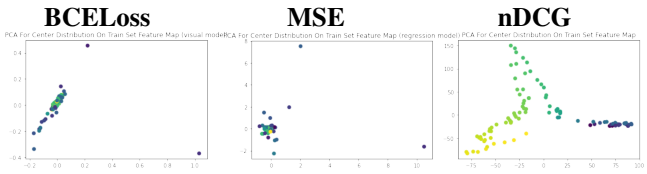

Fig. 1: Projecció PCA de l'espai vectorial d'una xarxa entrenada amb classificació (esquerra), regressió (centre) i retrieval (dreta). El color del punt indica el valor de la categoria, una variable continua, en aquest cas l'any. Observem com amb nDCG la distribució de centroides de les etiquetes a l'espai de sortida respecta la informació semàntica de les mateixes.

binaris, on un cert document i té una rellevancia $r(i)$ expresada en forma continua respecte una *query* q. L'ús d'aquesta funció en contrast a clasificadors o regressors clàssics provoca que l'espai vectorial de sortida de la xarxa neuronal mantingui l'informació semàntica en la seva estructura geomètrica; dit d'altre manera, les distàncies a l'espai semàntic expressat a través de la rellevancia $r(i)$ es mantenen proporcionals a l'espai de sortida Figura 1.

$$DCG_q \approx \sum_{i \in \Omega_q} \frac{r(i)}{\log_2 \left(2 + \sum_{j \in \Omega_q, j \neq i} \mathcal{G}(D_{ij}; \tau)\right)} \quad (1)$$

5.2 Base de Dades Documental

Una de les principals decisons de disseny ha sigut l'ús d'una base de dades documental. Això, a més d'un cert grau de flexibilitat, aporta una bona interacció amb Python, lo qual és molt rellevant com s'exposarà a 5.3.

S'ha triat treballar en `MongoDB`, una base de dades documental que permet un fàcil acoplament a aplicacions de tota mena, ja que les peticions actuen directament en forma de peticions a una API localitzada en un endpoint determinat, per exemple, `hist4xac.cvc.uab.cat/mongo/?q={example: {$exists: True}}`, podria llençar una consulta a través de l'API. D'aquesta manera acabem de deslligar la base de dades de l'aplicatiu a través d'un funcionament d'`API`. En general, les bases de dades documentals permeten el desenvolupament d'aplicacions més flexibles i que es poden comportar i canviar de forma dinàmica sense irrompre en la base de dades.

En segona instància, MongoDB és obert, fàcil d'instalar i desplegar a qualsevol màquina, així podem tenir un ull possat en el desplegament en plataformes cloud. Relacionat amb aquest mateix objectiu, MongoDB és altament escalable de forma horitzontal, provocant així que puguem realitzar operacions amb moltíssimes dades de forma ràpida i económica a través de replica-sets a instàncies o màquines de no gaire potència.

Perd el sentit parlar de models d'entitat-relació en termes de bases de dades documentals; nogensmenys algunes de els entitats principals que té sentit considerar al projecte venen representades a la figura 2. Noteu que en el cas de les imatges, s'ha decidir guardar tant la pròpia imatge com el camí al fitxer; això tot i semblar que pugui produir redundància aporta dos beneficis principals:

1. Quan parlem de servir imatges al front-end, és molt més ràpid donar el camí; però quan volem computar procesos sobre les imatges és convenient no haver de carregarles desde memòria secondaria per reduir l'overheat.

2. Si pel que sigi, cambiés el sistema de fitxers, seguim podent donar resposta a les peticions d'imatges de forma temporal mentres es soluciona, fins i tot tornar a descarregar la imatge al lloc on hauria d'estar.

Com s'ha exposat anteriorment, hem desacoblat l'aplicatiu de la base de dades a través d'una `API`. Algunes de les funcionalitats més rellevants de la mateixa són:

- Guardar i carregar dades serialitzades com els models o imatges. En comptes de guardar-les al propi document, produint un altísim cost en memória; guardem l'index a una base de dades auxiliar, d'aquesta manera evitem haver de carregar N imatges quan el que voliem realment són les seves metadades.

