

IA, recherche d'information et recommandation automatique

Patrice Bellot

▶ **To cite this version:**

Patrice Bellot. IA, recherche d'information et recommandation automatique : La diversification et la transparence des modèles comme rempart à l'uniformisation. Implications philosophiques, 2020, Dossier " Philosophie et numérique ". hal-03521645

HAL Id: hal-03521645
https://hal.science/hal-03521645

Submitted on 11 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

IA, recherche d'information et recommandation automatique

implications-philosophiques.org/ia-recherche-dinformation-et-recommandation-automatique/

April 27, 2020

La diversification et la transparence des modèles comme rempart à l'uniformisation

[box type= »bio »] Patrice Bellot Professeur des Universités en Informatique

Chargé de mission Fouille de textes à l'INS2I CNRS Aix Marseille Univ,

Université de Toulon, CNRS, LIS, Marseille, France

patrice.bellot@univ-amu.fr[/box]

[box type= »info »] Résumé : La recherche d'information et la recommandation correspondent à des processus cognitifs complexes. Pour être efficaces, les approches algorithmiques doivent s'adapter dynamiquement et disposer d'informations nombreuses, notamment sur les contextes d'utilisation et sur les préférences des utilisateurs. C'est ainsi que se sont développées des approches personnalisées d'apprentissage automatique, exploitant de très nombreuses données de façon peu transparente. Nous présentons les approches statistiques les plus populaires, du modèle vectoriel de recherche d'information aux approches connexionnistes à bases de réseaux neuronaux. Nous soulignons les limites inhérentes aux procédures d'apprentissage dans le cadre de tâches subjectives, notamment du point de vue des risques d'uniformisation contre lesquelles une pluralité de modèles est nécessaire. Nous proposons enfin d'étudier le comportement des modèles et des systèmes automatiques sous différents angles, selon des approches pluridisciplinaires.

Mots-clefs : intelligence artificielle, recherche d'information, recommandation, apprentissage automatique, réseaux neuronaux

Abstract: Information retrieval and automatic recommendation of contents correspond to complex cognitive processes. To be effective, algorithmic approaches need to be dynamic and take into account the preferences and the past actions of the users and the contexts of use, among others. This has led to the development of personalized machine learning based approaches, exploiting a large amount of data in a non-transparent way. We present the most popular statistical approaches, from vector-based information retrieval to neural networks. We emphasize the inherent limitations of the learning procedures in

subjective tasks and the risks of standardization against which a plurality of models is necessary. Lastly, we propose to study the behavior of the models and automatic systems according to different points of view, using a multidisciplinary approach.

Keywords: artificial intelligence, information retrieval, recommender systems, machine learning, neural networks.[/box]

Introduction

Les théories et techniques ayant conduit au développement de services de recherche d'information et de recommandation automatique s'inscrivent dans le cadre de l'intelligence artificielle, de la fouille de données et du traitement automatique des langues. Ces tâches cognitives complexes sont abordées par des approches probabilistes dans lesquelles les calculs de similarités et de corrélations jouent un rôle majeur et qui peuvent être modélisées par apprentissage statistique à partir de masses de données. Lorsque l'on cherche à atteindre un objectif par identification et construction d'une référence qui, à défaut d'être donnée *a priori* formellement, ne peut l'être que par observation et échantillonnage statistique, un danger survient qui peut ériger en vérité universelle toute forme majoritaire, au prix d'une uniformisation pour le moins regrettable, puisque le comportement moyen définit à la fois la référence et l'objectif. La captation de données comportementales pose des questions sociétales désormais bien connues parmi lesquelles celles de la protection de la vie privée et du droit à l'oubli. Les internautes ne sont pas de simples consommateurs d'objets en ligne mais aussi des acteurs du fait de leurs interactions avec les services proposés et les autres internautes (avis et notes publics, *playlists* partagées, messages personnels sur les réseaux sociaux…). Ils peuvent être désireux d'exister sur les réseaux en laissant des traces de leurs passages et craintifs des usages qui en sont faits. Les analyses de ces traces sont pourtant cruciales du fait des possibilités de compréhension des usages qu'elles offrent[1].

