Un modèle de recherche de fichiers personnels par contexte dans les systèmes d'étiquetage

Ngo Ba Hung[1,2], Frédérique Silber-Chaussumier[1], Christian Bac[1]

Institut National des Télécommunications[1]
9 rue de Charles Fourier
91011 Evry cedex
France
{hung.ngo_ba, frederique.silber-chaussumier, christian.bac}@it-sudparis.eu

Université de Cantho[2]
1 rue de LyTuTrong,
NinhKieu, Cantho
Vietnam

RÉSUMÉ. *Depuis peu, les étiquettes sont utilisées largement pour identifier des contenus aussi bien sur le bureau informatique des utilisateurs que sur les sites coopératifs du Web dit 2.0. Notre recherche se focalise sur l'organisation assistée des étiquettes personnelles afin d'améliorer la pertinence des recherches de fichiers personnels associés à des étiquettes. Notre proposition utilise la notion de contexte comme point central. Un contexte est constitué à partir d'un ensemble d'étiquettes affectées par un utilisateur à un fichier. Nous proposons une infrastructure qui permet à un utilisateur de naviguer à travers les contextes pour retrouver ses fichiers.*

ABSTRACT. *Recently, tagging systems are widely used on the Internet. On desktops, tags are also supported by some semantic file systems and desktop search tools. In this paper, we focus on personal tag organization to enhance personal file retrieval. Our approach is based on the notion of context. A context is a set of tags assigned to a file by a user. Based on tag popularity and relationships between tags, our proposed algorithm creates a hierarchy of contexts on which a user can navigate to retrieve files in an effective manner.*

MOTS-CLÉS : *étiquette, système d'étiquetage, recherche de fichiers personnels, système de gestion de fichiers personnels, contexte, recherche d'information.*

KEYWORDS: *tag, tagging system, personal file retrieval, personal information management, context, information retrieval.*