- Actualitzar documents donada una query: És una capa per sobre de la funcionalitat que proporciona mongoDB asegurant que es mantenen els formats adients.

- Trobar col·leccions d'imatges: Retorna una llista de les col·leccions disponibles, estalviant queries més complexes que s'hauran de repetir varies vegades.

Tota aquesta mena d'`API` en `Python` funciona com a punt intermig entre HIST4XAC i `MongoDB`, asegurant formats i centralitzant les operacions de mongo més freqüents per part de localitzar errors de cerca més fàcilment.

5.3 Infrastructura pel Procesament de Dades

Per a fer efectiu el desenvolupament de l'eina s'ha plantejat una infrastructura determinada. Com obervem Figura 3 s'ha intentat compartimentar al màxim els mòduls. Això vol dir que per evitar caure en desenvolupament d'aplicacions monolítiques, poc flexibles i no escalables cadascuna de les capes sols hauria de poder comunicar-se amb les capes immediatament superiors o inferiors, sense salts.

Referint-nos a la figura esmentada Figura 3 trobem les següents particularitats al nostre esquema:

1. El front-end es basa en `HTML, JS i CSS`, aquest front-end envia les peticions a `Flask`.

2. `Flask` fa la comunicació entre les funcionalitats i les peticions de la capa superior, enviant a respondre-la al módul f_n pertinent.

3. La funcionalitat f_n en qüestió es comunica amb l'`API` esmentada al punt Secció 5.2; demanant les dades necesaries.

4. Aquesta `API` fa la petició al servidor `MongoDB`, independentment de la seva localització física, ja que funciona a través de l'endpoint especificat.

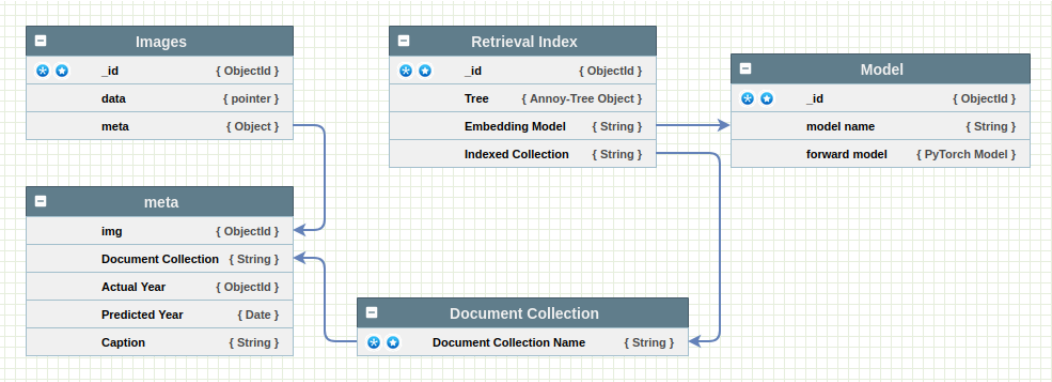

Fig. 2: Diagrama d'algunes de les relacions més rellevants del model de base de dades documentals.

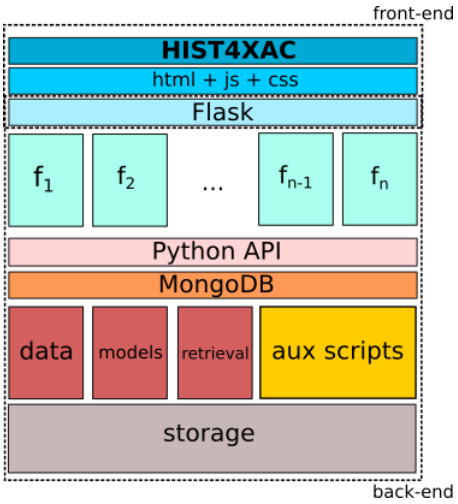

Fig. 3: Arquitectura modular de l'aplicatiu HIST4XAC. Diferenciem les parts amb interficies que les comuniquen.

