



Indexation de Documents Manuscrits Offline

Alessandro Vinciarelli

► To cite this version:

Alessandro Vinciarelli. Indexation de Documents Manuscrits Offline. Sep 2006, pp.49-53. hal-00111994

HAL Id: hal-00111994

<https://hal.science/hal-00111994>

Submitted on 6 Nov 2006

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.

Indexation de Documents Manuscrits Offline

Alessandro Vinciarelli^{1,2}

¹ Idiap Research Institute
CP592 - 1920 Martigny (Switzerland)

² Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne (Switzerland)

alessandro.vinciarelli@idiap.ch

Résumé : *Les systèmes de reconnaissance automatique de l'écriture permettent de transformer des collections de documents manuscrits en archives de documents numériques. L'avantage n'est pas tellement la réduction de l'espace nécessaire pour stocker les données, mais plutôt la possibilité d'appliquer les technologies de gestion du contenu normalement utilisées pour des textes numériques tels que pages web et e-mails. Le problème principal dans une telle démarche est que les transcriptions sont généralement bruitées, c'est-à-dire qu'elles sont caractérisées par un taux d'erreur qui peut atteindre, dépendamment des cas, les 50 pour cent. Cet article montre que cela ne constitue pas un problème majeur dans deux cas importants : Information Retrieval et Text Categorization. Une comparaison des résultats obtenus avec les mêmes technologies d'indexation sur les transcriptions manuelles (sans erreur) et automatiques (environ 40% de Term Error Rate) des mêmes documents montre en effet que l'impact sur la performance de l'application finale est négligeable.*

Mots-clés : Information Retrieval, Text Categorization, Modèles de Markov, Indexation, Reconnaissance automatique de l'écriture.

1 Introduction

La reconnaissance offline de l'écriture a atteint des résultats remarquables dans des domaines tels que la transcription des adresses postales ou la lecture des chèques bancaires. La raison principale est que les informations complémentaires (code postal et montant du chèque écrit en chiffres) et les contraintes expérimentales limitent la taille du lexique et la variabilité des données à reconnaître [PLA 00]. La situation est différente en ce qui concerne la transcription automatique de textes génériques. Dans ce cas, le lexique doit nécessairement inclure plusieurs dizaines de milliers de mots, et la seule connaissance *a-priori* disponible est le langage dans lequel le document est écrit. Les résultats des travaux qui s'occupent de reconnaissance de textes montrent un taux d'erreur qui parfois avoisine les 50% [MAR 01, VIN 04c].

Les transcriptions qui en résultent sont donc illisibles pour un être humain, mais elles contiennent quand même suffisamment d'information pour l'application des technologies basées sur l'indexation des textes. En fait, cet article montre que, au moins en ce qui concerne l'*Information Retrieval* (IR) [BAE 99] et la *Text Categorization* (TC) [SEB 02], l'ef-

fet des erreurs sur la performance est négligeable. En effet, l'effort supplémentaire demandé aux utilisateurs, dû à la présence d'erreurs, est acceptable (cf. Partie 4 pour plus de détails). Les deux raisons principales sont les suivantes : premièrement, l'indexation se base sur des statistiques de présence/absence des mots dans les documents et souvent, les mots les plus représentatifs du contenu sont écrits plusieurs fois. La redondance permet donc de faire face à une partie des erreurs de reconnaissance. Deuxièmement, les mécanismes de *query matching* en IR et les modèles de catégorie en TC se basent sur plusieurs mots, et même en présence d'un taux d'erreur élevé, la probabilité de les manquer tous est faible.

Jusqu'ici, les approches de modélisation du contenu des documents manuscrits se sont limitées, à notre connaissance, à l'application du *keyword spotting*, c'est-à-dire à l'identification des mots clés soumis par les utilisateurs dans les transcriptions [KWO 00, MAR 03, RUS 02] ou dans les images mêmes des textes manuscrits [JAI 03, KOL 00, RAT 03, RAT 04, TAN 02, TOM 02, UCH 99]. Cela correspond en IR à une approche booléenne dépassée par d'autres méthodes [BAE 99], mais surtout cela empêche d'utiliser d'autres technologies comme la TC, la sommairisation automatique, la *topic segmentation* et d'autres. Pour cette raison, l'application de l'indexing peut représenter une avancée importante.

Cet article est une synthèse de plusieurs travaux précédents [VIN 04b, VIN 04a, VIN 05b, VIN 05a] contenant tous les détails qui, pour des raisons de place, ne seront pas reportés ici. Dans le reste de l'article, la Partie 2 présente les techniques d'indexation, la Partie 3 décrit le système de reconnaissance de textes manuscrits utilisé, la Partie 4 montre les expériences et résultats et la Partie 5 contient les considérations finales.

