Marçal Rusiñol
Centre de Visió per Computador, Departament de Ciències de la Computació, Universitat Autònoma de Barcelona
marcal@cvc.uab.cat

Article rebut l'octubre de 2018; revisat el novembre de 2018.

Classificació semàntica i visual de documents digitals

Resum: En aquest article donem una visió de conjunt de la classificació automàtica de documents digitals. Veurem de quines maneres es poden descriure tant l'aparença visual com els continguts semàntics i textuals de documents, per poder entrenar models computacionals que siguin capaços de poder classificar, agrupar o fer cerques sobre documents digitals. Resumirem les tècniques clàssiques de la visió per computador i de processament del llenguatge natural, i veurem com els darrers avenços en l'aprenentatge profund (*deep learning*) han revolucionat aquest camp.
Paraules clau: Anàlisi de documents, visió per computador, processament del llenguatge natural, aprenentatge computacional, aprenentatge profund.

Clasificación semántica y visual de documentos digitales

Resumen: En este articulo presentamos una visión de conjunto sobre la clasificación automática de documentos digitales. Veremos de qué maneras se pueden describir tanto la apariencia visual como los contenidos textuales y semánticos de los documentos, para poder entrenar modelos computacionales que sean capaces de poder clasificar, agrupar o realizar búsquedas sobre documentos digitales. Resumiremos las técnicas clásicas de la visión por computador y de procesamiento del lenguaje natural, y veremos cómo los últimos avances en el aprendizaje profundo (*deep learning*) han revolucionado este campo.
Palabras clave: Análisis de documentos, visión por computador, procesamiento del lenguaje natural, aprendizaje computacional, aprendizaje profundo.

Semantic and visual classification of digitized documents

Abstract: This paper presents an overview of the problem of automatic classification of digitized documents. We will see the options available to describe both the visual appearance and the textual and semantic contents of these documents. We will review how these descriptions can be used for the classification, clustering or retrieval of digitized documents. We will summarise the state-of-the-art approaches both from the computer vision and natural language processing fields and will see how the latest breakthroughs in Deep Learning have revolutionized these fields.
Keywords: Document analysis, computer vision, natural language processing, machine learning, deep learning.

Introducció. La visió per computador i el processament del llenguatge natural

L'ús que fa de noves tecnologies la comunitat arxivística, bibliotecària i documentalista és una realitat en gairebé tots els passos de la cadena documental, des de les tècniques de conservació i els sistemes de digitalització fins a les eines informàtiques per a l'anotació de metadades o les bases de dades de consulta que faciliten la difusió de continguts. Tot i això, el procés documental requereix encara una gran quantitat de treball manual, cosa que pot arribar a convertir-se en un coll d'ampolla a l'hora de processar grans volums d'informació.

Gestionar aquests grans volums d'informació pot arribar a fer inviable, per motius de cost i temps, la tasca d'etiquetar i descriure acuradament cadascun dels documents amb les dades adequades. Com a conseqüència, gran part dels continguts dels fons documentals custodiats en biblioteques, hemeroteques o arxius estarà condemnada a romandre inaccessible als sistemes de cerca informàtica. En aquest sentit, un procés automàtic que pogués analitzar documents digitalitzats per descriure'n els continguts podria contribuir a facilitar-hi l'accés, a permetre'n la indexació automàtica i a fer accessibles els documents als motors de cerca.

La intel·ligència artificial ha esdevingut, en els darrers anys, una tecnologia emergent i ubiqua. L'abaratiment de la tecnologia, l'augment de la capacitat de càlcul dels ordinadors i l'accés a quantitats enormes de dades ho han fet possible. Avui dia, utilitzem serveis o productes basats en intel·ligència artificial de manera quotidiana. Quan fem una cerca als buscadors d'internet; quan una web ens recomana articles que ens poden interessar basant-se en els nostres gustos; quan el telèfon mòbil ens avisa al matí que avui trobarem trànsit per arribar a la feina, o quan engeguem el nostre robot aspirador perquè recorri casa nostra mentre no hi som, es fan servir algorismes d'aprenentatge computacional.