5. El propi `MongoDB` retorna la informació de la base de dades i, de ser necesari, desencadena scripts auxiliars que mouran fitxers o executaran procesos necesaris en segon pla; com s'ha esmentat aquests procesos auxiliars sols podran afectar al emmagatzenament local o a la pròpia base de dades, mai es comunicaran directament amb cap capa no adjacent.

6. Per últim, es torna a propagar aquesta cadena de comunicació fins el front-end, on les capes superiors aniran desempaquetant la informació de les inferiors.

Les funcionalitats triades com a móduls f_n seran discutides a la Secció 5.4. Notem que la majoria d'aquestes funcionalitats estan basades en tractament de models d'IA; pels quals `PyTorch` és una de les biblioteques dominants. Per aquest motiu, totes les capes es construiran amb la idea de que sigui senzill utilitzar aquests models. Així s'ha utilitzat `HTML` perque és el que facilita més la comunicació amb templates de `Flask + Jinja2`; així mateix `Flask` és desplega en funcions de `Python` que actuan com a endpoints. `MongoDB`, entre d'altres virtuts discutides a Secció 5.2, funciona bé amb `Python` a través de la biblioteca `PyMongo`.

Fig. 4: Espai de cerca proposat. Noteu que en el cas que qualsevol dels subespais no s'estiguin utilitzant, el seu volum equival al de la capa inmediatament superior.

5.4 Funcionalitats

Com s'ha exposat a la Secció 5.3, s'han desenvolupat els moduls amb les funcionalitats indicades a la Secció 3. Aquestes són completament independents entre elles, el que significa que el front-end es comunica amb qualsevol dels moduls segons la seva necesitat.

5.4.1 Cercador

Una de les necesitats més evidents en aquest tipus de sistemes és la de cercar informació segons alguns paràmetres. En aquest cas es proposa un espai de cerca Figura 4 on es pot efectuar cerca per contingut, per etiqueta (any del document) i filtrar el nombre de resultats.

Per efectuar la cerca per etiqueta és tan simple com fer una crida a l'`API` Secció 5.2, 5.3 demanant tots els documents que satisfan la condició de tenir disponible una etiqueta (predita o etiquetada) com la demanada.

Per a efectuar aquesta comana, en cas que s'hagi cercat addicionalment sobre l'espai d'imatges visualment similars, es crida al model dessitjat per processar la imatge. Un cop es té el vector de característiques s'ha utilitzat un index preprocesat amb `Annoy`[4]. S'ha de destacar que es pot fer la cerca no per una sola imatge, si no per un conjunt d'imatges; això és especialment interessant en el cas dels espais vectorials ordenats per data Secció 5.1 [16], ja que el centroide d'un grup d'imatges es correspon a la mitjana de les seves dates. Aleshores donada la llista d'imatges candidates L, la cerca pasa a ser la mateixa, però restringida al conjunt d'imatges retornades per l'índex.

[4] Annoy: https://github.com/spotify/annoy

5.4.2 Fletxa Històrica i Galeria

En termes de documentació històrica, es té un particular interés per estudiar la evolució d'una col·lecció particular en el temps. Per això es proposa la construcció de les linies temporals de les col·leccions amb les etiquetes conegudes o aproximades.

Observem un prototip de la fletxa temporal 5 amb el que es poden accedir a galeries dels diferents anys. Això permet tenir una vista general del tipus de document d'un any determinat, així com donar anotacions als documents que poden servir tant com fer el retorn cap a centres de recerca per models de captionning com per fer anotacions útils per recerca o divulació històrica.

