# Extension automatique d'annotation et classification de documents en utilisant un modèle graphique probabiliste[1]

**Abdessalem Bouzaieni**[*,**] **– Sabine Barrat**[***] **– Salvatore Tabbone**[*]

[*] *Université de Lorraine-LORIA, UMR 7503, Vandoeuvre-les-Nancy, France*
*Email: {abdessalem.bouzaieni,tabbone}@loria.fr*
[**] *Xilopix, Épinal, France*
[***] *Université François Rabelais- LI, Tours, France*
*Email: sabine.barrat@univ-tours.fr*

RÉSUMÉ. *Avec la prolifération des images de documents, l'annotation de documents est devenue un domaine de recherche d'un grand intérêt. L'annotation permet de décrire le contenu sémantique des documents et facilite leur utilisation et leur recherche. Toutefois, pour un grand nombre de documents, l'annotation manuelle de chaque document devient une tâche fastidieuse. Une solution est d'annoter une petite partie des documents et d'étendre automatiquement les annotations à l'ensemble du jeu de données. Dans cet article, nous proposons un modèle pour l'extension d'annotations et la classification de documents en utilisant un modèle graphique probabiliste. Dans ce dernier, nous combinons des caractéristiques visuelles et textuelles et montrons que l'intégration du retour utilisateur améliore la qualité de l'annotation.*

ABSTRACT. *With the fast growth of document images, document annotation has become a research area of great interest. Annotation allows to describe the semantic content of documents and facilitates their use and research. However, for a huge number of documents, the manual annotation of each document becomes a tedious task. A solution is to annotate a small part of the documents and to extend it automatically to the whole dataset. In this paper, we propose a model for annotation extension and document classification using a probabilistic graphical model. In this latter, we combine visual and textual characteristics and we show that the integration of the user feedback improves the annotation step.*

MOTS-CLÉS : *Annotation, classification, modèle graphique probabiliste, retour utilisateur.*

KEYWORDS: *Annotation, classification, probabilistic graphical model, user feedback.*

1. Traduction d'un article publié en anglais à ICDAR 2015 (Bouzaieni *et al.*, 2015a)

## 1. Introduction

La quantité de documents numériques stockée chaque jour par les entreprises et les archives personnelles sont en croissance permanente. En collaboration avec le web, ces archives atteignent des tailles inimaginables. La popularité de ces grandes collections de documents numériques dépend de leur facilité d'utilisation. Cependant, toutes ces bases de données ne sont pas souvent équipées d'informations d'indexation adéquates. Cela rend beaucoup plus difficile l'accès à des informations intéressantes pour l'utilisateur. Ainsi, pour effectuer la recherche d'information sur les images de documents, un mécanisme approprié est nécessaire pour caractériser le contenu du document de manière significative. Pour les images de documents numérisés, l'accès au niveau contenu se faisait traditionnellement à l'aide d'outils de reconnaissance optique des caractères (OCR). Cependant, malgré des efforts considérables, les OCR robustes ne sont pas disponibles pour beaucoup de types de documents. Le texte obtenu est, par conséquent, mal adapté pour l'indexation et la reconnaissance. Les inconvénients des OCR peuvent être surmontés par une approche fondée sur l'annotation manuelle en attribuant des mots-clés pertinents à un document pour décrire son contenu sémantique. L'annotation manuelle étant performante mais très coûteuse pour l'humain, une solution pour ce problème peut être d'annoter partiellement des documents et d'étendre automatiquement les annotations aux autres documents.

La tâche d'annotation des documents est d'un grand intérêt pour les utilisateurs. Elle permet de mieux comprendre le contenu sémantique des documents et permet aux utilisateurs une recherche plus rapide et plus robuste des documents. Contrairement aux documents du web, pour lesquels plusieurs outils d'annotation sémantique ont été développés comme SMORE (Kalyanpur *et al.*, 2006), Annotea (Kahan *et al.*, 2002) et Semtag (Dill *et al.*, 2003), moins d'efforts ont été consacrés à l'annotation des images des documents par mots-clés.

