Classification supervisée de documents

Jean Beney

LCI, Département Informatique, INSA

04 72 43 82 05, jean.beney@insa-lyon.fr

Les dernières années ont vu une grande amélioration des capacités des ordinateurs et de leurs périphériques. Tous ces gains de temps et de place permettent maintenant d'envisager de gérer électroniquement de grandes quantités de documents.

Gérer des documents, c'est d'abord les organiser pour que leur récupération soit facile. Donc les mettre dans des classes définies par les usagers ou calculées automatiquement. Cette définition recouvre le routage (envoi aux personnes concernées), le filtrage ou distribution d'information (envoi aux personnes intéressées) mais aussi la recherche d'information où le problème se réduit à 2 classes (les documents cherchés et les autres).

Lorsque les classes existent et que l'on dispose d'un assez grand nombre de documents déjà (pré) classifiés, on peut se baser sur ces exemples pour classifier de nouveaux documents : c'est la classification supervisée qui s'oppose d'une part à la taxinomie qui consiste à définir les classes et d'autre part à la classification à partir d'une définition de la classe ou d'index bibliographiques.

En général, rien n'empêche qu'un document appartienne à plusieurs classes; nous parlerons alors de multiclassification. Dans certains cas cependant, chaque document n'appartiendra qu'à une seule classe car ces classes ont été définies de façon à ne pas se recouvrir : c'est la monoclassification.

Si les exemples sont suffisamment nombreux, on les traite tous d'un coup pour obtenir un classifieur définitif (traitement par lots, batch). Sinon, on considérera que le classifieur doit être amélioré dès que possible, c'est-à-dire lorsque d'autres exemples arrivent. C'est la classification adaptative qui permet entre autre de tenir compte des évolutions du vocabulaire (nouveaux termes techniques, expressions à la mode).

Exemples d'applications:

- L'office européen des brevets (EPO) est chargé comme les offices nationaux (en France l'INPI) de certifier l'originalité d'une invention pour la protéger contre d'éventuelles copies. Il reçoit 160.000 demandes par an qui sont examinés par des experts répartis en 549 équipes (une par domaine technique) organisées en 44 directoires. Les demandes doivent être envoyées aux équipes concernées (multiclassification) et cet aiguillage ne peut être fait que par des personnes dotés d'un large savoir scientifique. L'automatisation du routage est facilement envisageable car on dispose de l'ensemble des demandes déjà examinées par les experts donc correctement préclassifiés. Une expérience avec Winnow sur 2000 documents par directoire a donné un taux d'erreur (20%) à peine supérieur à celui actuellement constaté.
- Une société néerlandaise distribue sur CD-ROM toutes les informations disponibles sur la fiscalité aux Pays-Bas. Pour aider les clients, les documents sont multiclassifiés en grand sujets. Semblablement, des portails internet (Yahoo, par exemple) proposent des liens organisés en catégories (immobilier, cinéma, ...), sous-catégories,...
  Pour ces documents fiscaux, il s'est avéré que l'automatisation était très difficile (moins de 60% de réussite) à cause du faible nombre de documents dans certaines catégories et à cause d'incohérences dans les données dues à des changements dans l'organisation des catégories. Le

travail d'automatisation a donc nécessité une redéfinition propre des catégories et un nettoyage de la classification des documents existants.

- Lorsqu'un utilisateur cherche un renseignement sur le web avec les moteurs actuels, il indique des mots dont il pense qu'ils apparaîtront dans les pages qui l'intéresse. Il n'y a donc pas de phase d'apprentissage automatique. Mais on peut imaginer des moteurs plus performants qui tiendraient compte des pages acceptées et refusées par l'utilisateur pour lui en proposer de plus pertinentes. Il s'agirait alors de traitement adaptatif.

Stratégies:

Les méthodes utilisées se basent toutes sur un modèle vectoriel. Dans un espace vectoriel dont les axes correspondent aux termes apparaissant dans les documents, chaque document est représenté par un point dont les coordonnées sur chaque axe est la force du terme dans le document (fonction de son nombre d'apparition). Pour des raisons d'efficacité, on cherche, pour chaque classe, une fonction linéaire de ces coordonnées (score du document) dont la valeur, comparée à un seuil, permettra de décider si le document appartient à la classe ou non.

Plusieurs stratégies sont utilisées pour trouver cette fonction :

- La modélisation consiste à construire, par des calculs statistiques (SBS) ou géométriques (Rocchio), un document-type qui représente la classe. Un nouveau document est alors mis dans la classe s'il est suffisamment proche du document-type, la similitude étant calculé par le produit scalaire.
- La séparation cherche un plan séparant la classe de son complément. Pour y parvenir, Winnow travaille par apprentissage successif (proche des réseaux de neurones) et SVM cherche les documents les plus proches de la limite en résolvant un problème d'optimisation quadratique.
- kNN est une méthode particulière en ce qu'elle ne fait pas d'apprentissage mais cherche les plus proches voisins d'un nouveau document et le classifie en fonction de leurs classes.
- Le boosting consiste à combiner des classifieurs médiocres mais adaptés à des cas particuliers.

Les méthodes par modélisation nécessitent moins de calculs pour l'apprentissage que les méthodes par séparation. KNN ne prend aucun temps à l'apprentissage mais la classification d'un nouveau document lui demande beaucoup plus de temps que pour les autres méthodes car il faut le comparer à tous les autres documents.