En els darrers anys vivim una època daurada de la intel·ligència artificial, bàsicament per l'explosió dels mètodes d'aprenentatge profund[1] (en anglès, *deep learning*) que han revolucionat aquest àmbit, ja que obtenen rendiments fins i tot per sobre dels que assoleixen els humans. Els algorismes d'aprenentatge profund funcionen amb un sistema per capes, simulant el funcionament bàsic del cervell i les neurones. És a dir, el conjunt de capes que formen l'aprenentatge profund representen les neurones del cervell.

Dins del camp de la intel·ligència artificial i de l'aprenentatge profund prenen especial rellevància dues línies de recerca en particular. Per una banda, la visió per computador, que és la disciplina de la informàtica que fa que les màquines hi vegin i puguin analitzar i entendre els continguts d'imatges,[2] i, per l'altra, el processament del llenguatge natural, que és la disciplina informàtica que s'encarrega de tractar computacionalment el llenguatge humà.

En aquest article veurem quin és l'estat de l'art, tant emprant mètodes clàssics com analitzant els darrers mètodes punters en l'àmbit de l'aprenentatge profund, perquè es puguin dur a terme de manera automàtica els processos de classificació, agrupament i cerca de documents digitals. Començarem per descriure les tasques de classificació, agrupament i cerca per després donar una visió de conjunt dels mètodes que ens permeten descriure els documents di-

1. Yann LeCun, Yoshua Bengio, Geoffrey Hinton, «Deep learning» [en línia]. *Nature*, v. 521, n. 7553 (28 May 2015), p. 436-444 (28 May 2015). <https://doi.org/10.1038/nature14539> [Consulta: 20/09/2018].
2. Alicia Fornés; *et al.* «La visió per computador com a eina per a la interpretació automàtica de fonts documentals». *Lligall: revista catalana d'arxivística*, n. 39 (2016), p. 18-44, <https://www.raco.cat/index.php/Lligall/article/view/340142/431080> [Consulta: 20/09/2018].

Un procés automàtic que pogués analitzar documents digitalitzats per descriure'n els continguts podria contribuir a facilitar-hi l'accés, a permetre'n la indexació automàtica i a fer accessibles els documents als motors de cerca.

gitals, tant de manera visual, és a dir, quina és la seva aparença, com a partir dels seus continguts semàntics, és a dir, definint de què parlen.

1. Classificació, agrupament i cerca de documents digitals

Què entenem per classificació, agrupament i cerca d'elements? Quan parlem de classificació de documents[3] entenem que prèviament s'ha definit un nombre finit de categories possibles i el que ha de fer l'algorisme és anar etiquetant de manera automàtica cada un dels elements a processar amb la categoria a la qual pertany. Es pot entendre, doncs, que el procés d'etiquetar els documents amb metadades a l'hora d'indexar-los per contingut es podria automatitzar amb un procés de classificació de documents. Per altra banda, els algorismes d'agrupació[4] troben clústers de documents que siguin similars entre ells sense necessitat de definir prèviament el concepte de tipus de document. En aquest cas, només s'ha d'especificar, generalment, quin és el nombre de clústers diferents que vols que trobi el sistema per agrupar els documents per similitud. Finalment, el procés de cerca de documents per similitud consisteix a involucrar l'usuari en el procés. Donada una col·lecció de documents digitals accessibles, qualsevol usuari pot venir a proposar una consulta. El sistema d'intel·ligència artificial haurà d'ordenar els documents de la base de dades de més a menys rellevants depenent de la consulta de l'usuari. En resum, en sistemes de classificació esperem que el sistema ens retorni etiquetes semàntiques donats els documents d'entrada; en sistemes d'agrupament, esperem que el sistema ens retorni els documents agrupats en diferents clústers significatius; i en sistemes de cerca, esperem que, donada una consulta, el sistema ens retorni una llista ordenada dels documents en funció de la rellevància. En qualsevol dels tres casos, una de les peces fonamentats és com fer-ho perquè la màquina tingui una representació semàntica dels continguts dels documents per a poder-los classificar, agrupar o ordenar de manera automàtica.