2 Analyse des Textes numériques

Un système d'analyse de textes numériques se compose de deux parties : la première est l'*indexation* qui convertit les textes en vecteurs, objets plus faciles à manipuler pour un ordinateur. La deuxième est l'application spécifique qui constitue le but de l'indexation comme par exemple, dans le cas de ce travail, IR [BAE 99] et TC [JOA 02]. Une description exhaustive de IR et TC peut être trouvée en [BAE 99] et [JOA 02] respectivement. Les prochaines sous-sections, nous décriront les techniques principales d'indexation et les

approches normalement utilisés en IR et TC.

2.1 Indexation

Le Vector Space Model (VSM) est l'approche la plus commune en indexation d'un corpus de documents. La raison de son nom est que les documents y sont représentés sous forme de vecteurs dans les quels chaque composante correspond à un mot du *dictionnaire* (voir la suite pour plus de détails). D'autres techniques se basent sur des modèles probabilistiques ou booléens, mais ils n'ont pas l'efficacité du VSM [VAN 79]. Cela peut provenir de la difficulté à trouver les informations pertinentes pour l'entraînement, ou c'est parce que le modèle ne tient pas compte d'informations aussi importantes comme la fréquence des mots dans le documents et force les utilisateurs à exprimer leurs requêtes sous forme d'opérations logiques, ce qui n'est pas toujours facile [BAE 99, VAN 79]. Pour les raisons précédentes, cet article se concentre sur le VSM.

La première opération de l'indexation est le *preprocessing* pendant lequel tout symbol qui ne correspond pas à une lettre de l'alphabet (points, virgules, traits d'union, etc.) est éliminé. Cette opération est motivée par le fait que ces caractères ne sont pas liés au contenu des documents et peuvent donc être négligés. La deuxième opération est le *stopping* qui correspond à la suppression de tous les mots qui sont trop fréquents (ils n'aident donc pas à distinguer entre les documents) ou jouent un rôle purement fonctionnel dans la construction des phrases (articles, prépositions, etc.). Le résultat du stopping est que le nombre de mots dans la collection, ce qu'on appelle la *masse des mots*, est réduit en moyenne de 50%. Les mots à éliminer, connus comme *stop-words*, sont recoltés dans la *stoplist* qui contient en général entre 300 et 400 éléments.

Pour un ordinateur, les différentes *variations morphologiques* d'un même mot, par exemple *cuisine*, *cuisinier* et *cuisiner*, sont des mots différents. Pour cette raison, l'opération suivante dans le processus qui mène à l'indexation, est le *stemming*, c'est-à-dire le remplacement de tous les mots par leur racine (*stem* en anglais). Les termes *cuisine*, *cuisinier* et *cuisiner* sont ainsi remplacés par *cuisin*. Le stemming n'a pas d'impact sur la masse des mots, mais réduit de 30% en moyenne la taille du *dictionnaire*, c'est à dire la liste $T = \{t_1, \dots, t_{|T|}\}$, $|T|$ étant le nombre d'éléments de T , des mots uniques qui apparaissent dans le corpus. Dans ce travail, le stemming a été effectué en utilisant l'algorithme de Porter [POR 80].

L'étape finale de l'indexation est l'extraction de la matrice *terms et documents* A où l'élément a_{ij} est une mesure de l'impact du terme t_i sur le document d_j . Il existe de nombreuses techniques qui permettent de calculer les valeurs a_{ij} (cf. [SAL 88] pour une étude exhaustive), mais en général on a toujours une expression du type :

$$a_{ij} = G(i)L(i, j) \quad (1)$$

où le facteur $G(i)$ est dit *global weight* parce qu'il ne dépend pas du document et qu'il se base sur des informations extraites sur la totalité du corpus. Le facteur $L(i, j)$ est dit *local weight* parce qu'il utilise seulement de l'information contenue dans le document d_j . La conséquence d'une telle

approche est que le même document est indexé différemment dans différentes collections. L'exemple le plus connu est le *tfidf* :

$$a_{ij} = tf(i, j) \cdot idf(i) = tf(i, j) \log \left(\frac{N}{N_i} \right) \quad (2)$$