Coüasnon et al. (Coüasnon *et al.*, 2007) ont proposé deux types d'annotations pour les documents d'archives manuscrits : annotations textuelles et annotations géométriques. Les annotations textuelles représentent toutes sortes d'informations sur lesquelles il est intéressant de faire une recherche (date, place, nom, *etc*). Les annotations géométriques représentent une position dans l'image comme une cellule, un champ, ou une zone, représentée par un rectangle ou un polygone. Pour produire automatiquement ces annotations, il est nécessaire, d'abord, de détecter les régions d'intérêt qui contiennent des informations intéressantes. Ensuite, un système de reconnaissance de l'écriture est appliqué. Li et al. (Li *et al.*, 2014) ont proposé un cadre d'apprentissage itératif pour l'étiquetage des symboles graphiques manuscrits. Un graphe relationnel entre les segments est construit. Les nœuds du graphe représentent les segments et les arcs représentent les relations spatiales entre ces segments. Ensuite, les segments sont regroupés pour construire un dictionnaire de codes visuels. Enfin, l'utilisateur donne des étiquettes à ces groupes, ce qui permet de diminuer l'effort manuel. Dans (Duthil *et al.*, 2014), les auteurs ont présenté une nouvelle approche d'annotation des documents administratifs. La méthode est basée sur l'annotation sémantique des documents suivant le texte contenu dans le document et/ou le logo qu'il contient.

La première étape consiste à extraire les logos contenus dans le document. Dans la seconde étape, chaque logo fait l'objet d'une requête sous forme d'image dans un moteur de recherche web (Google Images[1]) pour identifier le nom du logo. Un ensemble de documents web ayant un contenu similaire au logo requête est récupéré pour construire un lexique de mots. Dans la dernière étape, le document de départ est annoté avec le lexique des mots construit selon les logos qu'il contient. Chakravarthy et al. (Chakravarthy *et al.*, 2006) ont proposé l'outil AKTiveMedia pour l'annotation de documents textes, images et html. Cet outil utilise à la fois des annotations à base d'ontologies et de texte libre. Les annotations libres sont faites par l'auteur et le lecteur du document. Ces annotations sont ensuite complétées par des ontologies.

L'inconvénient majeur de ces méthodes est que les modèles proposés ne sont utilisés que pour la tâche d'annotation et non étendus dans un cadre de classification.

Par ailleurs, le problème de la classification de documents (Sharma et Gupta, 2012), (Shah et Mahajan, 2012), (Chen et Blostein, 2007) a été traité dans la majorité des travaux de recherche en utilisant l'information textuelle et/ou structurelle. Cependant, l'extraction de ces informations peut être compliquée ou irréalisable pour diverses raisons : documents de mauvaise qualité, en différentes langues, présentant peu d'information textuelle, *etc*. Dans ce cas, il est préférable d'utiliser des caractéristiques visuelles ou de combiner les caractéristiques. Malgré le grand effort dans les travaux de recherche concernant la classification des documents, peu de travaux utilisent une combinaison des caractéristiques visuelles et textuelles, contrairement au domaine des images, où les caractéristiques visuelles sont largement exploitées (Bouzaieni *et al.*, 2015b). Par exemple, Barrat et al. (Barrat et Tabbone, 2010) ont proposé un modèle de mélange de lois de Bernoulli et de mélanges de Gaussiennes (GMM-B) pour la classification et l'annotation des images. Dans ce modèle, l'ensemble des caractéristiques visuelles (variables continues) suit une loi dont la fonction de densité est une densité de mélange de Gaussiennes, et les variables discrètes (mots-clés) suivent une distribution de Bernoulli. Le travail présenté dans cet article est dans le même esprit que (Barrat et Tabbone, 2010) mais les modèles sont différents sur trois niveaux. Notre modèle est utilisé pour les images de documents alors que le modèle GMM-B est utilisé pour les images naturelles. Les caractéristiques visuelles utilisées dans les deux modèles sont différentes. La structure de notre modèle est différente de l'autre modèle et a l'avantage de réduire la complexité du réseau.