Cal tenir en compte també que, dins del camp de la intel·ligència artificial i de reconeixement de patrons, els diferents mètodes se separen en dues branques diferents.[6] Per una banda, hi ha els mètodes supervisats i, per l'altra, els mètodes no supervisats. La diferència entre aquestes dues aproximacions és que en els mètodes supervisats la màquina aprèn a dur a terme la tasca observant anotacions fetes per humans que li han estat donades al començament de l'aprenentatge. Si, per exemple, volem programar un algorisme de classificació d'imatges que sigui capaç de diferenciar entre fotos de diferents animals, haurem de donar a la màquina un conjunt de fotos prèviament anotades i separades de manera manual perquè la màquina pugui aprendre quines són les caracte-

3. Simone Marinai, «Page similarity and classification», En: David Doermann, Karl Tombre (eds.). *Handbook of document image processing and recognition,* New York, [etc.]: Springer, 2014, p. 223-253.
4. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, «Cap. 16: Flat clustering». En: *Introduction to information retrieval,* Cambridge, [etc.]: Cambridge University Press, 2008, p. 349-375, Accés en línia: <https://nlp.stanford.edu/IR-book/pdf/16flat.pdf> [Consulta: 20/09/2018].
5. Ricardo Baeza Yates, Berthier Ribeiro Neto, *Modern information retrieval: the concepts and technology behind search*, Harlow, [etc.]: Addison-Wesley; Pearson, 2011.
6. Christopher Bishop, *Pattern recognition and machine learning*, New York, [etc.]: Springer, 2006.

rístiques visuals que ens diferencien entre fotos de gossos, gats, cavalls i zebres. Per altra banda, si el que volem és tenir un algorisme que ens agrupi una col·lecció d'imatges en funció de la seva similitud, no caldria proporcionar a la màquina tot d'etiquetes descrivint els diferents continguts de les imatges, sinó que la màquina haurà de trobar de manera autònoma i sense cap mena d'informació externa quina és l'agrupació que minimitza la distància entre imatges. En aquest cas, doncs, estaríem parlant d'algorismes no supervisats.

Ara bé, tant si parlem d'algorismes de classificació, com d'agrupament o de cerca, la peça clau perquè els sistemes d'intel·ligència artificial es comportin adequadament és la manera com es representen els documents. Al cap i a la fi, les màquines només saben fer operacions numèriques i, per tant, hem de trobar la manera de poder representar els documents amb nombres. Aquestes representacions numèriques dels continguts dels documents, anomenades descriptors, han de poder permetre que el càlcul de distàncies o similituds entre representacions tingui correlació amb la nostra percepció de la similitud dels documents. Així doncs, donat un document d'entrada, el descriptor extraurà característiques discriminatòries del document per a obtenir-ne una representació numèrica. Aquests descriptors hauran de seguir els principis següents per poder ser útils: si tenim dos documents semblants i en calculem els descriptors, la diferència (distància) entre aquests haurà de ser un valor petit, i contràriament, si calculem els descriptors de dos documents prou diferents, la seva distància haurà de ser elevada. Ara bé, aquesta noció de similitud entre documents pot ser molt variada depenent de l'aplicació final.

Per exemple, imaginem-nos que volem un sistema de classificació de documents en l'àmbit administratiu. Partim de la base que tenim una vintena de formularis i que, per tant, a l'hora de classificar els documents volem veure a quin tipus de formulari correspon el document digitalitzat. En aquest cas, com que les plantilles

estan definides prèviament i com que qualsevol instància d'un formulari concret tindrà la mateixa aparença visual, siguin quins siguin els continguts, necessitarem que la representació numèrica d'aquests documents tingui en compte l'aparença visual que tenen i no tant el contingut textual. D'altra banda, imaginem-nos que volem trobar agrupacions d'articles de diaris que parlin dels mateixos temes. En aquest cas, no ens interessarà quin format tenen aquests documents, però sí que ens interessarà definir la noció de similitud d'acord amb els continguts semàntics textuals dels documents que s'analitzen.

A continuació analitzarem aquestes dues modalitats de representació de documents digitals i veurem alguns exemples de sistemes automàtics de classificació i cerca de documents, tant per contingut com per aparença visual.

ments que codifica de manera jeràrquica, dividint les imatges de documents en una graella amb les densitats dels píxels en cada una de les cel·les. El vector numèric resultant descriu l'aparença visual del document, com si ens el miréssim de lluny.