Dans ce qui suit nous allons tout d'abord réfléchir à la question de la représentation des données dans l'optique de leur utilisation à grande échelle. Nous nous pencherons ensuite sur les approches numériques de recherche d'information et de recommandation automatique et de prescription, briques essentielles des systèmes d'éditorialisation des plateformes culturelles en ligne[2]. La partie suivante sera consacrée aux approches d'apprentissage machine, notamment l'apprentissage neuronal, qui permettent une adaptation aux usages et une personnalisation des modèles numériques. Nous terminerons par des propositions en faveur de la pluralité des modèles et de la poursuite de recherches pluridisciplinaires pour qualifier les comportements des systèmes numériques.

I. Une question de représentation

Si le langage est « l'expression symbolique par excellence »[3], la façon de considérer la faculté de langage, que ce soit pour comprendre ou générer un texte varie. Les méthodes numériques s'opposent aux méthodes dites symboliques en cela qu'elles ne procèdent

pas par manipulation directe de symboles mais par comparaison ou mise en relation entre des représentations de ces mêmes symboles. Par exemple lorsqu'une méthode symbolique produit la forme au singulier d'un mot au pluriel par identification puis suppression de la marque du pluriel, une méthode numérique établit, pour toute forme lexicale envisagée, un score selon une ressemblance par rapport à des exemples fournis de couples singulier-pluriel. Dans le cas d'une méthode symbolique, les règles sont écrites puis les solutions déduites. Même s'il est vrai que les règles symboliques peuvent être apprises ou inférées à partir d'exemples, elles nécessitent en général un travail manuel important et sont peu flexibles.



Dans le cas d'une méthode numérique, les symboles sont représentés par des séries de vecteurs, qui permettent d'estimer des similarités, des corrélations, des degrés d'analogie. Les modèles de données peuvent être associés à des transformations, des exemples d'entrées et de sorties, qui permettent de généraliser par apprentissage automatique des traitements sur des symboles non encore observés, autrement dit d'induire automatiquement des règles de correspondance. Ces dernières sont, selon les approches et les représentations, plus ou moins explicites, c'est-à-dire mettant en relation des variables mathématiques associables à des descripteurs observables ou à des variables cachées latentes difficilement interprétables.

II. Représentations symboliques ou vectorielles

Les descriptions linguistiques traditionnelles considèrent le langage comme une composition d'éléments logiques catégorisés[4] et les textes, le discours, comme un ensemble de propositions et d'unités interconnectées ou de structures grammaticales peuplées d'entrées d'un lexique, obéissant à des règles syntaxiques, le tout plus ou moins dépendant de contextes sociaux, psychologiques, géographiques et temporels. De là peuvent être définies des représentations des phrases en constituants sur lesquels peuvent être appliquées des transformations, insertions, permutations et suppressions ou estimées des similarités. La question consiste à trouver la représentation la plus adaptée, c'est-à-dire à la fois la plus simple et qui conserve le plus d'informations utiles pour réaliser les traitements souhaités.

Dans de nombreux cas l'objectif est d'être capable d'établir des correspondances textuelles, entre des documents et des classes de documents, entre des requêtes et des documents, entre des documents et des profils utilisateurs. En recherche d'information, chaque document et chaque requête peuvent être représentés par des vecteurs dans un même espace multidimensionnel[5] dans lequel chaque direction correspond à un mot

différent de la langue. Il y a autant de dimensions que de mots possibles et la direction d'un vecteur représente la distribution statistique des mots qu'il contient. Pour chaque composante, i.e. pour chaque mot, il est tenu compte de son nombre d'occurrences dans le document et de sa rareté dans la langue ou dans un corpus de référence. Ceci permet de déterminer les mots les plus représentatifs.

III. La similarité, pivot de la pertinence en recherche d'information

Selon l'hypothèse distributionnelle[6], le thème d'un document est déterminé par la distribution des mots qui le composent. La conséquence de cette hypothèse est que la similarité thématique de deux textes peut être estimée à partir des deux vecteurs qui les représentent selon leur produit scalaire ou comme le cosinus de l'angle qui les sépare. Ces représentations et opérations s'avèrent efficaces pour de nombreuses tâches malgré leurs limites évidentes : l'ordre des mots n'est pas pris en compte, on parle de modèle sac de mots, un même thème peut être évoqué en employant des mots différents, un même mot peut avoir plusieurs sens etc. Des extensions des représentations vectorielles ont été proposées au moyen d'estimations variées des composantes des vecteurs ou de projection dans des espaces de dimension réduite (indexation sémantique latente, plongements lexicaux).