Dans (Kumar *et al.*, 2013), les descripteurs SURF (Bay *et al.*, 2008) sont utilisés pour calculer les caractéristiques visuelles d'un document. Le document est segmenté en plusieurs régions. Les relations spatiales entre ces régions sont représentées par un histogramme des mots de code visuel. Les caractéristiques visuelles et spatiales sont utilisées par un classificateur de forêt aléatoire pour la classification et la recherche des documents. Dans (Chen *et al.*, 2006), des descripteurs de couleurs et de texture calculés sur des figures contenues dans les documents sont utilisés pour construire un dictionnaire de mots visuels. Chaque sous figure est représentée par un mot visuel.

1. https ://images.google.com/

Ainsi, le document peut être décrit comme une séquence de mots visuels. Ces derniers sont utilisés par un classificateur naïve Bayes pour la classification de documents. Rusinõl et al. (Rusinõl *et al.*, 2012) ont présenté une nouvelle méthode de recherche des documents en utilisant des caractéristiques visuelles et textuelles. Les caractéristiques visuelles sont représentées par les descripteurs SIFT (Lowe, 2004), et les caractéristiques textuelles sont représentées par un vecteur de sac de mots fournis par un OCR. Des documents similaires sont récupérés en utilisant la distance classique du cosinus.

Dans cette direction, la première contribution de ce papier est de proposer un seul modèle pour deux tâches différentes : l'extension d'annotations et la classification de documents. La deuxième contribution est l'intégration de l'utilisateur dans la phase d'apprentissage pour réduire le coût de l'effort humain à travers de ce que nous avons appelé "apprentissage dans l'apprentissage". Nous présentons notre modèle d'annotation des documents en utilisant un modèle graphique probabiliste. Ce modèle permet de combiner des caractéristiques visuelles et textuelles afin d'étendre l'annotation aux documents partiellement annotés. Ce modèle ne nécessite pas que tous les documents de l'ensemble d'apprentissage soient annotés et peut gérer le problème des données manquantes. Le modèle peut également être utilisé pour la tâche de classification des documents en utilisant les caractéristiques visuelles et textuelles.

L'article est organisé de la manière suivante. Dans la section 2, nous présentons notre modèle d'annotation et de classification en utilisant un modèle graphique probabiliste. La section 3 est consacrée aux résultats expérimentaux et nous donnons des conclusions et les perspectives de recherche de notre travail dans la section 4.

## 2. Classification et annotation des documents

Dans cette section, nous détaillons notre méthode d'annotation et de classification de documents en utilisant un modèle graphique probabiliste.

Le modèle proposé est un mélange de distributions multinomiales et de mélanges de Gaussiennes. Le modèle proposé est présenté dans la Figure 1. Nous supposons que les caractéristiques visuelles sont considérées comme des variables continues. Elles suivent une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Les caractéristiques textuelles (mots-clés) sont considérées comme des variables discrètes. Elles suivent une distribution multinomiale. Les caractéristiques visuelles d'un document sont représentées par deux nœuds :

– Le nœud *Gaussian* est modélisé par une variable aléatoire continue qui est utilisée pour représenter les descripteurs calculés sur le document.

– Le nœud *Component* est modélisé par une variable aléatoire cachée qui est utilisée pour représenter le poids des Gaussiennes utilisées. Il peut prendre $g$ valeurs correspondant au nombre de Gaussiennes utilisées dans le mélange.