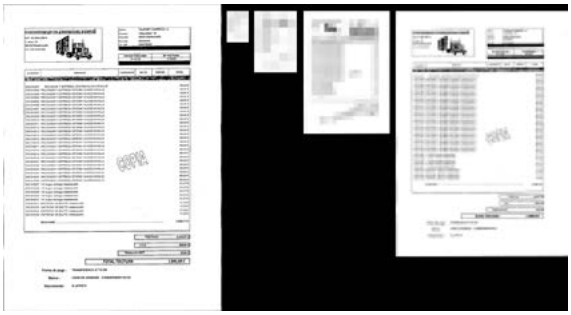

Figura 1. Exemple d'imatge de document i del descriptor jeràrquic visual proposat per Rusiñol, *et al.*

2. Descripció visual de documents

En aquest apartat ens centrarem a estudiar l'estat de la qüestió en les representacions numèriques de documents digitalitzats que codifiquin l'aparença visual dels documents.

2.1. Mètodes clàssics

Els mètodes basats en la visió per computador per a descriure l'aparença visual de documents digitalitzats processen els documents directament en format d'imatge. Sovint els descriptors més simples utilitzen estadístiques calculades a partir de funcions de baix nivell per codificar com es veuen els documents. Per exemple, Rusiñol, *et al.*[7] proposa un descriptor de documents que codifica de manera jeràrquica, dividint les imatges de documents en una graella amb les densitats dels píxels en cada una de les cel·les. El vector numèric resultant descriu l'aparença visual del document, com si ens el miréssim de lluny.

Aquests descriptors simples són útils quan es tracta de problemes en què els documents del mateix tipus són visualment similars encara que els continguts poden variar (p. ex., formularis).

Hi ha mètodes més elaborats que codifiquen la similitud del document en funció de la seva estructura. Les característiques estructurals s'obtenen a partir d'una anàlisi de disseny lògic o físic de les pàgines del document. L'anàlisi del disseny físic descompon les imatges del document en blocs i la similitud del document es pot expressar en termes de les relacions espacials entre aquests blocs. Com a exemple de família de mètodes que descriuen documents en termes de l'estructura dels elements lògics, citarem el treball presentat per Gordo.[8] En aquest cas, la descripció del document codifica com es localitzen els elements lògics i quines són les relacions espacials que tenen.

7. Marçal Rusiñol, *et al.* «Multipage document retrieval by textual and visual representations». En: International Conference on Pattern Recognition (21st: 2012: Tsukuba, Japan). *Proceedings of the 21st International Conference on Pattern Recognition, November* 11-15, 2012. [Tsukuba, Japan: International Association for Pattern Recognition, 2012], p. 521-524, <http://www.cvc.uab.es/~marcal/pdfs/ICPR12.pdf> [Consulta: 20/09/2018].
8. Albert Gordo, *et al.*, «A kernel-based approach to document retrieval» [en línia]. En: IAPR International Workshop on Document Analysis Systems (9th: 2010: Boston, MA), *DAS'10: proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, Massachusetts, USA, June 09-11, 2010*, [New York: Association for Computing Machinery, 2010], p. 377-384, <http://www.cvc.uab.es/~marcal/pdfs/DAS10b.pdf> [Consulta: 20/09/2018].

Els descriptors estructurals són molt més eficients per avaluar la semblança visual entre els diversos tipus de documents que no els mètodes basats en imatges. No obstant això, tenen l'inconvenient que un mateix ha de computar la distància entre les dues estructures de disseny i això és computacionalment car.