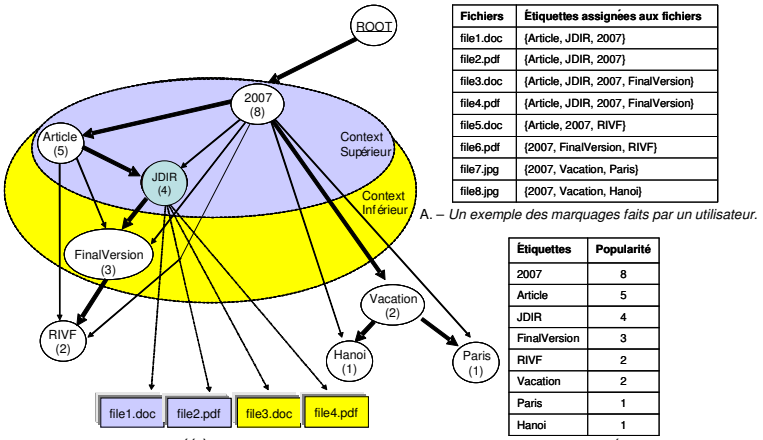
1. Introduction

Récemment, les systèmes d'étiquetage tel que (Delicious) sont largement utilisés sur Internet. Ces systèmes d'étiquetage permettent aux utilisateurs d'ajouter des mots clés (ou étiquettes) aux ressources de l'Internet pour les rechercher plus tard. Sur le bureau informatique, les étiquettes sont également utilisées par certains systèmes de fichiers sémantiques et des outils de recherche de fichiers. Les étiquettes, permettent de décrire de façon souple les opinions et les intérêts des utilisateurs dans un fichier (ou une ressource). Par conséquent, les fichiers intéressants pour un utilisateur - fichiers personnels - sont classés par étiquette et donc chaque utilisateur dispose d'un vocabulaire personnel matérialisé par les étiquettes. Les utilisateurs peuvent ensuite rechercher leurs fichiers en utilisant des expressions logiques qui portent sur ces étiquettes : *l'interrogation*. Par défaut, les systèmes d'étiquetage sont plus adaptés à la recherche de fichiers en utilisant *l'interrogation* que la *navigation*. Cependant, les expériences dans le domaine de gestion des informations personnelles (Barreau et Nardi, 1995), et (Khoo et al., 2007) montrent que la plupart des utilisateurs préfèrent la navigation que l'interrogation pour rechercher leurs fichiers dans un ordinateur personnel. C'est la raison pour laquelle, récemment, les systèmes d'étiquetage sur le bureau informatique se concentrent sur l'organisation des étiquettes dans des structures pour permettre aux utilisateurs de naviguer afin de trouver leurs fichiers. Notre proposition vise à améliorer la *recherche de fichiers personnels* dans les systèmes d'étiquetage en introduisant une notion que nous appelons *contexte*. Un contexte dans notre approche est un ensemble d'étiquettes assignées à un fichier (ou une ressource) par un utilisateur. Basé sur la popularité des étiquettes et les relations entre elles, nous proposons un algorithme qui crée une hiérarchie de contextes dans laquelle un utilisateur peut naviguer pour rechercher des fichiers d'une manière efficace. Dans cet article, nous présentons d'abord les techniques d'organisation des étiquettes dans la section 2; introduisons la notion du contexte basée sur des étiquettes et comment améliorer les systèmes d'étiquetage par la recherche de fichiers basée sur le contexte dans la section 3. Nous proposons un algorithme pour la création d'un *Graphe Acyclique Orienté des Étiquettes* (DAGoT - pour *Directed Acyclic Graph of Tags* en anglais), basé sur la popularité des étiquettes et les relations entre elles dans la section 4. Ce DAGoT est utilisé pour organiser des contextes dans une structure hiérarchique, afin que nous puissions améliorer la recherche de fichiers personnels dans les systèmes d'étiquetage avec le contexte. Une mise en œuvre et les résultats expérimentaux utilisant des données réelles sont présentés dans la section 5. Nos conclusions et perspectives sont dans la dernière section.

2. Techniques d'organisation des étiquettes

(Delicious) est un serveur très connu où les utilisateurs peuvent utiliser leurs propres étiquettes pour organiser en ligne des favoris et les rechercher plus tard.

Sous Delicious, deux ou plusieurs étiquettes associées à une même ressource sont considérées comme *relatives*. Le nombre de ressources associées à une étiquette par un utilisateur est appelé *la popularité de l'étiquette*. Quand une étiquette est choisie, une liste de ressources marquées avec cette étiquette et une liste d'étiquettes relatives à cette étiquette sont retournées comme résultat de recherche. Les étiquettes relatives permettent une naviguation pour revisiter des ressources intéressantes. Toutefois, lorsque le nombre de favoris et le nombre d'étiquettes augmentent, l'analyse du résultat de recherche pour un favori ou le choix pertinent d'une étiquette relative devient une tâche difficile pour un utilisateur. Sur le bureau informatique, des étiquettes sont également utilisées pour la recherche de fichiers personnels. Les utilisateurs de Spotlight (Apple Computer, 2005) peuvent assigner une étiquette à un ensemble de fichiers qui sont liés de manière à travers elle. Cette étiquette permet ensuite une recherche simple. Dans le domaine des systèmes de fichiers, LFS (Padioleau, 2005) permet aux utilisateurs d'associer des fichiers avec des étiquettes représentant les propriétés des fichiers. LFS utilise des axiomes entre étiquettes relatives comme des relations parent-enfant. Les utilisateurs peuvent créer manuellement des axiomes entre les étiquettes. À partir de ces axiomes, une taxonomie d'étiquettes est créée. Les utilisateurs peuvent naviguer sur la taxonomie pour rechercher des fichiers comme ils le font avec des répertoires traditionnels. TagFS (Bloehdorn et coll., 2006) organise les étiquettes en utilisant la notion d'étiquettes relatives comme le fait Delicious. Par conséquent, TagFS a le même inconvénient lorsque le nombre de fichiers et d'étiquettes augmente.



Fichiers	Étiquettes assignées aux fichiers
file1.doc	{Article, JDIR, 2007}
file2.pdf	{Article, JDIR, 2007}
file3.doc	{Article, JDIR, 2007, FinalVersion}
file4.pdf	{Article, JDIR, 2007, FinalVersion}
file5.doc	{Article, 2007, RIVF}
file6.pdf	{2007, FinalVersion, RIVF}
file7.jpg	{2007, Vacation, Paris}
file8.jpg	{2007, Vacation, Hanoi}

A. – *Un exemple des marquages faits par un utilisateur.*

Étiquettes	Popularité
2007	8
Article	5
JDIR	4
FinalVersion	3
RIVF	2
Vacation	2
Paris	1
Hanoi	1

B. – *Popupularité des étiquettes.*

C. – *Le DAGoT crée à partir des étiquettes dans le tabeau A.*

Figure 1. DAGoT créé à partir des marquages faits par un utilisateur.