où $tf(i, j)$ est la *term frequency*, c'est-à-dire le nombre de fois que le terme t_i apparaît dans le document d_j , et $idf(i)$ est l'*inverse document frequency*, c'est-à-dire le logarithme du rapport entre le nombre N de documents dans le corpus et le nombre N_i de documents qui contiennent le terme t_i . Ce schéma d'indexation donne plus de poids aux termes qui apparaissent avec une haute fréquence dans peu de documents. L'idée sous-jacente est que de tels mots aident à discriminer entre textes ayant différent sujet. Le *tfidf* a deux limites fondamentales : la première est que la dépendance de la term frequency est trop importante. Si un mot apparaît deux fois dans un document d_j , ça ne veut pas nécessairement dire qu'il a deux fois plus d'importance que dans un document d_k où il n'apparaît qu'une seule fois. La deuxième est que les documents plus longs ont typiquement des poids plus forts parce qu'ils contiennent plus de mots, donc les term frequencies tendent à être plus élevées. Pour aborder ces problèmes a été proposée une nouvelle technique d'indexation connue comme *Okapi Formula* [ROB 00] :

$$a_{ij} = \frac{tf(i, j)idf(i)}{[(1 - b) + b \cdot NDL(d_j)] + tf(i, j)} \quad (3)$$

où $NDL(d_j)$ est la *longueur normalisée* de d_j , c'est à dire sa longueur (nombre de mots qu'il contient) divisée par la longueur moyenne des documents dans le corpus.

2.2 Information Retrieval and Text Categorization

Une fois qu'un corpus a été indexé, il est possible d'appliquer à ces documents différentes technologies. Cette partie illustre IR et TC.

Un système de IR est censé trouver, dans un corpus donné, les documents pertinents par rapport à une requête soumise en langage naturel par un utilisateur. Cette tâche est accomplie en mesurant la similarité entre la requête, qui n'est rien d'autre qu'un court texte, et les documents du corpus. Cette opération permet d'associer une *Retrieval Status Value* $RSV(q, d)$ à chaque document d qui est calculée comme dans la formule suivante :

$$RSV(q, d_j) = \sum_{t_i \in q} a_{ij}. \quad (4)$$

Cela correspond à sommer les poids a_{ij} correspondants aux termes t_i qui apparaissent dans la requête q . Les documents qui n'ont aucun terme en commun avec la query ont donc une RSV nulle. Une fois que la RSV est disponible pour tous les documents du corpus, il est possible de faire un classement (en ordre descendant). On s'attend ainsi à trouver les documents pertinents en tête. Le système utilisé dans ce travail se base sur la formule d'Okapi qui permet d'obtenir des résultats comparables à l'état de l'art sur les tâches de retrieval plus fréquemment utilisées en littérature [BAE 99] comme benchmark.

○ ○ ○ ○ × ○ ○ ○ × ○ ○ × × ○ ○ ○ ○ ○ × × ○ × × × × ○ × × × ○ × × × ×
 × Does not belong to c
 ○ Belongs to c

_____ Recall = 52.9, Precision = 81.8%

_____ Recall = 94.1%, Precision = 57.1%

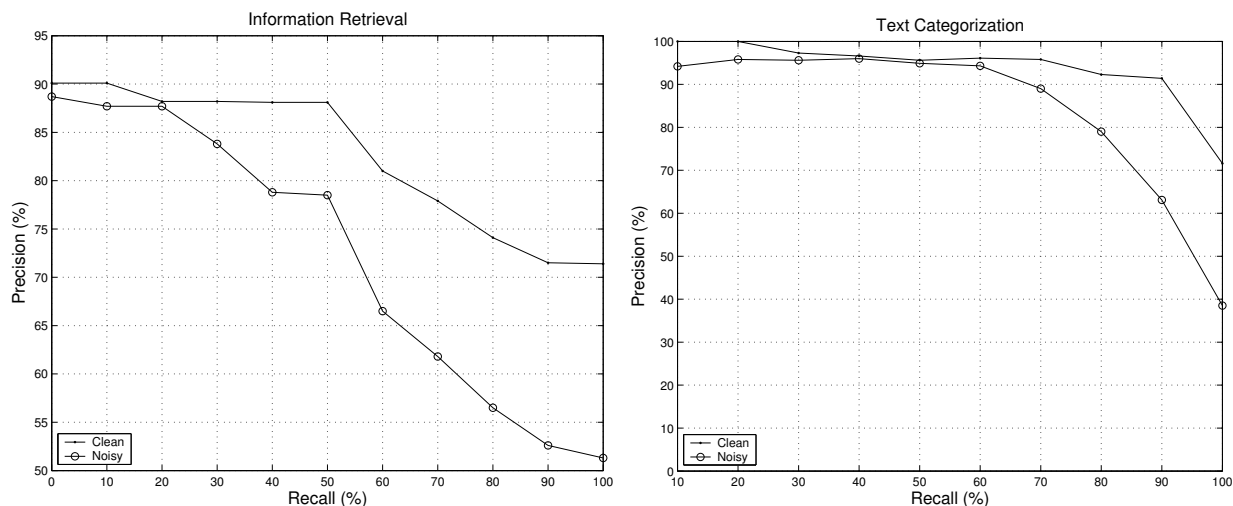


FIG. 2 – Resultats. Le plot à gauche montre les resultats des experience de IR, le plot à droite montre les resultats des experiences de TC.