2.2. Mètodes basats en aprenentatge profund

Els mètodes basats en aprenentatge profund per a la representació de l'aparença visual d'imatges de documents segueixen una estructura força similar. Un model computacional

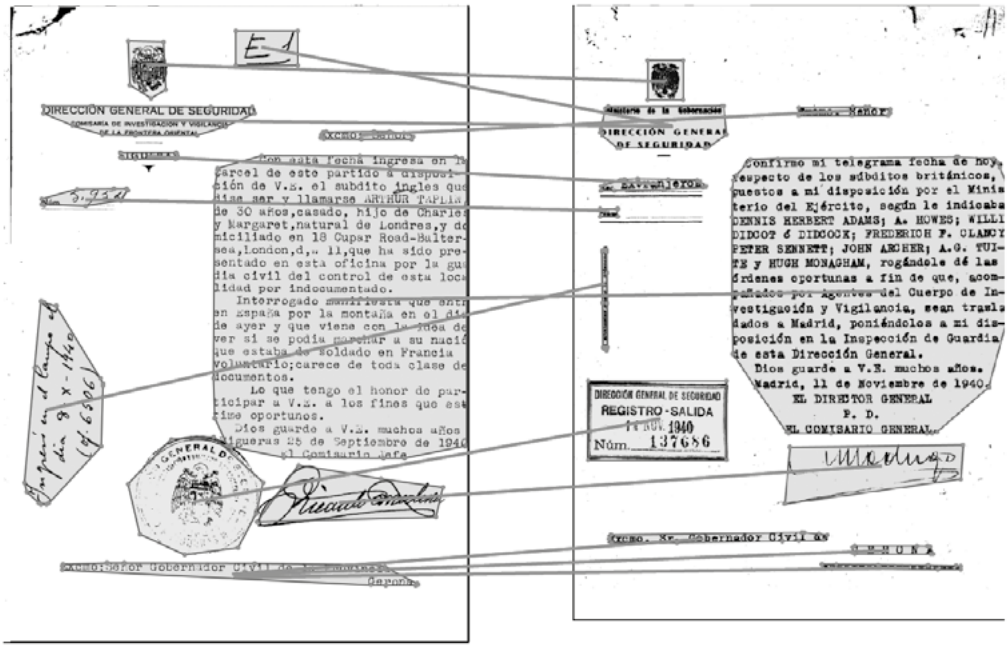

Figura 2. Exemple d'aparellament entre dos documents mitjançant descriptors d'estructura proposat per Gordo, *et al*.

De descriptors visuals de documents se n'han proposat centenars al llarg dels anys, cadascun amb els seus punts forts i punts febles. Recomanem al lector àvid de més detalls que faci un cop d'ull als articles de revisió de l'estat de la qüestió de Chen[9] o de Doermann.[10] Ara bé, com a la resta d'aplicacions de visió per computador, l'explosió de l'aprenentatge profund va revolucionar aquest àmbit i, avui dia, els descriptors d'imatges de documents basats en xarxes neuronals són els que predominen pel seu elevat rendiment.

conegut com a xarxa neuronal s'entrena amb un conjunt etiquetat de dades per a una tasca de classificació. És a dir, necessitem un conjunt força gran de documents etiquetats que es farà servir com a conjunt d'entrenament del sistema. Per poder entrenar de manera efectiva aquests models, normalment es necessiten centenars de milers o, fins i tot, milions d'imatges etiquetades. En aquest àmbit hi ha conjunts de bases de dades públiques i gratuïtes formades per documents prou heterogenis, cosa que permet que es puguin fer aquests entrenaments.

9. Nawei Chen, Dorothea Blostein, «A survey of document image classification: problem statement, classifier architecture and performance evaluation» [en línia], *International Journal on Document Analysis and Recognition*, v. 10, n. (2006), p. 1-16, <http://research.cs.queensu.ca/~blostein/IJDARSurvey2007.pdf> [Consulta: 26/09/2018].
10. David Doermann, «The indexing and retrieval of document images: a survey» [en línia], *Computer Vision Image Understanding*, v. 70, n. 3 (1998), p. 287-298, <https://pdfs.semanticscholar.org/bbf9/b8ece2d8c392e91eded2f2faf8174c7ab265.pdf> [Consulta: 20/09/2018].

La xarxa neuronal, un cop entrenada perquè pugui classificar bé aquests documents, en realitat, ha calculat i extret un seguit de característiques que són les que permeten diferenciar un document d'un altre. Així doncs, independentment de quina tipologia d'imatges de documents es vulgui tractar, o de si es volen els descriptors tant per classificar com per agrupar o fer cerques, podem emprar aquestes xarxes neuronals ja entrenades amb el conjunt de dades públiques per calcular uns descriptors que són molt efectius i discriminatoris.