3. Définition de Contexte basé sur des étiquettes

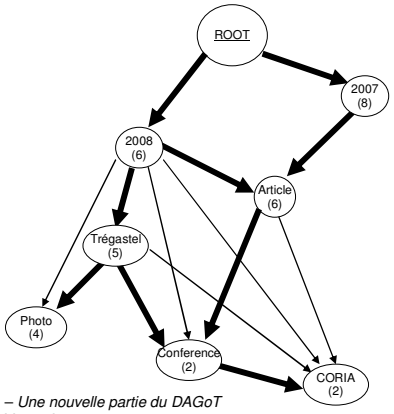
Un utilisateur d'un système d'étiquetage réalise *un marquage* en assignant un ensemble d'étiquettes à un fichier. Chaque étiquette représente un concept ou un objet concernant le fichier. De plus l'ensemble des étiquettes assignées au fichier représente un sujet ou un thème concernant le propriétaire du fichier. Par exemple, un utilisateur peut assigner l'ensemble des étiquettes {*vacances, Paris, 2007*} au fichier *myphoto.jpg* pour rappeler que la photo a été prise lors des vacances de l'été 2007, à Paris. Nous appelons cet ensemble d'étiquettes assignées à une ressource par un utilisateur *un contexte basé sur des étiquettes* (ou le contexte). Le sens d'un contexte est agrégé à partir de ses éléments. Un contexte est plus significatif qu'une étiquette. Par exemple, le contexte {*vacances, Paris, 2007*} est plus significatif que l'étiquette *2007*. En fait, lorsqu'il attribue un ensemble d'étiquettes à un fichier, un utilisateur souhaite classifier le fichier par le contexte représenté par l'ensemble de ces étiquettes. Les systèmes d'étiquetage devraient donc permettre une recherche plus pertinente s'ils autorisaient la recherche de fichiers par contexte. La figure 1.A est un exemple de marquages faits par un utilisateur. Si l'utilisateur fait une recherche de fichiers en fonction de l'étiquette *Article*, les cinq fichiers, *file1* à *file5*, sont retournés. Ces fichiers appartiennent à trois contextes {*Article, JDIR, 2007*}, {*Article, JDIR, 2007, FinalVersion*} et {*Article, 2007, RIVF*}. Nous trouvons qu'une étiquette participe habituellement dans des nombreux contextes. La figure 1.C montre que l'étiquette *JDIR* participe à deux contextes {*Article, JDIR, 2007*} et {*Article, JDIR, 2007, FinalVersion*}. Le premier contexte est plus général que le second. Dans la section suivante, nous proposons une méthode pour organiser des fichiers par leurs contextes et classifier des contextes dans une structure hiérarchique: des contextes généraux aux contextes spécifiques

4. Recherche de fichiers personnels par contexte

Nous proposons une méthode pour construire, à partir des étiquettes personnelles, un graphe acyclique orienté en fonction de la popularité des étiquettes et de leurs relations entre elles. Ce Graphe Acyclique Orienté des Étiquettes (DAGoT) est utilisé pour organiser automatiquement les fichiers dans des contextes appropriés, pour identifier le contexte le plus général contenant une étiquette, et pour permettre à l'utilisateur de naviguer d'un contexte à un autre pour rechercher des fichiers d'une manière efficace.