Les caractéristiques textuelles (mots-clés) sont modélisées par $N$ nœuds discrets, où $N$ est le nombre maximum de mots-clés utilisés pour annoter un document. Des

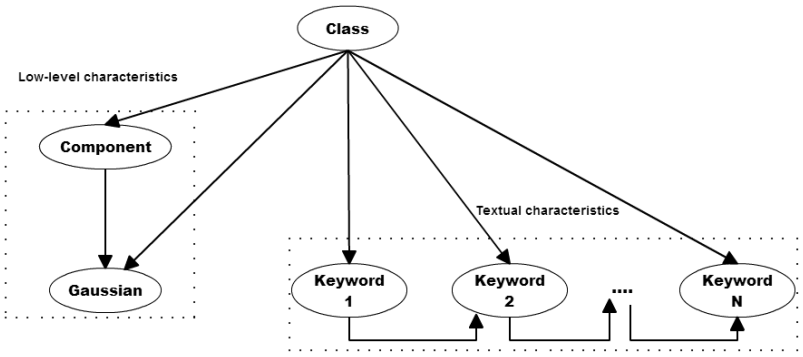


**Figure 1.** *Modèle de mélange de lois multinomiales et de mélange de Gaussiennes*

arcs sont ajoutés entre les $N$ nœuds pour représenter les dépendances conditionnelles entre les mots-clés. Un nœud racine *Class* est utilisé pour représenter le type de document, il peut prendre $k$ valeurs correspondant aux classes prédéfinies $C_1, ..., C_k$.

Pour un tel réseau, on peut écrire la probabilité jointe :

$$P(C, TC, LLC) = P(C) \prod_{i=1}^{M} P(LLC_i|C) \prod_{i=1}^{N} P(TC_i|C) \quad [1]$$

où TC représente les caractéristiques textuelles (les mots-clés $Kw_1, ..., Kw_N$) et $LLC_1, ..., LLC_M$ représentent les caractéristiques de bas niveau (caractéristiques visuelles).

Soit $Kw_1, ..., Kw_N$ l'ensemble des mots-clés dans un document. Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ peut être représentée par un espace vectoriel booléen des mots du vocabulaire :

$Kw_j = \{m_1, ..., m_n\}$, où $m_i = 0$ ou $1, \forall i \in \{1, ..., n\}$ et $\sum_{i=1}^{n} m_i = k$.

Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ suit une distribution multinomiale avec les paramètres $\Phi_{TC} = (k, p_1, ..., p_n)$, où $p_i$ est la probabilité associée à chaque valeur $m$ :

$$p(m_1 = p_1, ..., m_n = p_n) = \frac{k!}{m_1!m_2!...m_n!} p_1^{m_1} p_2^{m_2} ... p_n^{m_n} \quad [2]$$

Soit $D$ un ensemble de $m$ documents $(d_1, ..., d_m)$ et $g$ groupes $(G_1, ..., G_g)$ dont chacun a une densité Gaussienne avec une moyenne $\mu_l, \forall l \in \{1, ..., g\}$ et une matrice de covariance $\sum_l$.

Soit $\pi_1, ..., \pi_g$ les proportions des différents groupes, on note par $\theta_k = (\mu_k, \sum_k)$ le paramètre de chaque Gaussienne, et $\Phi_{LLC} = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ le paramètre

global du mélange. Alors, la densité de probabilité de $D$ conditionnellement à la classe $c_i, \forall i \in \{1, ..., k\}$ est définie par :

$$P(d, \Phi_{LLC}) = \sum_{l=1}^{g} \pi_l p(d, \theta_l) \quad [3]$$

où $p(d, \theta_l)$ est la Gaussienne multivariée définie par le paramètre $\theta_l$.