Pour la recherche d'information, la pertinence calculée d'un document, c'est-à-dire le score estimé automatiquement par un moteur de recherche, est définie, selon ce modèle, comme étant la similarité entre le vecteur qui le représente et le vecteur requête. Il s'agit d'une vision simplifiée qui s'accorde avec la théorie de la pertinence de Sperber et Wilson[7] cherchant à « déterminer quelle information particulière retiendra l'attention d'un individu »[8]. La pertinence, qui ne tient pas compte de la différenciation perceptive des lecteurs, n'est en aucun cas une estimation de la véracité des informations présentes mais une proximité lexicale induisant une proximité thématique et, par là même, la potentialité d'induire une lecture intéressante. Il s'agit bien d'une approximation puisque l'intention de l'auteur ne peut être entièrement incluse ni définie de façon non ambiguë dans l'accumulation d'indices réduits aux expressions lexicales utilisées. Ceci d'autant plus que les mots ne sont envisagés que comme des signes indépendants les uns des autres, ne faisant référence à aucune représentation mentale. Notons que cet aspect peut être pris en compte au moins partiellement dans les représentations lexicales continues distribuées, appelées plongements lexicaux[9], ou grâce à l'exploitation de ressources sémantiques telles que celles créées par le jeu sérieux JeuxDeMots[10].

Dans la lignée des propositions de L. de Saussure et de T. Wharton[11], nous sommes convaincus que les traces mémorielles engendrent des effets émotionnels à même de modifier la pertinence réelle ou supposée telle d'un document, pertinence qui devrait se doter d'une dimension pragmatique ignorée des modèles de recherche d'information vectoriels dans leur forme classique. Bien-sûr, une dimension pragmatique devrait inclure en outre une contextualisation large de la recherche, allant de l'estimation de l'intention de l'utilisateur (dans quel but cherche-t-il un document ?) jusqu'à la prise en compte de son niveau d'expertise pour estimer l'effort nécessaire à la lecture et de ses connaissances pour évaluer la nouveauté informationnelle. Tout au plus pouvons-nous

souligner que l'apprentissage automatique, comme nous le verrons plus loin, de profils personnalisés de lecteurs, via l'observation des documents lus en réponse à des requêtes, permet de créer des modèles qui incluent implicitement cette dimension pragmatique.

IV. Le filtrage collaboratif pour la recommandation de contenus

Les plateformes de diffusion et de vente exploitent des approches algorithmiques de recommandation automatique dont l'objectif est de deviner à quel point un humain apprécierait un contenu. Ces approches, fondées sur les comportements des utilisateurs et sur leurs jugements, sont abondamment décrites dans la littérature[12]. Si les algorithmes sont connus, ce n'est pas le cas des valeurs des paramètres employés, ni des critères pris en compte et des données recueillies[13]. Ces données sont pour la plupart fournies par les utilisateurs eux-mêmes, soit au fil des usages et des consultations, soit par l'incitation, plus ou moins diffuse, à écrire leurs avis, donner leurs préférences, annoter les contenus et attribuer des notes. Il s'agit d'une forme d'emprise industrielle de plus en plus prégnante qui contraint l'utilisateur à autoriser l'exploitation de ses traces numériques[14]. Cependant, une recommandation automatique indépendante de ces données, c'est-à-dire une recommandation non personnalisée, serait à la fois plus sensible au marché, aux mesures globales d'audience, et reflèterait davantage le caractère limité des catalogues des plateformes[15] (le ratio de titres qui intéressent les utilisateurs dans le catalogue général est la plupart du temps faible). Cela n'empêche que la protection de la vie privée et la maîtrise des données par les utilisateurs[16] sont au cœur de nouvelles approches informatiques qui présentent pourtant encore de nombreux verrous tant scientifiques que techniques et ergonomiques.

L'approche de recommandation par filtrage collaboratif, au sens où l'ensemble des utilisateurs est contributeur, est issue d'une représentation vectorielle des utilisateurs et des contenus du catalogue. L'idée est de regrouper ces vecteurs en une matrice où chaque ligne est associée à un utilisateur et chaque colonne à un contenu. Au croisement d'une ligne et d'une colonne se trouve une valeur numérique qui peut être soit 0 ou 1 selon que l'internaute a accédé ou non au contenu correspondant, soit une note donnée. L'idée est d'estimer, par calcul de similarité entre les lignes de la matrice, quels sont les utilisateurs similaires et, entre les colonnes, quels sont les contenus similaires. Pour un utilisateur, la note qu'il est supposé attribuer est composée à partir des notes des internautes similaires. Ce type d'approche fonctionne bien sûr d'autant mieux que l'on dispose d'un grand nombre de notes.

