Estimation de la pertinence multidimensionnelle en recherche d'information : évaluation de l'application d'un opérateur flou d'agrégation

Bilel Moulahi[*,**] **Lynda Tamine**[*] **Sadok Ben Yahia**[**,***]

[*] *Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France*
bilel.moulahi@irit.fr, lynda.tamine@irit.fr

[**]*Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH, 2092, Tunisie*
sadok.benyahia@fst.rnu.tn

[***]*Institut Mines-Télécom, Télécom SudParis, UMR CNRS Samovar, 91011 Evry Cedex, France*

RÉSUMÉ. *Nous proposons une nouvelle approche d'agrégation pour l'estimation de la pertinence multidimensionnelle. L'approche est basée sur un opérateur d'agrégation mathématique qui utilise une mesure floue permettant la quantification de l'importance estimée des critères ainsi que leur degré d'interaction ou d'interdépendance. Nous évaluons l'opérateur proposé dans le cadre de trois scénarios de recherche d'information, en l'occurrence une tâche de recherche de tweets, une tâche de recherche personnalisée dans les folksonomies et une tâche de recherche d'information contextuelle. Les résultats expérimentaux obtenus montrent l'impact de l'approche proposée sur les performances de recherche.*

ABSTRACT. *We propose a novel personalized aggregation approach to the multidimensional relevance aggregation. The approach is based on a mathematical aggregation operator relying on a fuzzy measure that allows quantifying the importance degree of each relevance dimension as well as the interaction existing between the criteria. Evaluation is carried out within three information retrieval settings referring to a tweet search task, a personalized information retrieval setting and a contextual suggestion task. Experimental results show the effectiveness of our approach on the search effectiveness.*

MOTS-CLÉS : *Personnalisation, préférences, pertinence, Choquet personnalisé, capacité.*

KEYWORDS: *Pesonalization, preferences, relevance, personalized Choquet, capacity.*

1. Introduction

De nombreux travaux en recherche d'information (RI) ont mis en exergue à la fois l'importance et la complexité du concept de *pertinence* (Borlund, 2003 ; Saracevic, 2007 ; Taylor *et al*., 2007). Son importance est liée au fait que la notion sous-jacente est le fondement des modèles d'ordonnancement de documents en réponse à une requête, qui est la finalité même d'un système de RI (Baeza-Yates, Ribeiro-Neto, 1999). Sa complexité est, quant à elle, subordonnée à deux propriétés. La première concerne la multiplicité de ses dimensions, vues comme des ensembles de critères, qui peuvent être de surcroît, interdépendantes ; même si de nombreux travaux du domaine se sont focalisés sur la dimension thématique seule, force est de constater que de nombreux autres travaux ont prouvé empiriquement l'impact conjoint de plusieurs dimensions sur l'estimation de la pertinence finale, comme la tâche et la situation de recherche (Borlund, 2003 ; Saracevic, 2007 ; Taylor *et al*., 2007). Considérons à titre d'exemple, une tâche de recherche de *tweets* ; des analyses expérimentales ont montré que la pertinence d'un *tweet* en réponse à une requête, est impactée principalement par la conjonction de trois dimensions qui sont le sujet et la fraîcheur du *tweet* et l'autorité du *tweeterer* qui l'a émis (Nagmoti *et al*., 2010). La seconde propriété concerne la subjectivité qui entoure ces dimensions ; en effet, la plupart d'entre elles ne sont pas basées sur des estimations objectives puisqu'elles sont fortement liées à la perception personnelle des utilisateurs impliqués dans la tâche de RI ; on cite à titre d'exemple les centres d'intérêt, l'expertise et les préférences des utilisateurs. La problématique scientifique est alors de définir des opérateurs capables d'agréger des scores de pertinence partiels (relatifs à chaque dimension) en tenant compte de leur interdépendance éventuelle. Cette problématique a été abordée dans diverses applications de RI comme la RI personnalisée (Sieg *et al*., 2007 ; Daoud *et al*., 2011 ; Costa Pereira *et al*., 2012), la RI mobile (Göker, Myrhaug, 2008 ; Bouidghaghen, Tamine, Boughanem, 2011), la RI sociale (Nagmoti *et al*., 2010) et la RI géographique (Mata, Claramunt, 2011). Cependant, ces travaux applicatifs ont généralement utilisé des opérateurs de calcul de moyenne pondérée ou de combinaison linéaire qui se basent sur l'hypothèse non réaliste d'additivité ou d'indépendance des dimensions. D'autres travaux fondamentaux récents, se sont intéressés en revanche à la définition d'opérateurs d'agrégation, indépendamment du cadre applicatif, qui permettent de traiter peu ou prou par le biais de l'interaction (Costa Pereira *et al*., 2012 ; Gerani *et al*., 2012 ; Eickhoff *et al*., 2013). Toutefois, ces opérateurs ne permettent pas : 1) de quantifier explicitement l'importance absolue des dimensions de pertinence compte tenu de la tâche de RI, 2) de tenir compte de la propriété de subjectivité qui peut se décliner à travers les différences entre les utilisateurs quant à l'importance accordée à chaque dimension de pertinence. Notre contribution, présentée dans ce papier, tente de répondre à ces objectifs. Plus précisément, nous proposons un opérateur flou d'agrégation basé sur l'intégrale de Choquet (Choquet, 1953 ; Grabisch, 1995), capable de pondérer et d'agréger les scores de différentes dimensions de pertinence pouvant être interdépendantes. Ces scores peuvent être, de surcroît, personnalisés de manière à considérer les préférences des utilisateurs.