On note par $\Phi$ le paramètre global de ce modèle :

$$\Phi = (\Phi_{LLC}, \Phi_{TC}) = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g, k, p_1, ..., p_n) \quad [4]$$

L'équation (1) peut être réécrite :

$$P(C, TC, LLC) = P(C) f(\pi_1, ..., \pi_g, \theta_1, ..., \theta_g, k, p_1, ..., p_n) \quad [5]$$

Les paramètres du modèle peuvent être appris d'un ensemble de documents d'apprentissage pour estimer la probabilité jointe de chaque document et chaque classe. Nous devons maximiser la log-vraisemblance $L_D$ de $D$ :

$$\begin{aligned} L_D &= log(P(C, TC, LLC) \\ &= \sum logP(C) + \sum logf(\pi_1, ..., \pi_g, \theta_1, ..., \theta_g) \\ &+ \sum logf(k, p_1, ..., p_n) \end{aligned} \quad [6]$$

Pour un document partiellement annoté, représenté par ses caractéristiques visuelles $LLC_1, LLC_2, ..., LLC_M$ et ses mots-clés existants $Kw_1, Kw_2, ..., Kw_n$, nous pouvons utiliser l'inférence bayésienne pour étendre l'annotation de ce document avec d'autres mots-clés. Nous pouvons calculer la probabilité *a posteriori*

$$P(Kw_i | LLC_1, LLC_2, ..., LLC_M, Kw_1, Kw_2, ..., Kw_n) \; \forall i \in \{1, ..., N\} \quad [7]$$

où $N$ est la taille du vocabulaire utilisé. Le mot-clé ayant la probabilité maximale sera retenu comme une nouvelle annotation du document. Nous pouvons également calculer la probabilité *a posteriori*

$$P(C_i | LLC_1, LLC_2, ..., LLC_M, Kw_1, Kw_2, ..., Kw_n) \quad [8]$$

dans le but d'identifier le type du document. Le document requête est affecté à la classe $C_i$ maximisant cette probabilité.

Le modèle proposé nécessite une étape préliminaire d'apprentissage où les documents sont annotés manuellement. Cette annotation pourrait être très coûteuse si l'ensemble d'apprentissage est annoté manuellement. Pour réduire ce coût et améliorer la

qualité de l'annotation obtenue, le retour de l'utilisateur est intégré dans notre modèle durant l'étape d'apprentissage. Dans un premier temps, l'annotation des documents d'apprentissage est automatiquement complétée par le modèle. Ensuite, l'utilisateur valide les nouvelles étiquettes des mots-clés correctes et rectifie certaines fausses étiquettes. Les documents avec leurs nouvelles étiquettes, après l'intervention de l'utilisateur, sont utilisés pour réapprendre le modèle. Ce processus peut être répété plusieurs fois. Autrement dit, nous avons fait une sorte "d'apprentissage dans l'apprentissage" et l'effort d'annotation est implicitement réduit puisque seulement quelques fausses étiquettes sont réajustées manuellement et itérativement (Voir Tableau 4 dans section 3.3).

## 3. Expérimentation

Dans cette section, nous présentons d'abord la base de documents utilisée dans nos expériences, et les différents critères d'évaluation des performances. Ensuite, nous présentons les résultats expérimentaux obtenus avec notre modèle.

### 3.1. *Base de documents et critères d'évaluation*

Nous avons effectué nos expériences sur des documents collectés à partir d'Internet. La base des documents est divisée en 100 documents pour l'apprentissage et 50 documents pour les tests. Les documents sont regroupés en 10 catégories (carte vitale, carte d'identité, passeport, fiche de salaire, facture de téléphone, facture d'électricité, permis de conduire, *curriculum vitæ*, articles scientifiques et articles de presse). Les documents d'appentissage et de test ont été annotés manuellement par 1 à 5 mots-clés en utilisant un vocabulaire de 46 mots.