Des approches de fouille de textes et de données peuvent être appliquées en complément : elles exploitent le texte des contenus, par exemple par reconnaissance de la parole ou indexation des sous-titres pour les films, mais aussi les critiques écrites par les utilisateurs pour en déterminer automatiquement le sentiment positif, négatif ou neutre[17] (les meilleures approches d'analyse de sentiment appliquées à des critiques de film fonctionnent correctement dans 80-90% des cas). Des projections de l'ensemble des données dans des espaces latents permettent en outre de mettre en relation des groupes d'utilisateurs avec des catégories de films[18].

V. Des modèles de données pour l'apprentissage machine

Pour de nombreuses applications, les règles qui permettent d'atteindre les objectifs sont inconnues ou trop complexes[19]. Exploiter des exemples de situations résolues est précieux, pour l'humain comme pour la machine. Ce sont là les bases de l'apprentissage statistique supervisé[20] : apprendre, du fait de la similarité entre les données en entrée, des corrélations significatives entre des observations initiales et les sorties attendues. Les méthodes sont nombreuses et connaissent un engouement nouveau du fait de la généralisation des applications permettant d'accumuler des données textuelles et comportementales et de l'accroissement spectaculaire de la puissance de calcul.

L'apprentissage supervisé est une approche constructiviste dans le sens où les modèles se construisent en observant l'humain en interaction avec la machine, en rejouant des séquences enregistrées ou en disposant de bonnes réponses préétablies *i.e.* d'une base d'apprentissage. Ces données mettent en relation un vecteur d'entrée, ensembles de caractéristiques observables, éventuellement sélectionnées par un humain (pour un texte ce sont les mots ou les caractères) et un vecteur de sortie, la réponse attendue (une classe dans un problème de classification, une série de scores pour la recherche d'information). Apprendre consiste à déterminer les valeurs des paramètres du modèle de façon à optimiser les sorties prédites, c'est-à-dire à minimiser l'écart avec les sorties attendues. Dans l'idéal, les données d'apprentissage sont fournies par des experts mais cela peut avoir un coût prohibitif lorsque plusieurs dizaines de millions d'exemples sont nécessaires. Plusieurs solutions sont envisagées : exploiter des exemples issus de la foule (données du Web et *crowdsourcing*) et profiter d'exemples comparables (apprentissage par transfert).

VI. Des approches plus ou moins transparentes

Étant donné les enjeux liés à leur utilisation, il semble intéressant de distinguer les approches selon leur degré de transparence, leur capacité à être interprétables. Parmi les approches les plus transparentes, celles qui s'appuient sur les arbres de décision sont les plus courantes. Elles consistent à déterminer hiérarchiquement les descripteurs les plus discriminants par minimisation de l'entropie pour décider de la décision à prendre, la sortie du système. Un tel système d'apprentissage peut expliquer à l'utilisateur la décision prise en listant les valeurs des descripteurs qui ont été décisives.

Les approches les moins transparentes sont les approches connexionnistes, dites neuronales car vaguement inspirées des architectures neuronales du cerveau, qui ont été esquissées dès 1943[21] mais sont aujourd'hui les plus performantes pour presque toutes les tâches de traitement de l'information. Puisque ces dernières sont trop complexes pour pouvoir être modélisées par des relations causales directes entre des observations, par exemple les mots d'une phrase à traduire ou les pixels d'une image, et des sorties attendues, l'idée est de faire intervenir des variables intermédiaires organisées en couches cachées. Ces dernières ne correspondent pas à des propriétés interprétables ou directement observables mais à des combinaisons d'observations pour la première couche cachée ou à des combinaisons de variables d'autres couches

cachées à partir de la seconde couche. Lorsque ces couches sont en nombre élevé, on parle de réseau profond. Plus le réseau est profond, plus le lien entre une sortie et une entrée correspond à une combinaison complexe de variables intermédiaires. Ce degré d'expressivité permet de rendre compte de corrélations plus ou moins spécifiques à certaines configurations rares. Plus l'architecture du réseau est complexe, plus il pourra décrire des relations fines et complexes, au prix d'un coût d'apprentissage élevé. Quantité de variantes et d'architectures ont été proposées ces dernières années (auto-encodeurs, réseaux convolutionnels, réseaux récurrents, transformeurs…) toutes plus ou moins applicables à différents domaines, du traitement de l'image au traitement des langues naturelles en passant par la reconnaissance vocale. Le choix d'une architecture ou d'une autre et des valeurs des hyper-paramètres associés sont actuellement des étapes très empiriques, nécessitant une ingénierie très importante et l'exploitation de modèles de langue pré-entraînés, en plus des données de l'application cible, que seuls les plus gros industriels et quelques institutions peuvent développer[22].