Per exemple, Csurka, *et al.*[11] van mostrar com es poden entrenar models de xarxes neuronals prou coneguts (com per exemple, AlexNet o GoogleNet) per poder representar característiques discriminatòries en imatges de documents. Aquests models, entrenats amb una base de dades de centenars de milers de documents, han après quines són les característiques visuals importants per poder-los discernir. El pas sorprenent és, però, que aquests models, un cop entrenats, es poden emprar per tractar lots de documents completament diferents dels mostrats en l'entrenament i, tot i així, aportar rendiments, tant en classificació com en cerca, molt superiors als prèviament proposats amb tècniques basades en anàlisi d'imatges.

3. Descripció semàntica dels continguts dels documents

En l'apartat anterior hem entrevist diferents maneres de poder descriure numèricament l'aparença visual dels documents. Ara bé, en molts casos ens interessarà descriure els documents en funció del tema que tracten, és a dir, dels continguts semàntics i textuals en lloc dels purament visuals. En aquest apartat, ens centrarem a estudiar l'estat de la qüestió en representacions numèriques de documents digitalitzats que codifiquin els continguts textuals. Començarem per parlar del reconeixement òptic de caràcters, en aquells casos en què els documents s'hagin hagut de digitalitzar amb un escàner i, per tant, d'entrada no els tinguem en format textual.

3.1. Reconeixement òptic de caràcters

Es pot considerar que l'origen de l'anàlisi de documents mitjançant la visió per computador apareix als anys seixanta, amb els primers sistemes de reconeixement òptic de caràcters (ROC). Els sistemes de ROC parteixen d'una imatge d'un document digitalitzat per produir un fitxer de text electrònic editable. Aquests

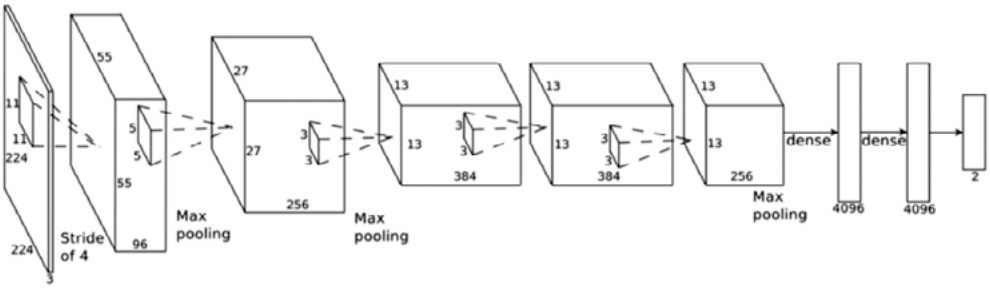

Figura 3. Exemple de l'arquitectura de la xarxa neuronal coneguda com a AlexNet, proposada per a codificar imatges de documents per Csurka, *et al.*

11. Gabriela Csurka, *et al.*, «What is the right way to represent document images?» [en línia], [arXiv preprint (Dec. 5, 2016)], <https://arxiv.org/abs/1603.01076> [Consulta: 20/09/2018].

sistemes integren, en primer lloc, un model de la forma de les lletres i un model lingüístic sobre les probabilitats que aquestes es combinin segons el llenguatge d'escriptura. Per tant, els programes de ROC reconeixen agrupacions de píxels com a lletres i, en un nivell superior, validen les interpretacions conjuntes per acabar transformant una imatge en un arxiu editable de paraules.

El programari de ROC ha evolucionat molt i avui dia té bones prestacions, especialment pel que fa a documents impresos i digitalització de qualitat. Les aplicacions d'ofimàtica i els escàners domèstics acostumen a incorporar un programari de ROC que permet transcriure automàticament els documents quotidians. Comercialment, grans corporacions com Nuance (OmniPage), Abbyy (FineReader) o Google (Tesseract) ofereixen bons sistemes que després altres empreses de serveis adapten a determinats escenaris, com ara el processament postal, la lectura de xecs bancaris i la incorporació de factures a sistemes de planificació de recursos empresarials (ERP).