La suite du papier est organisée comme suit : la section 2 présente un aperçu des travaux du domaine et situe notre contribution dans ce contexte. La section 3 détaille les principes de l'opérateur d'agrégation ainsi que l'algorithme d'apprentissage des mesures d'importance. Les sections 4 et 5 décrivent le cadre expérimental puis les résultats de l'application de l'approche proposée dans deux tâches TREC dédiées à la RI personnalisée en l'occurrence "TREC [1] *Contextual Suggestion*" (Dean-Hall *et al*., 2013) et la recherche de tweets dans le cadre de la tâche TREC Microblog (Soboroff *et al*., 2012). Nous présentons également les résultats dans le cadre d'une tâche de RI personnalisée dans les folksonomies (Vallet, Castells, 2012).

2. Synthèse des travaux

La littérature concernant le domaine de la RI a connu un très grand nombre de publications portant sur le concept de pertinence durant les deux dernières décennies. Dans cette section, nous présentons un aperçu des travaux qui se sont intéressés à ce concept, tant au niveau de la définition que sur le plan de son application dans divers cadres de RI. Ensuite, nous situons notre contribution dans ce cadre.

2.1. Pertinence multidimensionnelle

Le concept de pertinence est incontestablement au centre d'une activité de RI comme en témoignent les nombreux travaux qui en ont fait l'objet d'étude (Saracevic, 1976 ; Borlund, 2003 ; Saracevic, 2007). L'un des résultats phares qui ressort de ces travaux est que la pertinence est estimée en globalité selon un ensemble de dimensions qui s'apparentent à des familles de critères ; parmi ces différentes dimensions, on cite les plus reconnues dont : la pertinence thématique (contenu et méta-contenu), la pertinence situationnelle (temps et géolocalisation) et la pertinence cognitive (expertise, centres d'intérêts). Un autre résultat important est l'interdépendance de ces dimensions pour inférer la pertinence globale d'un document (Nagmoti *et al*., 2010 ; Saracevic, 2007). En clair, un utilisateur juge de la pertinence d'un document en tenant compte conjointement de l'ensemble des critères de pertinence ; à titre d'exemple, un document est d'autant plus pertinent du point de vue du contenu que l'expertise de l'utilisateur est en lien avec ce contenu. Historiquement, la dimension thématique est particulièrement considérée dans le domaine. La prise en compte de la propriété de multiplicité des dimensions de pertinence a particulièrement émergé dans des cadres applicatifs de la RI comme :

– la RI mobile (Göker, Myrhaug, 2008 ; Bouidghaghen, Tamine, Boughanem, 2011) : un document est d'autant plus pertinent pour une requête qu'il en est proche thématiquement et qu'il comporte des liens vers des lieux géographiquement proches de l'utilisateur qui est en situation de mobilité;