Construire (apprendre) un réseau de neurones consiste à estimer la force des associations, à la manière des connexions synaptiques du cerveau, entre des variables d'entrée et des variables de sortie, en passant par un grand nombre de variables intermédiaires. Chaque exemple de la phase d'apprentissage renforce, par activation et désactivation progressive des connexions par rétro-propagation du degré d'erreur constaté[23], les connexions conduisant à la bonne sortie. Une fois l'apprentissage terminé, c'est-à-dire lorsqu'on estime que le réseau fait suffisamment peu d'erreurs, chaque nouvelle entrée provoque, selon les valeurs de ses descripteurs, une réactivation des connexions et conduit, de proche en proche, aux sorties les plus probables.

Expliciter le fonctionnement d'un réseau neuronal reviendrait à fournir l'ensemble des vecteurs neuronaux et de leurs connexions[24] en soulignant que tel ou tel chemin est plus fort qu'un autre, sans que l'on puisse en donner une interprétation autrement que par de très longues combinaisons linéaires entre variables non observables. Un parallèle avec l'humain reviendrait à expliciter les comportements en fonction des connexions neuronales, issues des expériences antérieures et de l'organisation neuronale initiale, généralement aléatoire pour les réseaux artificiels même si cela peut ne pas être optimal[25]. À l'inverse, un système numérique peut être étudié lui aussi sous un angle psychologique (comportement global face à telle ou telle situation), neurologique (étude des connexions entre variables) et psychanalytique en recherchant par l'écoute, une génération automatique de phrases[26], les éléments de l'historique qui expliquent le comportement observé. Bien sûr, interpréter ne suffit pas et il faut être en mesure de corriger les erreurs par une modification des connexions neuronales sans qu'il soit nécessaire de trop multiplier les contre-exemples. Cela revient à savoir comment oublier ce qui a été appris par erreur, par exemple selon des règles et des transformations *a posteriori*[27].

VII. Un objectif plus ou moins subjectif

L'apprentissage par induction pose la question de la justesse et de la représentativité des couples d'entrées-sorties donnés en exemple. Mais c'est aussi le cas de l'apprentissage par renforcement[28], qui permet de façon interactive et itérative de tenir compte des retours des utilisateurs pour modifier le modèle appris. Il est d'autant plus efficace que les sorties erronées identifiées sont représentatives des données générales. Rappelons que l'objectif de l'apprentissage automatique est de réduire en moyenne l'écart entre les sorties voulues et les sorties obtenues automatiquement. Ce processus est viable dans le cas de tâches dont les sorties voulues font l'objet d'un consensus large (reconnaissance de la langue d'un texte, de la polarité du sentiment et du rôles des mots d'une phrase ou en encore validation ou non d'une traduction). Pour ces tâches, les méthodes d'apprentissage profond obtiennent des résultats très supérieurs à toutes les autres approches et, pour certaines d'entre elles, font jeu égal avec l'humain[29].

Pour d'autres tâches, la pertinence des sorties est plus subjective, dépendante d'un contexte difficilement cernable, liée à des opinions divergentes sans qu'elles puissent dites vraies ou fausses ni même peu ou beaucoup pertinentes : c'est le cas de la pertinence d'une liste de documents pour des requêtes difficiles pour lesquelles l'intention et le profil de l'utilisateur sont insuffisamment identifiés, de la prédiction des émotions qui pourraient être ressenties durant la visualisation d'un film mais aussi de toute recommandation employant des modèles non personnalisés. Promouvoir la majorité comme référence possède un caractère normatif incrémental qui tend à reproduire des comportements influençant les internautes, lesquels accentuent alors l'avis majoritaire et ainsi de suite. Malheureusement, lorsqu'aucune majorité ne se dégage et que les désaccords sont trop grands, l'apprentissage ne parvient pas à converger vers un modèle stable et il faut se replier sur des approches plus simples mais plus robustes.