Ara bé, tenir un fitxer amb el text electrònic automàticament extret, no acaba de solucionar l'accessibilitat sobre col·leccions de documents. Encara necessitem trobar com es poden codificar de manera numèrica continguts textuals per poder classificar documents semànticament. És a dir, poder tenir models computacionals que siguin capaços d'agrupar documents depenent dels conceptes semàntics que tractin.

3.2. De les estadístiques de paraules a la semàntica latent

Pel que fa a la descripció textual, el model de descripció més emprat es basa simplement en el càlcul d'estadístiques bàsiques sobre l'aparició de les diferents paraules en un text. Aquesta representació, coneguda com a «sacs de paraules» (*bag-of-words model*) ja s'emprava als anys cinquanta.[12] Tot i que són simples, aquests models funcionen prou bé en molts casos, encara que comencen a tenir limitacions a l'hora de tractar corpus de dades de certa envergadura. El principal problema radica en el fet que dos documents poden parlar de la mateixa cosa i fer servir paraules completament diferents.

Per fer front a aquesta limitació, apareixen tècniques computacionals que intenten extreure la semàntica latent que hi ha als documents. És a dir, els documents no es defineixen segons les paraules que hi apareixen, sinó segons els temes semàntics de què tracten i de quina manera els tracten. Aquestes tècniques, proposades a principis dels noranta per Deerwester, *et al.*,[13] foren ràpidament adoptades i són a la base de tots els motors de cerca per internet. El seu punt fort radica en el fet que es poden recuperar documents encara que cap de les paraules formulades a la consulta apareguin al text original. Així es van solucionar els problemes que tenien els mètodes de sacs de paraules amb la sinonímia i la polisèmia. Aquests mètodes també tenen el gran avantatge de poder tractar textos de llargada arbitrària, amb la qual cosa es poden processar consultes sobre documents de diverses pàgines.[14]

12. Zellig Harris, «Distributional structure» [en línia], *Word*, v. 10, n. 2 (1954), p. 146-162, <https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520> [Consulta: 20/09/2018].
13. Scott Deerwester, *et al.*, «Indexing by latent semantic analysis» [en línia], *Journal of the American Society for Information Science*, v. 41, n. 6 (1990), p. 391-407, <http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf> [Consulta: 20/09/2018].
14. Albert Gordo, *et al.*, «Document classification and page stream segmentation for digital mailroom applications» [en línia], En: International Conference on Document Analysis and Recognition (12th: 2013: Washington, DC), *ICDAR 2013: proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, 25-28 August 2013*, Washington, DC: IEEE Computer Society, cop. 2013, <http://www.cvc.uab.es/~marcal/pdfs/ICDAR13c.pdf> [Consulta: 20/09/2018].

3.3. Mètodes basats en aprenentatge profund

Tal com ha passat en l'àmbit de la visió per computador, l'aparició de les xarxes neuronals també ha revolucionat el camp del processament del llenguatge natural. Tot i que les tècniques basades en l'anàlisi de la semàntica latent es continuen emprant pel seu bon rendiment i pel fet que no es necessiten dades etiquetades per entrenar-los, ja que es tracta de mètodes no supervisats, apareixen mètodes basats en l'aprenentatge profund que són molt prometedors, com és el cas del mètode de *word2vec* proposat per Mikolov, *et al*.[15]

El mètode *word2vec* és capaç de representar numèricament paraules capturant-ne el significat semàntic. El mètode es basa en una xarxa neuronal que, donades unes quantes paraules d'un text, ha de predir quina serà la paraula següent. Aquesta xarxa arriba a tenir una representació numèrica del significat de les paraules molt rica que, per tant, es pot emprar per a representar de què parlen els documents.

L'aparició de les xarxes neuronals també ha revolucionat el camp del processament del llenguatge natural.