5.4.3 Visualitzador 2D

Una de les funcionalitats amb la que s'ha mostrat major entusiasme es la projecció de col·leccions en dues dimensions, o visualitzador 2D. Aquest visualitzador tracta de projectar l'espai de sortida dels models disponibles a un espai 2D que podem visualitzar, $\mathbb{R}^n \rightarrow \mathbb{R}^2$. Aquesta transformació es fa mitjançant una transformació d'anàlisi de components principals (PCA). S'ha decidit utilitzar PCA per sobre de TSNE perque el primer acostuma a comportar-se de forma més laxa amb les estructures locals, mentres que la geometria global es conserva millor. Això ens interesa perque pels models basats en ordenació d'etiquetes, és rellevant conservar l'estructura global on l'espai vectorial és proporcional a l'espai semantic Secció 5.1. De totes maneres en molts altres tipus de models, és més interesant observar els comportaments locals, per això canviar entre PCA i TSNE és relativament senzill gràcies a estar utilitzant `sklearn`.

Al selector de visualitzador veiem que es realitzen les projeccions de tots els models disponibles amb totes les col·leccions pujades a l'aplicatiu. Això provoca que certes combinacions no tinguin gaire sentit (com models de premsa històrica per col·leccions de fotografies d'esports). Al botó que accedeix a cadascuna de les visualitzacions es mostra un overview de l'espai projectat, podent així identificar facilment la modularitat dels agrupaments de les dades.

A nivell tècnic, s'ha adaptat una eina basada en `leaflet` on es genera un `HTML` capaç de mostrar imatges projectades a un pla de "tiles"amb un zoom determinat. Donada aquesta eina es preparen les "tiles"per una col·lecció i model determinats i s'envia a procesar el mapa en segon pla. Un cop fet el mapa està accesible a l'endpoint de l'aplicació proporcionat per `/visualizer/<model>/<col·lecció>`. La contribució del projecte ha sigut la gestió i integració d'aquesta eina [19] a un numero indeterminat de models i col·leccions.

Noteu que amb aquest mecanisme, es pot portar a procesar qualsevol tupla de (model, col·lecció) per tal de generar la projecció. Als exemples Figura 6 observem el comportament de models de caracterítiques visuals i models d'ordenació temporal Secció 5.1; nogensmeys no hi ha cap limit a quin model es pot aplicar: un model de word spotting podria agrupar segons trobi paraules clau, un classificador hauria d'agrupar categories... Alguns models són "més compatibles"amb certes col·leccions en el sentit de generar espais vectorials amb un bon grau de modularitat depenent la natura de les dades i el model, la çompatibilitat"entre models es visualitza de forma senzilla ja que en models amb poca compatibilitat amb una base de dades; els documents segueixen una distribució uniforme en ambdós eixos (condició de màxima entropia) significant que no hi ha correlació entre els atributs trobats a les dades. En casos de models compatibles amb el dataset, en canvi, s'observen numerosos grups, mostrant modularitat i, per tant, agrupacions segons característiques útils.

5.4.4 Distribucions

També s'han implementat unes visualitzacions exploratives més senzilles; consistent en l'histograma per a la distribució d'etiquetes de la base de dades i la projecció de les mitjanes d'aquestes etiquetes a l'espai de dues dimensions esmentat a la Secció 5.4.3. Els punts de les dues visualitzacions són interactuables i porten a la galeria de l'any en particular esmentada a la Figura 5 i Secció 5.4.2. Això és rellevant ja que en els dos casos una densitat molt alta d'etiquetes a una regió de l'espai 2D pot estar indicant que el model Secció 5.1 no està representant correctament la variabilitat de les dades.

6 Evaluació i Consideracions

6.1 Evaluació tècnica

Per a tenir una visió general del grau de maduresa de la tecnologia, evaluem dos aspectes importants de l'aplicatiu, la seva útilitat com a eina de perfilat de dades i la latència d'algunes de les funcions principals.