Pour évaluer notre modèle d'annotation de documents, nous utilisons les quatre mesures d'évaluation standards utilisées dans l'annotation des images. Nous annotons automatiquement par notre modèle chaque document dans la base de test par 5 mots et nous calculons le rappel, la précision, $F1$ et $N+$. Supposons qu'une étiquette soit présente $m_1$ fois dans les documents de la vérité terrain, et apparaisse dans $m_2$ documents lors des tests à partir desquels $m_3$ prédictions sont correctes. La précision ($P$) est le rapport entre les documents correctement annotés par un mot-clé et tous les documents annotés par ce mot-clé par le modèle : $P = m_3/m_2$. Le rappel ($R$) est le rapport entre les documents correctement annotés par un mot-clé et tous les documents annotés par ce mot-clé dans les documents de vérité terrain : $R = m_3/m_1$. $N+$ est le nombre de mots qui sont correctement affectés à au moins un document de test (nombre de mots avec rappel strictement positif). La mesure $F_1$ est une moyenne harmonique entre le rappel et la précision : $F_1 = 2(PR)/(P + R)$.
De même, nous évaluons la classification avec les trois critères : rappel, précision et $F1$.

$$Rappel_i = \frac{nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

$$Précision_i = \frac{nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

$$Rappel = \sum_{i=1}^{C} \frac{Rappel_i}{C} \ ; \ Précision = \sum_{i=1}^{C} \frac{Précision_i}{C}$$

où $C$ est le nombre de classes.

### 3.2. *Caractéristiques visuelles*

Nous avons utilisé les deux descripteurs LBP (Ojala *et al.*, 1996) et l'histogramme des longueurs des séquences (Gordo *et al.*, 2013). Le descripteur de texture LBP compare le niveau de luminance d'un pixel avec les niveaux de ses voisins. Grâce à son pouvoir discriminant et la simplicité de calcul, LBP est devenu populaire dans diverses applications. L'histogramme des longueurs des séquences est un descripteur visuel où une séquence est une série de pixels de la même valeur. Cet histogramme est utilisé pour l'analyse et la classification des documents et il est rapide à calculer.

### 3.3. *Résultats*

Le Tableau 1 illustre l'extension des annotations de certains documents. Dans la première, la quatrième et la septième lignes, nous trouvons les documents à annoter. Chaque document est un exemple d'une catégorie. Dans la deuxième, la cinquième et la huitième ligne, les étiquettes de la vérité terrain sont données. Dans les autres lignes, on trouve les résultats de notre extension d'annotation (mots-clés en gras). Par exemple, le deuxième document (*curriculum vitæ*) a été annoté par deux mots-clés au début, trois nouveaux mots clés "compétences", "e-mail" et "adresse" sont automatiquement ajoutés après l'extension d'annotation.

Les tableaux 2 et 3 représentent respectivement les résultats de la classification et de l'annotation des documents de test suivant le taux d'annotation manuelle effectuée par l'utilisateur au début. Dans la première colonne de chaque tableau, nous présentons le nombre d'annotations dans les documents d'apprentissage. Dans les autres colonnes, nous présentons les performances de l'annotation et de la classification des documents de test. A partir de ces tableaux, nous pouvons remarquer que les performances sont améliorées lorsque le nombre d'annotations manuelles augmente. L'augmentation des performances est d'environ 50%, passant de 2 mots-clés par document à 5. Par exemple, pour l'annotation, nous passons d'une précision de $0,20$ et un rappel

**Tableau 1.** *Exemples d'annotation de documents*

| | | |
|---|---|---|
|  |  |  |
| net à payer | formation, études | numéro, consommation, nom |
| net à payer, **salaire**, **entreprise, date, numéro** | formation, études, **compétences, e-mail, adresse** | numéro, consommation, **adresse, net à payer, date** |
|  |  |  |
| nom, adresse, catégorie | journal, article | nom, numéro, sexe, taille |
| nom, adresse, catégorie, **signature, photo** | journal, article, **auteur**, **paragraphe, publicité** | nom, numéro, sexe, taille, **signature** |
|  |  |  |
| titre, nom, résumé | nom, opérateur | nationalité |
| titre, résumé, **introduction, travail, expérimentation** | nom, opérateur, **e-mail, consommation, adresse** | nationalité **numéro, sexe, taille préfecture** |

de $0,31$ (taux d'annotation manuelle de 2 mots-clés par document), à une précision de $0,35$ et un rappel de $0,5$ avec 5 mots-clés.