Si toutes ces méthodes sont théoriquement parfaites sur des classes séparables et des exemples représentatifs, on constate que dans la pratique Rocchio donne des résultats assez bon même avec peu d 'exemples mais Winnow et SVM sont meilleurs lorsque l'on dispose de beaucoup d'exemples.

Résultat de la classification

Dans la plupart des cas, le résultat d'un classifieur est simplement booléen : tel document d appartient ou n'appartient pas à telle classe c. Ce résultat peut s'accompagner d'une valeur qui représente la confiance qu'on lui accorde.

Dans certains cas, on présentera comme conclusion une liste ordonnée de longueur fixée ou non : soit la liste des documents les plus probables dans chaque classe, soit la liste des classe les plus probables pour chaque document.

Mesure de la qualité

L'évaluation de la qualité du résultat dépend de son type, donc de l'application traitée.

On utilise généralement deux mesures : l'erreur de rappel (proportion de documents pertinents manquants) et l'erreur de précision (proportion de documents sélectionnés à tort). Ces deux mesures varient en sens inverse en fonction du seuil d'acceptation.

Pour évaluer les listes ordonnées, on peut utiliser la moyenne des précisions obtenues pour différentes longueur de la liste. Si l'on se contente d'une classe par document, c'est la précision du premier choix.

Les fonctions d'utilité visent à refléter la satisfaction d'un utilisateur en donnant des points pour chaque document pertinent sélectionné et en en enlevant pour chaque document non pertinent.

Pour la recherche d'information, on regardera le rang du premier document qui contient la réponse à la question posée.

Préparation des données

Un documents est une suite de mots éventuellement structurée par des balises qu'il faut donc représenter par un vecteur de forces. Ce passage consiste à déterminer les termes qui vont représenter le document et à leur attribuer une force.

Extraction des termes

Le plus simple est de considérer chaque mot, tel qu'il apparaît, comme un terme, mais il peut être positif de leur faire subir un traitement linguistique :

- extraction du radical ou d'une forme standard. Cela va diminuer le nombre de termes en groupant des mots au sens évidemment liés mais il faut faire attention aux mots dont les différentes formes ont des sens assez différents (ciseau, ciseaux).
- marquage de la catégorie pour distinguer des homonymes *(visons* est nom ou verbe*)*.
- recherche de groupes de mots (p.ex. *machine à vapeur*).
- extraction de notions : on se base sur une ontologie des notions du domaine à laquelle est associée la liste des synonymes représentant chaque notion (Cf. EuroWordnet). La difficulté est l'ambiguïté introduite par les homonymes.

Des outils existent pour ces traitements mais ils ne sont pas parfaits. Ce qui explique qu'ils n'apportent pas de gain de précision car celui-ci est contrebalancé par les erreurs introduites.

Sélection des termes

Les mots différents étant très nombreux (jusqu'à 1.500.000 dans un ensemble de brevets en Français), il est bon d'en faire une sélection a priori pour accélérer l'apprentissage. Cette sélection se fait par une liste de mots-outils (conjonctions ...) ou sur des critères statistiques : on élimine les mots trop rares (1 ou 2 apparitions) ou trop fréquents. Des méthodes plus sophistiquées (χ^2 , incertitude) permettent d'éliminer les termes qui n'apportent aucune information car ils apparaissent dans de nombreuses classes.

Force des termes

Différentes formules sont utilisables pour quantifier l'importance d'un terme dans un document :

- un booléen indiquant la présence ou l'absence du terme ;
- un entier : le nombre d'apparition du terme ;
- la brutalité de cette mesure peut être atténuée en en prenant le logarithme ;
- pour diminuer l'importance des mots très fréquents, on divisera cette valeur par le nombre total d'apparition du terme ou son logarithme (TF-IDF, LTC).

Conclusion

La classification supervisée de document a fait beaucoup de progrès ces dernières années. Lorsque les documents préclassifiés sont nombreux, les résultats sont comparables à ceux obtenus par des experts.

Dans le cas d'une nouvelle application, l'étude manuelle préalable de milliers de documents coûterait trop chère. Nous cherchons actuellement à diminuer ce coût en sélectionnant les documents qui apportent le plus d'information pour obtenir la même qualité avec beaucoup moins de documents préclassifiés.

Parallèlement, nous espérons que les progrès réalisés dans les techniques linguistiques pourront apporter un réel gain dans la qualité des classifications automatiques.

Bibliographie

Avi Arampatzis, Jean Beney, Cornelis H.A. Koster, Th.P. van der Weide, "Incrementality, Decay, and Threshold Optimization for Adaptive Filtering Systems", The Ninth Text REtrieval Conference (TREC-9), Gaithersburg, Maryland, November 13-16, 2000. http://www.cs.ru.nl/peking/trec9.ps.gz

Cornelis H.A. Koster, Marc Seutter and Jean Beney (2003), "Multi-Classification of Patent Applications with Winnow", Proceedings PSI 2003, Springer LNCS 2890, pp 545-554. http://www.cs.ru.nl/peking/psi2003.ps.gz

Jean Beney and Cornelis H.A. Koster (2003), "Classification supervisée de brevets: d'un jeu d'essai au cas réel", Actes du workshop RI au XXIeme congrès Inforsid, pp.50-59. http://inforsid2003.loria.fr/ActesWorkshopRI.pdf