6.1.1 Perfilat de Dades

Com s'ha comentat durant l'article a les Seccions 1 3 l'objectiu principal era construir una eina útil a nivell de perfilat de dades. Com observem al "White Paper"pel perfilat de dades [20] aquest projecte s'englobaria dins del contexte de l'exploració i anàlisis de dominis. Tot i que és una mica complicat i pot resultar controversial l'aplicació d'aquestes pràctiques en dades desestructurades i no tabulars, observem que es compleixen les característiques suficients per considerar l'aplicatiu dins de l'anàlisi de dominis Taula 7. Com s'observa l'aplicatiu es capaç de donar cobertura a tres de les quatre particularitats que defineixen aquest contexte del perfilat de dades; donant un excelent servei en termes d'identificació de valors conceptuals, ja que a través de el visualitzador Secció 5.4.3 es poden identificar camps semantics de diferent mena i desde diferents perspectives depenent el model Figura 6.

6.1.2 Anàlisi de Rendiment

És important considerar que, donat que l'aplicatiu gira entorn l'ús d'una base de dades, s'ha d'analitzar la latència de les operacions relacionades, ja que podrà arribar a actuar com a coll d'ampolla de les operacions principals. Als annexos o documents addicionals es podrà trobar el resultat del perfilat de l'intèrpret de `Python` que s'ha realitzat; però de forma sintetitzada s'han plantejat les següents situacions[5]:

[5]Els percentatges (%) poden variar molt lleugerament respecte els annexos donat que diferents execucions porten a lleugeres variances.

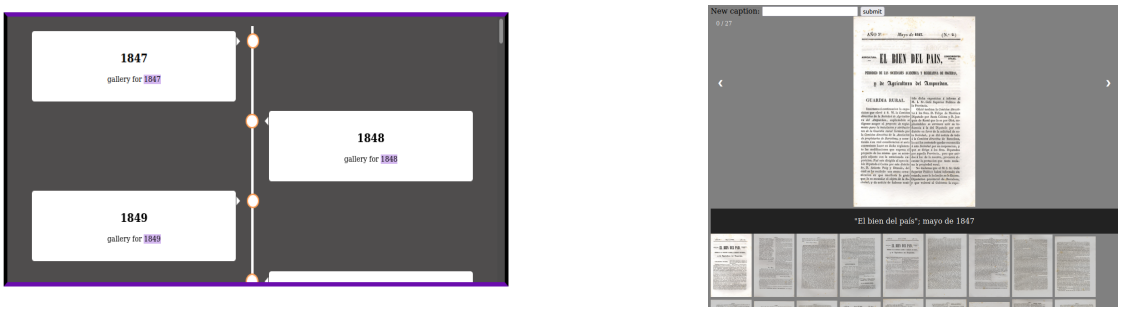

Fig. 5: Prototip de fletxa temporal donada una col·llecció (dreta). Es pot accedir a les galeries d'imatges (esquerra) per any fent click sobre les pestanyes. Aquesta galèria anàloga a sales de mouseus poden contenir descripcions que es guardaran permanentment com a metadades de la imatge.