L'annotation manuelle est une tâche fastidieuse pour l'être humain. Pour réduire l'effort humain, nous avons intégré, tel que mentionné dans la section précédente, l'utilisateur dans le processus. Nous avons fait une sorte "d'apprentissage dans l'apprentissage". Plus précisément, nous menons l'expérience suivante dont les résultats

**Tableau 2.** *Performances de la classification*

| annotation manuelle (mots-clés) | P | R | F1 |
|---|---|---|---|
| 200 (2 par document) | 0.74 | 0.70 | 0.719 |
| 300 (3 par document) | 0.78 | 0.72 | 0.748 |
| 500 (5 par document) | 0.92 | 0.91 | 0.915 |

**Tableau 3.** *Performances de l'annotation*

| annotation manuelle (mots-clés) | P | R | F1 | N+ |
|---|---|---|---|---|
| 200 (2 par document) | 0.20 | 0.31 | 0.243 | 23 |
| 300 (3 par document) | 0.28 | 0.43 | 0.339 | 28 |
| 500 (5 par document) | 0.35 | 0.59 | 0.439 | 35 |

sont présentés dans le Tableau 4. Nous annotons d'abord manuellement chaque document d'apprentissage par 2 mots-clés. Nous souhaitons annoter par 5 mots-clés chaque document et, pour cela, au lieu de faire toute l'annotation manuellement, nous proposons de combiner une extension d'annotation itérative et un retour utilisateur. Ainsi, un troisième mot-clé est automatiquement étendu aux 100 documents d'apprentissage. Sur ces 100 mots-clés ajoutés, 19 annotations sont correctes, mais les 81 autres ont besoin de l'intervention de l'utilisateur. Ce processus est répété deux fois pour ajouter le quatrième et le cinquième mot-clé. Le nombre de corrections de l'utilisateur est de 72 et 63 respectivement. Ainsi, à partir de 200 mots-clés et 216 $(81+72+63)$ corrections de l'utilisateur, nous obtenons 500 mots-clés. Nous gagnons $16,8\%$ $(84/500)$ d'effort manuel.

**Tableau 4.** *Effort de l'utilisateur en tenant compte de ses retours*

| annotation manuelle (mots-clés) | corrections de l'utilisateur (mots-clés) | annotations correctes (mots-clés) |
|---|---|---|
| 200 | 81 | 300 |
| 281 | 72 | 400 |
| 353 | 63 | 500 |

## 4. Conclusion et perspectives

Nous avons proposé un modèle graphique probabiliste pour l'extension d'annotations et la classification de documents. Ce modèle est un mélange de distributions multinomiales et de mélange de Gaussiennes. Pour réduire le coût de l'annotation manuelle, nous améliorons itérativement l'apprentissage en ajoutant l'utilisateur dans le processus. Plus précisément, nous intégrons les évaluations de l'utilisateur pour apprendre le modèle. Les résultats expérimentaux montrent que notre modèle est efficace. Nos futurs travaux seront consacrés à utiliser les hiérarchies sémantiques pour enrichir l'annotation des documents.


## Remerciements

Ce travail est réalisé dans le cadre d'un contrat CIFRE avec la société Xilopix d'Épinal.


## 5. Bibliographie


Barrat S., Tabbone S., « Modeling, classifying and annotating weakly annotated images using bayesian network », *Journal of Visual Communication and Image Representation*, vol. 21, n° 4, p. 355-363, 2010.