VIII. De la nécessité du pluralisme par la pluralité des modèles

Promouvoir le pluralisme des modèles ne consiste pas seulement à reconnaître et accepter des opinions différentes. De fait, cela est déjà le cas lorsque l'apprentissage est effectué à partir de données non triées ou dans lesquelles seules les opinions les plus extrêmes sont filtrées, parfois automatiquement, souvent manuellement en employant des milliers de modérateurs très rapidement formés. Mais à quoi bon accepter des opinions différentes si elles sont ensuite invisibles ? Bien-sûr, le problème n'est pas simple car rendre visible cette diversité d'opinions nécessite d'abord de les identifier et ensuite de les proposer de façon explicite (sous quels labels ?). Les verrous sont nombreux tant au niveau de l'identification (à partir de quel degré peut-on estimer que deux opinions divergent ? faute de référence sur des classes d'opinion, on est réduit à mesurer des écarts deux à deux) que de l'ergonomie des systèmes : doit-on présenter autant de listes de réponses que d'opinions ? Le choix des données d'apprentissage est fondamental : elles doivent être les plus représentatives possible de la réalité et de sa diversité, et nombreux sont les biais possibles, biais d'échantillonnage (les corpus de données issus des réseaux sociaux correspondent souvent à des données qui n'ont été produites que par une infime partie de leurs membres[30] et qui peuvent s'avérer en outre être de faux messages c'est-à-dire générés par des robots), de genre[31], cognitifs, sociaux ou linguistiques[32] qui doivent être identifiés les uns après les autres, etc.

Apprendre sur un corpus de textes d'une période ou d'une autre, d'un pays ou d'un autre, produit des modèles différents qui auront plus ou moins de mal à se généraliser et à reproduire les biais de l'apprentissage. Apprendre sur l'un ou l'autre des réseaux sociaux, sur Wikipédia ou sur le Web général, produit des modèles reflétant les catégories d'utilisateurs, les usages et les opinions s'y rapportant.

Présenter aux utilisateurs les résultats d'un seul modèle omnipotent est un choix qui doit être remis en question. Un tel modèle ne peut être suffisant car, au-delà des biais d'apprentissage et de l'attraction majoritaire centrale pour assurer la convergence de l'apprentissage, toute approche algorithmique reflète des points de vue sur la nature des données et sur la notion de pertinence. Il ne s'agit pas de problèmes propres aux traitements automatiques ou à l'intelligence artificielle : chaque humain a lui aussi ses *a priori*, ses opinions et sa méthode, ses algorithmes conscients et inconscients. Ça n'est pas tant la mainmise des algorithmes sur notre quotidien qui présente un danger que l'unicité des algorithmes et des modèles, associée à la non perception des orientations idéologiques et morales souvent involontaires induites par les différents biais, les algorithmes et les choix sur les données à observer, conserver ou supprimer.

Conclusion

L'étude des comportements des modèles et des systèmes, notamment face à des situations critiques ou marginales, est une voie pour caractériser, évaluer et interpréter les résultats obtenus. Des études sont conduites dans le cadre de compétitions internationales et selon des jeux de données standards auxquelles participent industriels et académiques[33]. Cependant, elles se concentrent sur des critères de performance en moyenne ne dépendant pas de la nature des données traitées ou des questions sociétales sous-jacentes à certains usages critiques, du degré de gravité des erreurs, de la diversité des opinions et de l'impact de nombreux biais possibles.

Les difficultés pour interpréter les modèles et les sorties des systèmes numériques utilisant l'apprentissage profond viennent du fait qu'ils sont le produit de combinaisons d'un nombre très élevé de variables, elles-mêmes peu ou pas interprétables que par ces mêmes combinaisons. L'étude de telles approches *boîte noire* n'est pas sans rappeler l'étude du comportement humain sous le prisme du seul cerveau dont on ignore le détail de toutes les étapes de sa vie personnelle et pour lequel le schéma des connexions neuronales qui le déterminent en grande partie est trop complexe pour pouvoir être visualisé quand bien même l'on parviendrait à le cartographier numériquement. Nous en sommes réduits à observer les comportements, à estimer la pertinence des réponses. La subjectivité naturelle de certaines tâches rend en outre difficile l'optimisation des modèles, à moins de ne considérer l'opinion majoritaire comme seule porteuse de vérité ou de concevoir des modèles individualisés mais alors intrusifs. Pour les industriels, cela est une incitation supplémentaire à recueillir et à exploiter toujours plus de données personnelles, caractérisant au mieux la variabilité inter-individus. Notons ironiquement que lorsque celle-ci sera réduite à sa portion congrue, c'est-à-dire lorsque les modèles

personnalisés seront similaires au modèle général, il ne sera plus utile d'exploiter des données personnelles et l'on pourra s'enorgueillir de participer à la protection de la vie privée[34].