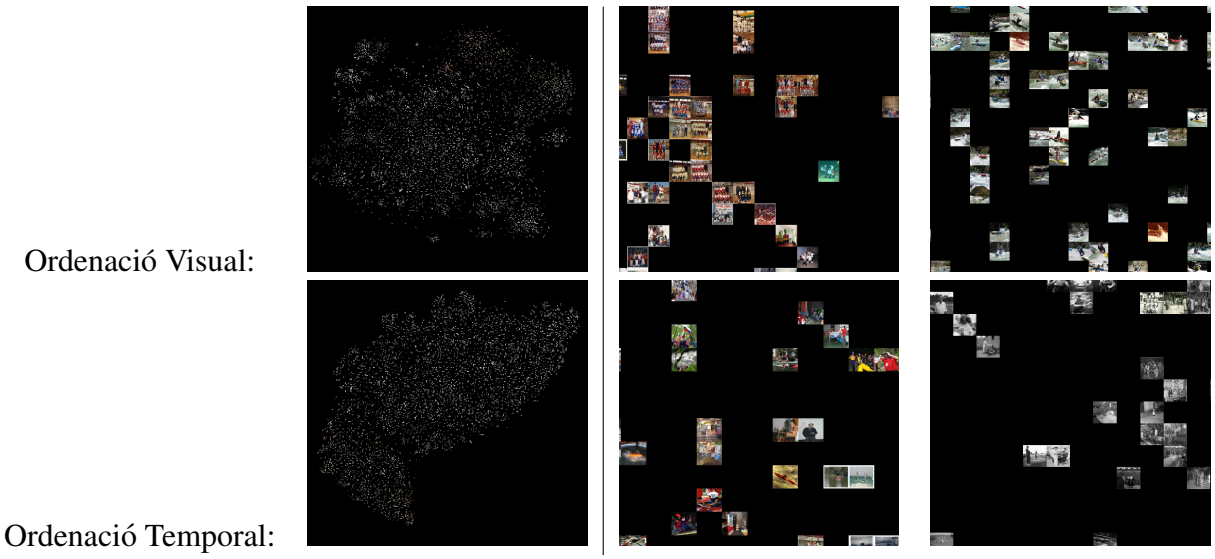

Fig. 6: Visualitzador 2D per col·lecció d'esports a la história. Ordenades segons similaritat visual [13] i segons ordenació temporal [16]. Observem que en el segon cas el contingut semàntic de les imatges no és rellevant per les seves agrupacions, S'observa regions d'imatges plenes de color i gradualment les imatges perden resolució com s'explica a la Secció 5.1
.

Identificació de Dades Referencials	✗
Anàlisi de Rangs Restringits*	✓
Identificació de Valors Conceptuals	✓
Anàlisi d'Estructures Abstractes	✓

Fig. 7: Particularitats de l'anàlisi de dominis en el perfilat de dades cobertes (✓) o no (✗) per l'aplicatiu. *Es pot interpretar com el rang dels anys de les imatges de forma literal o com el rang semàntic proporcionat per la explicabilitat de les dades segons els diferents models.

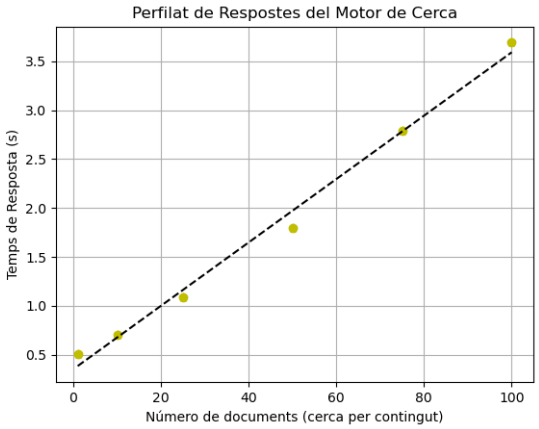

Fig. 8: Perfilat dels temps de resposta del motor de cerca. Observem un ordre lineal $O(n)$ respecte el nombre d'imatges a procesar.

1. **Cerca d'imatges**: Si desgarreguem de la base de dades totes les imatges que corresponen a una `query`, el 99% del temps s'inverteix en aconseguir aquestes imatges.
2. Això provoca que si perfilem en funcionalitats com **la fletxa històrica** Secció 5.4.2 el 81% del temps s'inverteix a aquesta mateixa operació.
3. **Si desglosem aquesta operació**: Observem que desempaquetar les dades serialitzades sols representa un 0.1% del temps invertit. El 77% s'inverteix en llistar els documents indexats a la base de dades en un 5% del temps. Aconseguir la base de dades amb els fitxers serialitzats `GridFS` representa un significant 10% del temps.