Com a exemple del poder discriminatori a escala semàntica del mètode *word2vec*, ens podem fixar en les relacions entre les representacions vectorials d'algunes paraules. Quan inspeccionem aquests casos, es fa evident que els vectors capturen informació semàntica

15. Tomas Mikolov, *et al*., «Efficient estimation of word representations in vector space» [en línia]. En: Interational Conference on Learning Representations (2013: Scottsdale, AZ), *Proceedings of the International Conference on Learning Representations* (ICLR 2013), [s.l.: ICLR, 2013], <https://arxiv.org/abs/1301.3781> [Consulta: 20/09/2018].

general i, de fet, força útil de les paraules i les relacions que tenen entre elles. És interessant veure que determinades àrees en l'espai vectorial induït s'especialitzen cap a relacions semàntiques, com poden ser les relacions entre masculí i femení, de temps verbals o, fins i tot, relacions de país i capital entre les paraules, tal com s'il·lustra en la figura següent.

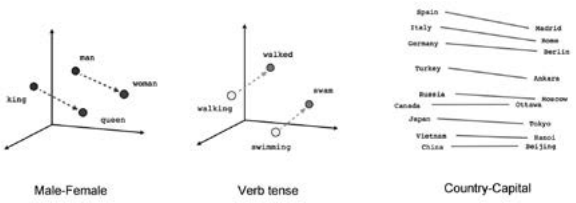

Figura 4. Exemple de relacions semàntiques entre paraules amb el mètode de word2vec de Mikolov, *et al.*

Com que amb el mètode de *word2vec* cada paraula es representa amb un nombre, es poden fer operacions aritmètiques entre paraules per poder visualitzar el poder de representació dels continguts semàntics, com per exemple:

Rei + (home - dona) = Reina

Què codifica una relació del tipus x és a y el que z és a…? Altres exemples d'aquestes relacions tenen a veure, tal com hem dit anteriorment, amb relacions país-capital, etc.

París + (França - Itàlia) = Roma

Einstein + (científic - pintor) = Picasso

Sushi + (Japó - Alemanya) = bratwurst

Microsoft + (Windows - Android) = Google

Aquestes representacions semàntiques numèriques tan riques en significat han revolucionat, de nou, el camp del processament del llenguatge natural, i no sols s'empren per a poder classificar o fer cerques semàntiques en fons documentals, sinó que també són en la base de sistemes de reconeixement de la parla o de traducció automàtica.

4. Conclusions

En els darrers anys s'ha avançat força tant en la visió per computador com en el processament del llenguatge natural, i en particular, en l'anàlisi de documents. L'avenç de les noves tecnologies, tant de programari com de maquinari, hi ha ajudat. Des de la recerca en l'àmbit de les enginyeries s'ha transferit coneixement que s'ha integrat a productes o serveis que avui dia estan integrats en arxius, biblioteques o plataformes de gran consum. Tanmateix, encara hi ha diversos reptes científics i tecnològics de cara a disposar de serveis de lectura universal. El que fa anys era el tractament de documents en estacions individuals, avui dia ha evolucionat cap a grans volums documentals (l'era de les dades massives). Els ciutadans del futur hauran de poder fer cerques per internet darrere de les quals hi haurà milions de documents provinents d'arxius d'arreu del món. Els algorismes han de ser prou eficients per no baixar el rendiment en aquesta nova dimensió. Lligada a aquest accés universal s'obre la necessitat de desenvolupament dels sistemes intel·ligents, és a dir, no només de reconeixement i transcripció literal, sinó també d'assoliment de la capacitat d'interpretació dels continguts. Només així els sistemes podran relacionar termes en diferents tipologies de documents i en diferents llengües. Aquest problema va més enllà de la visió per computador i requereix altres experteses d'intel·ligència artificial, com ara sistemes de modelatge i gestió del coneixement, representacions amb ontologies, etc.

Agraïments

Aquest treball ha estat finançat en part pel projecte TIN2014-52072-P, el programa CERCA de la Generalitat de Catalunya i el projecte aBSINTHE de la Fundació BBVA. Agraïm a NVIDIA Corporation la donació de la GPU Titan Xp, que s'utilitza en el marc d'aquesta investigació.

***Plataforma de préstec en línia d'audiovisuals**, cinema, sèries, documentals, curts, concerts, animació i cursos, creada exclusivament per a les biblioteques.*

EFILM

efilm.online