Un DAGoT dispose de trois types de nœud : étiquette, ressource et racine. *Un nœud étiquette* représente une étiquette créée par un utilisateur. Il a un label et une popularité. Un nœud étiquette peut avoir de nombreux nœuds parents et de nombreux nœuds enfants. *Un nœud ressource* représente un fichier annoté par un utilisateur. Il a une localisation, telle qu'une URL, d'où le fichier peut être accédé. Un nœud ressource est un nœud feuille. Il a un ou plusieurs nœuds étiquettes comme parents. *Un nœud racine* est la racine de l'arbre et représente la plus populaire des

nœuds étiquettes. Il existe trois types d'arc : arc relatif, arc parent et arc ressource. Deux étiquettes affectées au même fichier sont dites relatives. *Un arc relatif* relie deux nœuds relatifs : du plus populaire (*nœud supérieur*) au moins populaire (*nœud inférieur*). Si deux nœuds étiquettes ont la même popularité, celui qui a le label le plus petit est le nœud supérieur. Une étiquette sans nœud supérieur s'associe au *nœud racine* comme nœud supérieur. Ainsi, un nœud étiquette a toujours un nœud supérieur. Pour un nœud étiquette, ses nœuds supérieurs sont distribués dans les différents *sous-graphes entiers interrelation*. Les nœuds dans chaque sous-graphe sont relatifs les uns aux autres. Le moins populaire des nœuds dans chaque sous-graphe devient un nœud parent du nœud. Et les arcs relatifs reliant un nœud et ses nœuds parents deviennent les *arcs parents*. Quand une étiquette est assignée à un fichier, *un arc ressource* est créé à partir du nœud étiquette vers le nœud ressource. La figure 1.C représente le DAGoT créé à partir des marquages dans la figure 1.A. La popularité de ces étiquettes est représentée dans la figure 1.B. Les flèches représentées par des traits fins, épais et pointillés représentent respectivement l'arc relatif, l'arc parent et l'arc de ressource. Pour être concis, nous ne montrons que les fichiers associés à l'étiquette *JDIR*. Le DAGoT montre que l'étiquette *JDIR* accepte l'étiquette *Article* comme parent et *FinalVersion* comme enfant. *JDIR* participe à deux contextes {*2007, l'article, JDIR*} et {*2007, l'article, JDIR, FinalVersion*}. Le premier est le contexte le plus populaire de *JDIR* et contient les étiquettes supérieures de *JDIR*. C'est ce contexte que le système donne comme résultat quand un utilisateur fait une recherche avec l'étiquette *JDIR*. A partir de *JDIR*, l'utilisateur peut raffiner sa requête en se déplaçant vers son enfant *FinalVersion*. Les arcs parents maintiennent une relation hiérarchique entre les contextes. Ils sont utilisés comme des guides pour l'utilisateur pour naviguer d'un contexte à l'autre.



C. – *Une nouvelle partie du DAGoT créée après l'ajouter des marquages dans la table A.*

Fichiers	Étiquettes assignées aux fichiers
file9.pdf	{Conference, CORIA, 2008, Trégastel}
file10.doc	{Conference, CORIA, 2008, Article}
file11.jpg	{Trégastel, 2008, Photo}
file12.jpg	{Trégastel, 2008, Photo}
file13.jpg	{Trégastel, 2008, Photo}
file14.jpg	{Trégastel, 2008, Photo}

A. – *Des marquages ajoutés par l'utilisateur.*

Étiquettes	Popularité
2008	6
Article	6
Trégastel	5
Photo	4
Conference	2
CORIA	2

B. – *Popupularité des étiquettes.*

Figure 2. *Exemple où l'étiquette Conference n'a plus qu'un seul parent.*

Nous supposons que l'utilisateur continue d'ajouter de nouveaux marquages comme dans la figure 2.A. La nouvelle popularité des étiquettes est décrite dans la figure 2.B. La popularité de l'étiquette *Article* dans la figure 2 augmente à 6. Dès que le nœud *Article* et le nœud *2008* ont la même popularité, le nœud *2008* est considéré comme le nœud supérieur du nœud *Article*, parce que le label du nœud *2008* est le plus petit que celui du nœud *Article*. C'est la même situation pour deux étiquettes *Conference* et *CORIA*. La comparaison des labels est non sensible à la casse. Le nœud *Conference* est un exemple intéressant de nœud étiquette ayant plusieurs parents. Le nœud *Conference* a trois nœuds supérieurs *2008*, *Article*, et *Trégastel*. Ils sont distribués dans deux sous-graphes entiers interrelation: {*2008, Article*} et {*2008, Trégastel*}. Dans le premier sous-graphe, *Article* est le moins populaire nœud. Dans le deuxième, c'est *Trégastel*. C'est pourquoi *Article* et *Trégastel* sont acceptés comme deux parents du nœud *Conference*. De la même façon, *Article* accepte *2007* et *2008* comme ses deux parents.