Addicionalment, com observem a la Figura 8 el temps de cercar una `query` al cercador Secció 5.4.1 depenent del nombre d'imatges que ens interesa procesar per la cerca visual. Aquest comportament és polinomic, però d'ordre 1 (lineal) respecte el temps. Això és interesant ja que tot i haver marge de millora, es una complexitat que no creix de forma massa dràstica.

6.2 Consideracions

En conclusió, pensem que l'eina compleix com a prototip d'un programari per al perfilat de dades históriques. L'exploració visual marca un precedent prometedor en l'anàlisi de dades a les institucions archivístiques, ja que es pot fàcilment visualitzar una mateixa col·lecció desde diferents perspectives o punts de vista interesants per l'arxiver.

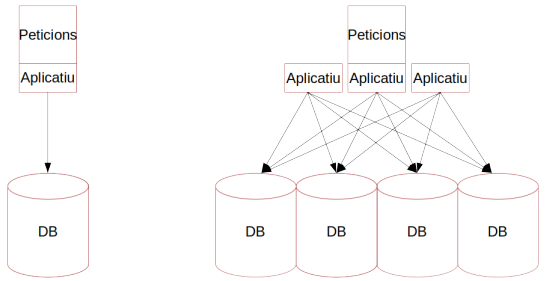

Fig. 9: Infrastructura base (esquerra) i infrastructura proposada (dreta), amb dos costats d'escalabilitat horitzontal. La probabilitat de coincidència d'una petició (sense balanceig de càrrega, totalment estocàstic) escala quadràticament.

En quant la planificació plantejada a l'inici, ha sigut ambiciosa ja que no s'ha pogut construir una aplicació visualment atractiva per falta de temps o priorització de tasques; nogensmenys en no ser un projecte focalitzat en el programari no és una gran pèrdua.

S'ha pogut comprovar que l'aplicació és fàcilment portable a través de contenidors `Docker`, ja que gràcies a les divisions plantejades a la Secció 5.3 és molt senzill desacoblar el sistema en una base de dades d'un servidor extern a la màquina que serveix les funcionalitats a través de `Flask`. Gràcies a aquesta separació es facilita l'escalabilitat horitzontal requerida a la Secció 3 ja que moltes instàncies de HIST4XAC poden comunicar-se a una mateixa base de dades, alliberant les funcionalitats d'una gran càrrega a un sol servidor. De la mateixa manera `MongoDB` té mecanismes propis d'escalabilitat horitzontal a través de `replica-sets`; alliberant així una càrrega desmesurada al mateix servidor de base de dades. Així hem aconseguit escalar horitzontalment ambdues parts, donant peu a un prometedor treball proper de desplegament d'instàncies de l'aplicació per proves pilot i casos d'èxit. De forma simplificada podem observar el següent comportament. Normalment quan escalem una aplicació horitzontalment, dividim la càrrega en N instàncies; fent que la probabilitat de que una instància acomuli una petició pasi de 100% a $100/N$. En tenir aquest comportament d'escalabilitat horitzontal desplegat en dues parts, N instàncies d'HIST4XAC i M de `MongoDB`, quan ambdues convergeixen (en l'infinit), observem que la probabilitat de que una petició caigui a la mateixa instància de la base de dades i aplicatiu disminueix no el doble, si no quadràticament $N \cdot M = O(n^2)$.

Aleshores considerem aquest treball un bon resultat en els termes exposats a la Secció 6.1 i requerits a la Secció 3.

Com s'ha comentat es requereix més desenvolupament en termes d'interfície, bugs i refinament del rendiment de funcionament; així com aprofitar el interesant precedent i facilitats de desmplegament que proporciona l'aplicació; però en termes generals considerem que s'han complert els objectius exposats a l'inici del projecte.