En conséquence, il semble important de suivre une réflexion spéculative qui imagine l'impact des hypothèses et des biais induits par les approches en œuvre et les vérifie expérimentalement. Une telle démarche permettrait d'établir un profil de chaque modèle, profil psychologique qui devrait être confronté de façon continue à ses états antérieurs et aux profils des autres modèles. Cela constituerait les bases d'une théorie de l'esprit mécanique permettant de représenter, au fil du temps, l'ensemble des croyances, des préférences et des capacités des systèmes automatiques et aux internautes de choisir en toute conscience les services qu'ils préfèrent. Autant de questions soulevées qui ne pourront avoir de réponse satisfaisante qu'en suivant une démarche profondément interdisciplinaire.

---

[1] Françoise Pasquienséguy, « La notion d'usage est-elle stratégique pour les industries créatives ? », *tic&société*, Vol. 4, n° 2, 2010. URL : http://journals.openedition.org/ticetsociete/895 ; DOI : 10.4000/ticetsociete.895.

[2] C. Lipsyc, M. Ihadjadene, « Architecture de l'information et éditorialisation », *Études de communication,* (41), 2013, p. 102-118. DOI : 10.4000/edc.5406, cité par Vincent Bullich et Benoit Lafon, Dailymotion : le *devenir média* d'une plateforme. Analyse d'une trajectoire sémio-économique (2005-2018) », *tic&société*, Vol. 13, N° 1-2, 2018, URL : http://journals.openedition.org/ticetsociete/3540 ; DOI : 10.4000/ticetsociete.3540.

[3] Émile Benveniste, *Problèmes de linguistique générale*, Gallimard, 1966, t. 1, p. 28.

[4] Ferdinand de Saussure, *Cours de linguistique générale*, Payot, 1995 (1re éd. 1916), XVIII-520 p., ISBN 2-228-88942-3.

[5] G. Salton, A. Wong, C.S. Yang, « A Vector Space Model for Automatic Indexing », *Communications of the ACM*, *18*(11), 1975, p. 613-620.

[6] Zellig S. Harris, « Distributional Structure, WORD », 10:2-3, 1958, p. 146-162, DOI: 10.1080/00437956.
1954.11659520.

[7] D. Sperber et D. Wilson, *Relevance. Communication and cognition*, Oxford, Basil Blackwell, 1986.

[8] Cité, d'après la traduction en français, dans Louis Quéré, « La pertinence », Communication et cognition (Don Sperber et Deirdre Wilson), *Réseaux. Communication – Technologie – Société*, 42, 1990, p. 110-111.

[9] T. Kenter, M. de Rijke, « Short Text similarity with Word Embeddings », *Proceedings of the 24th ACM international on conference on information and knowledge management*, ACM, 2015, p. 1411-1420.

[10] http://www.jeuxdemots.org/

[11] Louis de Saussure, Tim Wharton, « La notion de pertinence au défi des effets émotionnels », *TIPA. Travaux interdisciplinaires sur la parole et le langage*, 35, 2019, URL : http://journals.openedition.org/tipa/3068 ; DOI : 10.4000/tipa.3068.

[12] Voir les actes des conférences *ACM Conf. On Recommender Systems*.

[13] Voir tout de même la compétition organisée par Netflix à la fin des années 2000 https://www.netflixprize.com. Voir aussi R. Bell, Y. Koren, « Lessons from the Netflix Prize Challenge », *SiGKDD Explorations*, 9(2), 2007, p. 75-79 et X. Amatriain, J. Basilico, « Recommender Systems in Industry : A Netflix Case Study », in *Recommender systems handbook*, Boston, MA, Springer, 2015, p. 385-419.

[14] Voir par exemple Françoise Paquienséguy, « Le glissement de la prescription dans les plateformes de recommandation », *Études de communication*, 2017/2 (n° 49), 2017, p. 13-32.

[15] Joëlle Farchy, Cécile Méadel, Arnaud Anciaux, « Une question de comportement. Recommandation des contenus audiovisuels et transformations numériques », *tic&société*, Vol. 10, N° 2-3, 2017, URL : http://journals.openedition.org/ticetsociete/2136, DOI : 10.4000/ticetsociete.2136.

[16] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, Joris van Hoboken, « Designing for the better by Taking Users into Account: a Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems », *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*, New York, NY, USA, ACM, 2019, p. 69-77, DOI: https://doi.org/10.1145/3298689.3347014.