1. http://trec.nist.gov

– la RI sociale (Nagmoti *et al*., 2010) : un document (ou ressource sociale) est d'autant plus pertinent pour une requête qu'il en est proche thématiquement, qu'il émane d'un acteur socialement important et qu'il est recommandé par un ami;

– la RI personnalisée (Sieg *et al*., 2007 ; Daoud *et al*., 2011 ; Costa Pereira *et al*., 2012) : un document est d'autant plus pertinent pour une requête qu'il en est proche thématiquement et qu'il est en adéquation avec les centres d'intérêts de l'utilisateur;

– RI géographique (Daoud, Huang, 2013) : un document est d'autant plus pertinent pour une requête qu'il en est proche thématiquement et qu'il comporte des liens vers des lieux géographiquement proches des lieux cités dans la requête;

2.2. Estimation de la pertinence multidimensionnelle

La plupart des travaux exploitant des critères des pertinences se basent sur des opérateurs classiques de produit, de moyenne pondérée et de combinaison linéaire. D'autres travaux (Palacio *et al*., 2010) exploitent des opérateurs de combinaison inspirés de la fusion des données. Cependant, ces opérateurs répondent à la problématique de l'agrégation en se basant sur l'hypothèse d'additivité ou d'indépendance des dimensions de pertinence. D'autres travaux récents ont particulièrement examiné le principe d'agrégation de dimensions interactives indépendamment du cadre applicatif (Costa Pereira *et al*., 2012 ; Gerani *et al*., 2012 ; Eickhoff *et al*., 2013). da Costa Pereira et *al.* (2011) ont proposé un opérateur d'agrégation multidimensionnelle mettant en jeu quatre critères des pertinence : contenu, couverture, adéquation et fiabilité en définissant deux opérateurs d'agrégation prioritaire en l'occurrence, "*And*" et "*Scoring*". Ces opérateurs modélisent un ordre de priorité entre les critères de pertinence sur la base d'un mode de calcul de poids associés qui favorise la satisfaction du critère d'ordre supérieur ; les travaux présentés dans (Bouidghaghen, Tamine, Pasi *et al*., 2011) ont montré l'efficacité de ces opérateurs dans un cadre de RI mobile. Gerani et *al.* (2012) ont proposé un opérateur qui ne nécessite pas la satisfaction de la condition de comparabilité des scores partiels de pertinence. Ils utilisent à cet effet un algorithme de transformation de scores basé sur l'algorithme *Alternating Conditional Expectation* et le modèle *BoxCox*. Plus récemment, Eickhoff et *al.* (2013) ont proposé une approche statistique basé sur la méthode *Copulas* qui traite spécifiquement la complexité des dépendances des critères de pertinence. Les auteurs ont montré que la méthode *Copulas* permet de modéliser des relations de dépendances complexes entre les différentes dimensions de pertinence. Leur approche a été évaluée dans trois tâches de RI à savoir, la recherche d'opinions dans les *blogs*, la RI personnalisée dans les folksnonomies et la recherche *Web* adaptée aux enfants.

Une autre direction de recherches qui a été largement exploitée dans la littérature est celle de l'apprentissage d'ordonnancement (Burges *et al*., 2005 ; Cao *et al*., 2007 ; Li, 2011). L'objectif consiste à optimiser une fonction d'ordonnancement automatiquement en se basant sur un ensemble d'apprentissage et suivant une mesure de pertinence bien définie (e.g., précision, NDCG). Cet ensemble inclut typiquement un ensemble de paires (documents, requêtes) avec une vérité de terrain qui leur est asso-

ciée. Chaque paire est représentée par un vecteur dans l'espace de termes et de documents qui les constituent. L'objectif est donc de combiner les scores afin d'apprendre la fonction optimale qui permet de donner le meilleur ordonnancement de documents. Cette méthodologie a été appliquée dans différent travaux en RI. Par exemple, dans les travaux de (Duan *et al*., 2010) et (Metzler, Cai, 2011), les auteurs ont proposé des algorithmes d'apprentissage d'ordonnancement pour la recherche de tweets afin de combiner plusieurs critères de pertinence. Toutefois, en dépit de leur large exploitation, ces méthodes ne permettent pas de donner une idée, pour les décideurs, sur la dépendance ou l'importance des critères utilisés (Eickhoff *et al*., 2013).