[1] $Res(t) \leftarrow \{r \in R \mid (r,t) \in P\}$: *Les arcs ressources partant d'une étiquette.*
[2] $Tag(r) \leftarrow \{t \in T \mid (r,t) \in P\}$: *Les arcs ressources arrivant à une ressource.*
[3] $Pop(t) \leftarrow Card(Res(t))$: *La popularité d'une étiquette.*
[4] $Rel(t1,t2) \leftarrow \exists r \in R \mid (r,t1) \in P \ \& \ (r,t2) \in P$: *Vérifier s'il existe un arc relatif entre deux étiquettes.*
[5] $Upper(t') \leftarrow \{t \mid Rel(t',t) \ \& \ Pop(t) > Pop(t')\}$: *Les arcs relatifs arrivant à l'étiquette t'.*
[6] $Upper(t') \leftarrow \{t \mid Rel(t',t) \ \& \ Pop(t) = Pop(t') \ \& \ Label(t) < Label(t')\}$: *Si deux étiquettes relatives ont la même popularité, l'arc relatif part de l'étiquette ayant le label le plus petit*
[7] $Parent(t) \leftarrow \{p \mid p \in Upper(t) \ \& \ !\exists p' \ \& \ p' \in Upper(t) \ \& \ p \in Upper(p')$: *Les noeuds parents d'un noeud étiquette.*
[8] $Children(t) \leftarrow \{ c \in T \mid t \in Parent(c)\}$: *Les noeuds enfants d'un noeud étiquette.*
[9] $Rsat(t) \leftarrow \{r \in Res(t) \mid Tag(r) \subseteq (Upper(t) \cup \{t\}) \}$
[10] $Empty(t) \leftarrow Card(Rsat(t))=0 \ \& \ Card(Children(t))=1 \ \& \ Card(Parent(t))=1$
[11] $Cbfr(t) \leftarrow [Rsat(t), Psat(t), Csat(t)] \mid !Empty(t)$
[12] $Cbfr(t) \leftarrow Cbfr(c) \mid Empty(t) \ \& \ c \in Children(t)$
[13] $Psat(t) \leftarrow \{ p \in Parent(t) \mid !Empty(p) \}$
[14] $Psat(t) \leftarrow Psat(p) \mid p \in Parent(t) \ \& \ Empty(p)$
[15] $Csat(t) \leftarrow \{c \in Children(t) \mid !Empty(c)\}$
[16] $Csat(t) \leftarrow Csat(c) \mid c \in Children(t) \ \& \ Empty(c)$

Tableau 1. *Le modèle formel d'un DAGoT.*

Pour chaque utilisateur, un système d'étiquetage est formellement représenté comme un tuple *U: = (R, T, P)*, où *R* et *T* sont des ensembles finis qui représentent les ressources (ou fichiers) et les étiquettes gérées par un utilisateur. *P* représente les

marquages faits par l'utilisateur. Un marquage représente la relation entre une ressource et une étiquette, $P = RxT$. Le modèle formel pour un DAGoT est décrit dans le tableau 1. Étant donné une étiquette *t*, la recherche de fichiers par contexte *Cbfr(t)* contient trois types d'informations: un ensemble de fichiers *Rsat(t)* qui satisfont le contexte le plus populaire contenant *t*; une liste d'étiquettes parentes *Psat(f)* vers les contextes plus généraux, et une liste d'étiquettes enfants *Csat(f)* vers des contextes plus spécifiques. Dans l'exemple ci-dessus, nous avons *Cbfr(JDIR) = [{file1.doc, file2.pdf}, {Article}, {FinalVersion}]*. En fait, *Rsat* ne retourne pas toujours une valeur pour chaque étiquette. Il y a quelques étiquettes pour lesquelles *Rsat* est vide. Une étiquette *t* est *vide* si son *Rsat(t)* est vide et s'il n'a qu'un seul parent et un seul enfant. Nous proposons que le résultat d'une recherche d'une étiquette vide soit automatiquement remplacé par celui de son unique enfant. Dans le DAGoT de la figure 1, *Article* est une étiquette vide. Par conséquent *Cbfr(Article)* est automatiquement remplacé par *Cbfr(JDIR)*. En outre, les rôles parent et enfant d'une étiquette vide sont également remplacés par son parent et enfant.