Bay H., Ess A., Tuytelaars T., Van Gool L., « Speeded-Up Robust Features (SURF) », *Comput. Vis. Image Underst.*, vol. 110, n° 3, p. 346-359, June, 2008.

Bouzaieni A., Barrat S., Tabbone S., « Automatic annotation extension and classification of documents using a probabilistic graphical model », *13th International Conference on Document Analysis and Recognition,* p. 316-320, 2015a.

Bouzaieni A., Tabbone S., Barrat S., « Automatic Images Annotation Extension Using a Probabilistic Graphical Model », *Computer Analysis of Images and Patterns*, p. 579-590, 2015b.

Chakravarthy A., Ciravegna F., Lanfranchi V., « Cross-media document annotation and enrichment », *1st Semantic Web Authoring and Annotation Workshop (SAAW2006)*, 2006.

Chen N., Blostein D., « A Survey of Document Image Classification : Problem Statement, Classifier Architecture and Performance Evaluation », *International Journal of Document Analysis and Recognition*, vol. 10, n° 1, p. 1-16, 2007.

Chen N., Shatkay H., Blostein D., « Exploring a new space of features for document classification : figure clustering », *Proceedings of the conference of the Center for Advanced Studies on Collaborative research*, p. 369-372, 2006.

Coüasnon B., Camillerapp J., Leplumey I., « Access by content to handwritten archive documents : generic document recognition method and platform for annotations », *International Journal of Document Analysis and Recognition*, vol. 9, p. 223-242, 2007.

Dill S., Eiron N., Gibson D., Gruhl D., Guha R., Jhingran A., Kanungo T., Rajagopalan S., Tomkins A., Tomlin J. A. *et al.*, « SemTag and Seeker : Bootstrapping the semantic web via automated semantic annotation », *Proceedings of the 12th international conference on World Wide Web*, ACM, p. 178-186, 2003.

Duthil B., Coustaty M., Courboulay V., Ogier J. M., « Annotation sémantique de documents administratifs », *EGC*, p. 47-52, 2014.

Gordo A., Perronnin F., Valveny E., « Large-scale document image retrieval and classification with runlength histograms and binary embeddings », *Pattern Recognition*, vol. 46, n° 7, p. 1898-1905, 2013.

Kahan J., Koivunen M.-R., Prud'Hommeaux E., Swick R. R., « Annotea : an open RDF infrastructure for shared Web annotations », *Computer Networks*, vol. 39, n° 5, p. 589-608, 2002.

Kalyanpur A., Hendler J., Parsia B., Golbeck J., « SMORE-semantic markup, ontology, and RDF editor », *Defense Technical Information Center*, 2006.

Kumar J., Ye P., Doermann D., « Structural Similarity for Document Image Classification and Retrieval », *Pattern Recognition Letters,* p. 119-126, November, 2013.

Li J., Mouchère H., Viard-Gaudin C., « An annotation assistance system using an unsupervised codebook composed of handwritten graphical multi-stroke symbols », *Pattern Recognition Letters*, vol. 35, p. 46-57, 2014.

Lowe D. G., « Distinctive Image Features from Scale-Invariant Keypoints », *International Journal of Computer Vision*, vol. 60, n° 2, p. 91-110, November, 2004.

Ojala T., Pietikinen M., Harwood D., « A comparative study of texture measures a with classification based on feature distributions », *Pattern Recognition*, n° 29, p. 51-59, 1996.

Rusinõl M., Karatzas D., Bagdanov A. D., Lladós J., « Multipage document retrieval by textual and visual representations », *International Conference on Pattern Recognition*, p. 521-524, 2012.

Shah N., Mahajan S., « Document Clustering : A Detailed Review », *International Journal of Applied Information Systems*, vol. 4, n° 5, p. 30-38, October, 2012.

Sharma S., Gupta V., « Recent Developments in Text Clustering Techniques », *International Journal of Computer Applications*, vol. 37, n° 6, p. 14-19, January, 2012.