2.3. Aperçu de la contribution et positionnement

La cadre général de nos travaux concerne l'agrégation de dimensions de pertinence, qu'elles soient interdépendantes ou indépendantes que nous évaluons dans trois tâches de RI. Plus spécifiquement, nous présentons une approche d'agrégation (personnalisée) des scores de pertinence basée sur l'usage d'une mesure floue, appelée capacité, sous-jacente à l'opérateur de Choquet (Choquet, 1953). Cette mesure est à la base de la quantification de l'importance estimée de chaque dimension pour chaque utilisateur ainsi que leur degré d'interaction ou d'interdépendance ; elle est estimée selon un algorithme d'apprentissage, qui infère les mesures optimales en utilisant une vérité de terrain évaluable à l'aide de la métrique précision de la recherche.
Nous effectuons une évaluation approfondie sur des adaptations du modèle avec et sans personnalisation des critères. Dans un premier temps, nous faisons une évaluation de notre approche non personnalisée dans un cadre de recherche de tweets, sur une collection de test standard fournie par la tâche Microblog de TREC 2011 et 2012. Ensuite, nous évaluons l'opérateur d'agrégation personnalisé dans deux contextes de RI différents, dont l'un en nous basant sur un scénario de RI dans les folksonomies et l'autre en utilisant un contexte de RI contextuelle. Dans ces deux derniers scénarios, nous exploitons respectivement une collection de signets collectées à partir d'un système d'annotation sociale ainsi que la collection de test standard fournie par la tâche TREC *Contextual Suggestion* (Dean-Hall *et al*., 2013). Pour ces deux cadres de RI, nous montrons l'impact de la prise en compte des dépendances entre les critères de pertinence ainsi que l'impact de leur personnalisation sur les performances de recherche. Comparativement aux travaux antérieurs proches (Costa Pereira *et al*., 2012 ; Gerani *et al*., 2012 ; Eickhoff *et al*., 2013) ainsi qu'à nos précédentes contributions (Moulahi, Tamine, Yahia, 2014), le travail présenté dans ce papier s'en distingue selon les principaux points clés suivants :

1. une approche générale pour l'estimation de pertinence pouvant être appliquée indépendamment du cadre de RI;

2. une agrégation pondérée par les préférences des utilisateurs quant à chacune des dimensions agrégées, contrairement aux travaux de l'état de l'art présentés dans (Costa Pereira *et al*., 2012 ; Gerani *et al*., 2012 ; Eickhoff *et al*., 2013) ainsi que dans notre précédente contribution (Moulahi, Tamine, Yahia, 2014) ; ces travaux proposent de déployer des opérateurs produisant des scores de pertinence dépendant seulement

des dimensions de pertinence agrégées, indépendamment des utilisateurs;

3. un nouvel algorithme d'apprentissages des mesures d'importance des critères;

4. une évaluation expérimentale étendue tant dans l'objectif que dans la méthodologie, qui montre, comparativement à celle menée dans (Moulahi, Tamine, Yahia, 2014 ; Moulahi, Tamine, Ben Yahia, 2014) [2], à la fois l'intérêt de l'agrégation et de la personnalisation des préférences des utilisateurs sur les performances de recherche.

3. Agrégation personnalisée de la pertinence multidimensionnelle

3.1. Formalisation de l'opérateur d'agrégation

Nous introduisons le problème d'agrégation de pertinence multidimensionnelle comme étant un problème de prise de décision multicritères où les critères considérés sont les dimensions de pertinence. En effet, le défi majeur dans le problème d'agrégation est : (1) l'estimation de l'importance des critères : identifier les critères devant avoir un poids d'importance plus élevé que d'autres ; (2) l'agrégation : combiner efficacement les critères de pertinence en tenant compte des dépendances pouvant exister entre eux.
Soient $\mathcal{D}$ un ensemble de documents, $\mathcal{C}$ l'ensemble des critères de pertinence et q une requête donnée. La tâche de combinaison des critères notée $RSV^u_{c_i}(q,d_j)$, d'un document $d_j \in \mathcal{D}$, obtenu suivant chaque critère de pertinence $c_i \in \mathcal{C}$ en réponse à une requête d'un utilisateur u, est appelé *agrégation*. La fonction $\mathcal{F}$ qui calcule le score de pertinence personnalisé a la forme suivante :

$$\mathcal{F} : \begin{cases} \mathbb{R}^N \longrightarrow \mathbb{R} \\ (RSV^u_{c_1}(q,d_j) \times \ldots \times RSV^u_{c_N}(q,d_j)) \longrightarrow \mathcal{F}(RSV^u_{c_1}(q,d_j), \ldots, RSV^u_{c_N}(q,d_j)) \end{cases}$$