AGRAÏMENTS

Aquest treball ha estat parcialment suportat pels projectes espanyols RTI2018-095645-B-C21 i FCT-19-15244, i els

projectes catalans 2017-SGR-1783, el Departament de Cultura de la Generalitat de Catalunya i el programa CERCA / Generalitat de Catalunya.

Agraïments especials per la supervisió als Drs. Oriol Ramos, Josep Lladós i Lluís Gomez i al Centre de Visió per Computador.

REFERÈNCIES

[1] A. Capellades, "Xiè laboratori d'arxius municipals "de la digitalització al servei digital"," https://arxivers.com/ladada/xie-laboratori-darxius-municipals-de-la-digitalitzacio-al-servei-digital/.

[2] "Interfaces to data for historical social network analysis and research," visited 9-June-2022. [Online]. Available: https://sonar.fh-potsdam.de/

[3] "Infografia de la xarxa d'arius comarcals," last visited 20-June-2022. [Online]. Available: https://xac.gencat.cat/web/.content/xac/05_Que_es_la_XAC/Infografia-XAC-2020rev.pdf

[4] "2021 gartner® magic quadrant™ for cloud database management systems recognizes google as a leader." [Online]. Available: https://cloud.google.com/blog/products/data-analytics/google-named-a-leader-in-2021-gartner-mq-for-cloud-dbms

[5] T. King, "What's changed: 2021 gartner magic quadrant for cloud database management systems." [Online]. Available: https://solutionsreview.com/data-management/whats-changed-2021-gartner-magic-quadrant-for-cloud-database-management-systems/

[6] "5g technology now powers augmented reality on barcelona tourist bus." [Online]. Available: https://www.catalannews.com/tech-science/item/5g-technology-now-powers-augmented-reality-on-barcelona-tourist-bus

[7] J. Andrés, J. R. Prieto, E. Granell, V. Romero, J. A. Sánchez, and E. Vidal, "Information extraction from handwritten tables in historical documents," in *Document Analysis Systems*, S. Uchida, E. Barney, and V. Eglin, Eds. Cham: Springer International Publishing, 2022, pp. 184–198.

[8] J. Andrés, A. H. Toselli, and E. Vidal, "Approximate search for keywords in handwritten text images," in *Document Analysis Systems*, S. Uchida, E. Barney, and V. Eglin, Eds. Cham: Springer International Publishing, 2022, pp. 367–381.

[9] P. Riba, A. Molina, L. Gomez, O. Ramos-Terrades, and J. Lladós, "Learning to rank words: Optimizing ranking metrics for word spotting," in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 381–395.

[10] A. Molina, L. Gomez, O. Ramos Terrades, and J. Lladós, "A generic image retrieval method for date estimation of historical document collections," in *International Workshop on Document Analysis Systems*. Springer, 2022, pp. 583–597.

[11] C. B. Monroc, B. Miret, M.-L. Bonhomme, and C. Kermorvant, "A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents," in *International Workshop on Document Analysis Systems*. Springer, 2022, pp. 429–444.

[12] E. D. M. Council, "Data management capability assessment model (dcam)."

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[16] A. Molina, P. Riba, L. Gomez, O. Ramos-Terrades, and J. Lladós, "Date estimation in the wild of scanned historical photos: An image retrieval approach," in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 306–320.

[17] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman, "Smooth-ap: Smoothing the path towards large-scale image retrieval," in *European Conference on Computer Vision*. Springer, 2020, pp. 677–694.

[18] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Asian conference on computer vision*. Springer, 2016, pp. 198–213.

[19] L. Gomez, A. Bagdanov, and D. Karatzas, "deeparchive.io," http://www.deeparchive.io/.

[20] D. Loshin and DATAFLUX, "The practitioner's guide to data profiling, a dataflux white paper." [Online]. Available: http://hosteddocs.ittoolbox.com/datafluxwp070thepractitionesguidetodataprofiling01112011.pdf