4.2 IR experiments

L'ensemble des requêtes (query set) Q utilisé dans ce travail contient 20 requêtes (on peut les trouver dans [VIN 05a]) qui ont, en moyenne, moins de quatre documents pertinents. Puisque cela correspond à moins de 2% du corpus, la probabilité de repérer les documents pertinents au hasard est faible et la tâche peut être qualifiée de difficile [BAE 99]. Le même système de IR est testé sur les transcription manuelles sans erreurs (clean) et sur les transcriptions automatiques avec un TER moyen de 40.7% (noisy), les résultats sont illustrés par le graphique gauche de la Figure 2.

Puisqu'on a en moyenne quatre document pertinents, une Precision de 50% à 100% de Recall (comme dans le cas des transcriptions automatiques) signifie qu'on trouve tous les documents pertinents dans les huit premières positions. Dans le cas de la transcription clean, la Precision à 100% de Recall est à peu près 75%, ce qui correspond à dire qu'on trouve, en moyenne, tous les documents pertinents dans les cinq premières positions. Cela signifie donc que les erreurs de transcription n'ont pas un impact majeur sur la perception de l'utilisateur. En fait, les moteurs de recherche les plus communément utilisés présentent les résultats en groupes de dix et, dans le cas des experiences présentées ici, les documents pertinents seraient tous sur la première page dans les deux cas. Pour cette raison, l'effort de l'utilisateur serait à peu près le même.

4.3 TC experiments

Une SVM différente a été entraînée pour chacune des dix catégories les plus représentées dans le corpus Reuters-21578. Cela correspond à une configuration expérimentale normalement utilisée dans la littérature [SEB 02] et permet d'être certain que chaque catégorie est représentée dans le corpus de 200 documents utilisés comme test set. La courbe à droite de la Figure 2 montre le macroaverage des courbes de Precision en fonction du Recall obtenus pour les différentes catégories.

La différence de π est modeste jusqu'à un niveau de Recall de 70%. La raison de l'écart après un tel seuil est que

pour chaque catégorie il y a quelques documents pour lesquels le TER est plus haut que la moyenne et, pour cette raison, se retrouvent à la fin du classement. Cela suffit à baisser de plusieurs points π . En général, ce phénomène concerne des documents courts dans lesquels il n'y a pas beaucoup de termes liés à la catégorie. Quelques erreurs de reconnaissance sont donc suffisants à rendre le texte difficile à classer correctement pour les SVM. Dans le cas des documents de longueur moyenne (ou élevée), même en présence d'un haut TER, il y a toujours plusieurs termes liés à la catégorie qui sont correctement transcrits et cela suffit à les garder à la tête du classement.

5 Conclusion

Ce travail a montré les résultats de l'application de technologies habituellement utilisées pour des textes numériques aux transcriptions des documents manuscrits. Les expériences concernent notamment Information Retrieval et Text Categorization, deux applications qui essaient respectivement de modéliser le contenu des documents afin de repérer les documents pertinents à une requête et d'attribuer les documents à une catégorie prédéfinie. Les transcriptions automatiques ont, en moyenne, un TER de 40.7% (cf. Partie 3). Les résultats montrent que, pour IR ainsi que pour TC, la dégradation due aux erreurs de transcriptions a un impact modéré sur la perception des utilisateurs. Dans la tâche de IR utilisée, chaque requête a, en moyenne, quatre documents pertinents et il peuvent être repérés dans les huit premières positions (cinq premières dans le cas des transcriptions manuelles). Dans le cas de TC, la différence se remarque surtout pour un Recall supérieur à 70% et c'est l'effet de quelques documents qui, à cause des erreurs, se retrouvent à des positions plus basses dans le classement. Les limitations de ces experiences sont au nombre de deux : premièrement, les documents sont écrits par une seule personne, et deuxièmement, le corpus ne contient que 200 documents. Les résultats semblent donc montrer que les techniques d'indexation sont robustes pour des taux d'erreur élevés, mais des confirma-

tions ultérieures sont nécessaires en employant des bases de données plus larges et en incluant plus d'écrivains.

Remerciements L'auteur remercie les organisateurs de la conférence pour l'invitation et Agnès Just pour sa collaboration.