Modèle Delicious		**Notre modèle basé sur DAGoT**	
Étiquettes par utilisateur	Moyen: 717	Contextes par utilisateur	Moyen: 145
	Intervalle: 2-4590		Intervalle: 2-663
Ressources par étiquette	Moyen: 4.6	Ressources par contexte	Moyen: 2.2
	Intervalle: 1-1426		Intervalle: 1-96
Étiquettes relatives par étiquette	Moyen: 16.5	Parents par étiquette	Moyen: 1.1
	Intervalle: 1-3715		Intervalle: 1-23
		Enfants par étiquette	Moyen: 2.7
			Intervalle: 1-113

Tableau 2. *Comparaison entre le modèle de Delicious et modèle DAGoT.*

5. Résultats expérimentaux

Premièrement, nous avons téléchargé les marquages de 46 personnes choisies aléatoirement depuis le site de (Delicious) pour calculer le nombre d'étiquettes et de ressources gérées par un utilisateur et le nombre de ressources et d'étiquettes relatives à une étiquette. Ensuite, nous avons fait des statistiques sur les 46 DAGoT créés dans le but de valider notre approche. Le tableau 2 compare les valeurs utiles pour la recherche de fichiers dans les deux modèles. Les valeurs moyennes des caractéristiques obtenues dans le modèle basé sur les contextes utilisant DAGoT sont toutes inférieures à celles du modèle de Delicious utilisant les étiquettes brutes. Ainsi, un utilisateur de Delicious dispose d'environ 717 étiquettes pour choisir et notre traitement nous permet d'obtenir seulement 145 contextes. Ce premier résultat

montre que le modèle DAGoT aide mieux les utilisateurs en réduisant l'espace de recherche. Les intervalles des valeurs de caractéristiques comparées dans notre modèle sont petits. Cela montre que le DAGoT dispose d'une structure plus équilibrée. Cela permet d'éviter les cas où il y a des milliers de ressources ou des centaines d'étiquettes relatives retournées pour une étiquette donnée.

6. Conclusion et perspectives

Nous avons proposé d'améliorer la recherche de fichiers personnels en introduisant une recherche par contexte. Nous supposons que chaque utilisateur dispose d'un vocabulaire personnel basé sur des étiquettes qui sont sémantiquement regroupées dans des contextes différents. L'ensemble des étiquettes associées à un fichier par un utilisateur crée un contexte. Nous avons proposé un algorithme pour créer un DAGoT basé sur la popularité et les relations entre étiquettes. Ce DAGoT est utilisé pour identifier automatiquement le contexte satisfait par une étiquette donnée. En utilisant un DAGoT, un utilisateur peut naviguer d'un contexte à l'autre pour rechercher les fichiers d'une manière efficace. À l'avenir, nous allons intégrer ce modèle au système de fichiers sémantique basée sur des ontologies (Ngo et al., 2007) et nous souhaitons proposer une méthode complète pour la recherche de fichiers dans laquelle nous prenons en compte à la fois les sémantiques associées au contenu et à l'environnement d'un fichier.

Bibliographie

Apple Computer, Inc: Tiger Developer Overview Series - Working with Spotlight, 2005, http://developer.apple.com/macosx/spotlight.html.

Barreau D., Nardi B., « Finding and reminding: file organization from the desktop ». ACM's *Special Interest Group in Computer-Human Interaction Bulletin*, 27(3), 1995, p. 39-43.

Bloehdorn S., Görlitz O., Schenk S., Völkel M., « TagFS --- Tag Semantics for Hierarchical File Systems ». *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, Graz, Austria, September 2006.

Delicious. http://del.icio.us/

Khoo C., Luyt B., Ee C., Osman J., Lim H.H., Yong S, « How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour », *Information Research*, 12(2), 2007, p. 293.

Ngo H.B., Bac C., Silber-Chaussumier F., « Toward ontology based semantic file systems », *Proceedings of the 5th International Conference on Research, Innovation & Vision for the Future*, 2007, Hanoi, Vietnam.

Padioleau Y., Logic File System, un système de fichier basé sur la logique. Thèse de doctorat, Université de Rennes 1, 2006.