IRCCyN/IVC **P9-S1**

Sujet de master recherche « Architectures logicielles distribuées » 2005-2006

Indexation d'une base de données de documents manuscrits
Identification du scripteur

Encadrant principal :
Emilie CAILLAULT
Courriel : Emilie.Caillault@univ-nantes.fr

Co-encadrant :
Christian VIARD-GAUDIN
Courriel : Christian.Viard-Gaudin@univ-nantes.fr

Tél : 02.40.68.30.40
Ecole Polytechnique de l'Université de Nantes

Objectif du stage :

On observe aujourd'hui l'émergence d'un nouveau type de documents, les documents manuscrits en-ligne, ou dynamiques. Ceux-ci sont produits à partir des nouveaux dispositifs de saisie que sont les tablettes PC et les papiers numériques couplés à l'utilisation de stylos digitaux. La production et la gestion d'une quantité de plus en plus grande de tels documents dans un contexte de système de traitement de l'information supposent des fonctionnalités d'indexation. En particulier, une des informations importantes à associer à un document à archiver dans une base de données concerne l'identité de son auteur. C'est précisément l'objectif initial d'un tel projet.

Nous souhaitons développer des modèles de recherche d'information dans des bases de données constituées de documents manuscrits en-ligne pour la tâche d'identification du scripteur à partir de requêtes constituées elles-mêmes d'un échantillon d'écriture manuscrite en-ligne.

Travail à réaliser :

La problématique envisagée peut se définir selon deux niveaux [Sai 00] :

- une tâche d'identification du scripteur parmi un ensemble de scripteurs connus du système,
- une tâche de vérification, consistant à déterminer si oui ou non, deux échantillons d'écritures proviennent de la même main.

Nous souhaitons aborder successivement ces deux tâches complémentaires. Pour la première fonction, nous souhaitons la modéliser explicitement comme un processus de Recherche d'Information (RI) et envisager les différents modèles proposés dans ce cadre [Son 99]. Le modèle vectoriel (VSM : Vector Space Model) développé par Salton apparaît ici tout à fait applicable [Fen 03], mais d'autres approches pourraient aussi être étudiées (modèle booléen, modèle probabiliste). Nous devons également définir une métrique efficace définissant la distance entre une requête et un document de la base. Nous proposerons des métriques adaptées à ce problème en nous appuyant sur des références dans ce domaine (Cosinus, Okapi, …). La seconde tâche, celle de vérification, est plus spécifique. Elle imposera sans doute une caractérisation plus fine des styles d'écritures dans les deux documents (la référence et la requête).

En amont de ce travail d'indexation, il sera primordial de construire un espace de représentation de l'écriture qui permette d'atténuer la variabilité intra-scripteur tout en conservant une bonne différenciation de la variabilité inter-scripteur. Les caractéristiques qui pourront être retenues seront aussi bien de type texturel que structurel. L'expertise que nous possédons dans le domaine de le reconnaissance de l'écriture manuscrite sera mise à profit pour développer cette partie.

Mots-clés : *Ecriture manuscrite en-ligne, indexation, segmentation, reconnaissance des formes, apprentissage.*

Pré-requis : bonne maîtrise des environnements de développement en Java ou C++.

Possibilité de thèse en continuité de ce travail.

Références :

[Ben 04] Bensefia A., Paquet T., Heutte L., "Handwriting Analysis for Writer Verification", International Workshop on Frontiers in Handwriting Recognition", pp. 196-201, Tokyo, oct. 2004.

[Fen 03] Feng D., Siu W.C., Zhang H.J., "Multimedia information retrieval and management", Springer edition, 2003.

[Sai 00] Said H.E.S., Tan T.N., Baker K.D. "Personal identification based on handwriting", Pattern Recognition, vol. 33, pp 149-160, 2000.

[Sch 04] Schomaker L., Bulacu M., Franke K., "Automatic Writer Identification Using Fragmented Connected-Component Contours", International Workshop on Frontiers in Handwriting Recognition", pp. 185-190, Tokyo, oct. 2004.

[Sch 04] Schomaker L., Bulacu M., "Automatic Writer Identification Using Fragmented Connected-Component Contours and edge-based features of upper-case western script", IEEE Trans. On PAMI, 26(6): 787-798, 2004.

[Son 99] Song F., Bruce W., "A general language model for information retrieval", Eight International Conference on Information and Knowledge Management (ICIKM'99).