Références

- [APT 94] APTÉ C., DAMERAU F., WEISS S., Automated learning decision rules for text categorization, *ACM Transactions on Information Systems*, vol. 12, n° 3, 1994, pp. 233-251.
- [BAE 99] BAEZA-YATES R., RIBEIRO-NETO B., *Modern Information Retrieval*, Addison Wesley, 1999.
- [BUR 98] BURGESS C., A tutorial on Support Vector Machines for pattern recognition, *Data Mining and Knowledge Discovery*, vol. 2, n° 2, 1998, pp. 121-167.
- [JAI 03] JAIN A., NAMBOODIRI A., Indexing and Retrieval of On-line handwritten documents, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 655-659, 2003.
- [JEL 97] JELINEK F., *Statistical Methods for Speech Recognition*, MIT Press, 1997.
- [JOA 02] JOACHIMS T., *Learning to Classify Text using Support Vector Machines*, Kluwer, 2002.
- [KOL 00] KOLCZ A., ALSPECTOR J., AUGUSTEIJN M., CARLSON R., VIOREL POPESCU G., A line oriented approach to word spotting in handwritten documents., *Pattern Analysis and Applications*, vol. 3, 2000, pp. 153-168.
- [KWO 00] KWOK T., PERRONE M., RUSSELL G., Ink retrieval from handwritten documents, *Proceedings of Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference*, pp. 461-466, 2000.
- [LEW 92] LEWIS D., An evaluation of phrasal and clustered representations on a text categorization task, *Proceedings of 15th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37-50, 1992.
- [MAR 01] MARTI U., BUNKE H., Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, n° 1, 2001, pp. 65-90.
- [MAR 03] MARINAI S., MARINO M., SODA G., Indexing and retrieval of words in old documents, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 223-227, 2003.
- [PLA 00] PLAMONDON R., SRIHARI S., Online and off-line handwriting recognition : a comprehensive survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 1, 2000, pp. 63-84.
- [POR 80] PORTER M., An algorithm for suffix stripping, *Program*, vol. 14, n° 3, 1980, pp. 130-137.
- [RAT 03] RATH T., MANMATHA R., Features for Word Spotting in Historical Manuscripts, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 218-222, 2003.
- [RAT 04] RATH T., MANMATHA R., LAVRENKO V., A search engine for historical manuscript images, *Proceedings of ACM SIGIR conference*, pp. 369-376, 2004.
- [ROB 00] ROBERTSON S., WALKER S., BEAULIEU M., Experimentation as a way of life : Okapi at TREC, *Information Processing and Management*, vol. 36, n° 1, 2000, pp. 95-108.
- [RUS 02] RUSSELL G., PERRONE M., CHEE Y., Handwritten Document Retrieval, *Proceedings of International Workshop on Frontiers in Handwriting Recognition*, pp. 233-238, 2002.
- [SAL 88] SALTON G., BUCKLEY C., Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, vol. 24, 1988, pp. 513-523.
- [SEB 02] SEBASTIANI F., Machine learning in automated text categorization, *ACM Computing Surveys*, vol. 34, n° 1, 2002, pp. 1-47.
- [TAN 02] TAN C., HUANG W., YU Z., XU Y., Imaged Document Text Retrieval Without OCR, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n° 6, 2002, pp. 838-844.
- [TOM 02] TOMAI C., ZHANG B., GOVINDARAJU V., Transcript mapping for historic handwritten document images, *Proceedings of International Workshop on Frontiers in Handwriting Recognition*, pp. 413-418, 2002.
- [UCH 99] UCHIASHI S., WILCOX L., Automatic index creation for handwritten notes, *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp. 3453-3456, 1999.
- [VAN 79] VAN RIJSBERGEN C., *Information Retrieval*, Butterworth, 1979.
- [VIN 04a] VINCIARELLI A., Effect of Recognition Errors on Information Retrieval Performance, *Proceedings of International Workshop on Frontiers in Handwriting Recognition*, pp. 275-279, 2004.
- [VIN 04b] VINCIARELLI A., Noisy Text Categorization, *Proceedings of International Conference on Pattern Recognition*, pp. 554-557, 2004.
- [VIN 04c] VINCIARELLI A., BENGIO S., BUNKE H., Off-line Recognition of Large Vocabulary Cursive Handwritten Text, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, n° 6, 2004, pp. 709-720.
- [VIN 05a] VINCIARELLI A., Application of Information Retrieval Techniques to Single Writer Documents, *Pattern Recognition Letters*, vol. 26, n° 14-15, 2005, pp. 2262-2271.
- [VIN 05b] VINCIARELLI A., Noisy Text Categorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, n° 12, 2005, pp. 1882-1295.