[17] N. Kawamae, N., « Predicting Future Reviews: Sentiment Analysis Models for Collaborative Filtering », *Proceedings of the fourth ACM Int. Conference on Web Search and Data Mining*, ACM, 2011, p. 605-614.

[18] Y. Koren, R. Bell, C. Volinsky, « Matrix Factorization Techniques for Recommender Systems », *Computer*, (8), 2009, p. 30-37.

[19] Hubert L. Dreyfus, *What Computers Can't Do*, New York, MIT Press, 1972.

[20] L'apprentissage non supervisé cherche lui à modéliser des ensembles de données, éventuellement en les séparant en différentes classes, indépendamment de toute catégorie établie par des humains.

[21] W. S. McCulloch, W. Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity », *The Bulletin of Mathematical Biophysics*, 5(4), 1943, p. 115-133. DOI : 10.1007/BF02478259.

[22] Voir les modèles distribués par Facebook (https://ai.facebook.com/tools/), ceux construits à partir de Twitter (https://nlp.stanford.edu/projects/glove/) ou de Wikipédia distribué par Google (https://github.com/google-research/bert).

[23] D. E. Rumelhart, G. E. Hinton et R. J. Williams, « Learning Representations by Back-propagating Errors », *Nature*, 323, 1986, p. 533-536.

[24] Des visualisations sont formulées pour certains types de réseaux neuronaux destinés à l'analyse d'images. Par exemple W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.R. Müller, « Evaluating the Visualization of what a Deep Neural Network has Learned », *IEEE transactions on neural networks and learning systems*, 28(11), 2016, p. 2260-2673. Voir aussi D. Erhan, A. Courville, Y. Bengio, « Understanding Representations Learned in Deep Architectures », *Universite de Montréal, QC, Canada, Tech. Rep*, vol. 1355, 2010.

[25] J. F. Kolen, J. B. Pollack, « Back Propagation is Sensitive to Initial Conditions », *Advances in neural information processing systems*, 1991, p. 860-867.

[26] Woon Sang Cho et al., « Towards Coherent and Cohesive Long-form Text Generation », *Proceedings of the First Workshop on Narrative Understanding*, ACL Anthology, 2019.

[27] Z. Hu, X. Ma, Z. Liu, E. Hovy, E. Xing, « Harnessing Deep Neural Networks with Logic Rules », *arXiv preprint arXiv:1603.06318*, 2016.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, S. Petersen, « Human-level Control through Deep Reinforcement Learning », *Nature*, 518 (7540), 529, 2015.

[29] Voir les résultats des compétitions GLUE sur de multiples tâches (https://gluebenchmark.com) et A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S.R. Bowman, « Glue : A Multi-task Benchmark and Analysis Platform for Natural Language Understanding », arXiv preprint arXiv:1804.07461, 2018.

[30] R. Baeza-Yates, « Bias on the Web », *Communications of the ACM*, *61*(6), 2018, p. 54-61, qui souligne, entre autres, que dans un jeu de données issu de Facebook en 2009, 50% des contenus n'étaient produits que par 7% des contributeurs, que dans un autre ensemble issu de Twitter en 2012, 50% des tweets provenaient de 2% de 12 millions d'utilisateurs, et que sur un jeu de données de 2013 provenant d'Amazon, 50% des commentaires sur provenaient de 4% des contributeurs. Ces jeux de données étaient pourtant censés être représentatifs des utilisateurs de ces réseaux sociaux. Que valent alors les modèles appris sur ces données ?

[31] Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan, « Semantics derived automatically from language corpora contain human-like biases », *Science*, 356.6334, 2017, p. 183-186.

[32] 50% des sites Web sont en anglais alors que l'anglais n'est la langue natale que de 5% de personnes.

[33] Voir par exemple TREC (http://trec.nist.gov) organisé par le NIST et le *U.S. Commerce Department*, CLEF (http://www.clef-initiative.eu), SemEval (http://alt.qcri.org/semeval2020/).

[34] Pour une analyse des enjeux de l'utilisation des données personnelles, de leur marchandisation et de leur protection, voir Brigitte Juanals, « Protection des données personnelles et TIC au cœur des enjeux de société et de la mondialisation : les mécanismes d'un contrôle distribué », *tic&société*, Vol. 8, N° 1-2, 2014, URL : http://journals.openedition.org/ticetsociete/1475, DOI : 10.4000/ticetsociete.1475.