Dans ce qui suit, nous allons nous baser sur l'intégrale de Choquet comme un opérateur d'agrégation de pertinence multidimensionnelle. Cette fonction mathématique est construite à l'aide d'une mesure floue (ou *capacité*) μ, définie comme suit.

DÉFINITION 1. — *Soit $I_\mathcal{C}$ l'ensemble de tous les sous ensembles de critère de $\mathcal{C}$. Une mesure floue est une fonction monotone normalisée μ de $I_\mathcal{C}$ à $[0,1]$ tels que :*
$\forall I_{C_1}, I_{C_2} \in I_\mathcal{C}$, *si* $(I_{C_1} \subseteq I_{C_2})$ *alors* $\mu(I_{C_1}) \leq \mu(I_{C_2})$, *avec* $\mu(I_\varnothing) = 0$ *et* $\mu(I_\mathcal{C}) = 1$.

Pour simplifier la notation, $\mu(I_{C_i})$ sera dénotée par μ_{C_i}. La valeur de μ_{C_1} peut être interprétée par le degré d'importance de l'interaction entre les critères inclus dans le sous ensemble C_1. La fonction d'agrégation de pertinence personnalisée basée sur l'intégrale de Choquet est définie comme suit :

2. Ce papier est une version étendue du papier publié sur les actes d'INFORSID'2014 (Moulahi et *al.* 2014a)

DÉFINITION 2. — *$RSV^u_{\mathcal{C}}(q, d_j)$ est le score de pertinence personnalisé de d_j pour l'utilisateur u suivant l'ensemble des critères de pertinence $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ défini comme :*

$$RSV^u_{\mathcal{C}}(q, d_j) = Ch_\mu(RSV^u_{c_1}(q, d_j), \ldots, RSV^u_{c_N}(q, d_j))$$
$$= \sum_{i=1}^{N} \mu^u_{\{c_i,\ldots,c_N\}} \cdot (rsv^u_{(i)j} - rsv^u_{(i-1)j})$$

Où Ch_μ la fonction d'agrégation de Choquet, $rsv^u_{(i)j}$ est le $i^{\text{ème}}$ élément de la permutation $RSV(q, d_j)$ sur le critère c_i, tel que ($0 \leq rsv^u_{(1)j} \leq ... \leq rsv^u_{(N)j}$) et $rsv^u_{(0)j} = 0$, μ est la mesure floue déjà définie.
De cette manière, nous sommes capables d'ajuster les paramètres du modèle d'ordonnancement automatiquement pour chaque utilisateur, rendant ainsi les résultats dépendants de ses préférences sur les critères considérés. Notons que si μ est une mesure additive, l'intégrale de Choquet correspond à la moyenne pondérée. Sinon, elle demande moins de 2^N mesures de capacité dans le cas où la mesure est $k-$additive, *i.e.,* $\mu_A = 0$ pour tous les sous ensembles de critères $A \subseteq \mathcal{C}$ avec $|A| > k$. D'un point de vue théorique, l'intégrale de Choquet dispose d'un nombre de propriétés qui semblent être pertinentes pour un domaine tel que la RI ; étant donné qu'elle est construite à partir du concept de mesure floue, elle permet la modélisation des relations d'interaction flexibles en considérant des relations de dépendance complexes entre les critères (Grabisch *et al*., 2000). Nous distinguons trois types d'interactions qui peuvent être modélisées par la mesure floue.

– *Interaction positive*, appelée aussi *synergie positive*, quand le poids global de deux critères est supérieure à leur poids individuels : $\mu_{\{c_i,c_j\}} > \mu_{c_i} + \mu_{c_j}$. Cette inégalité peut être interprétée comme suit : *"la contribution de c_j à toute combinaison de critères contenant c_i est strictement supérieure à la contribution de c_j à la même combinaison quand c_i est exclu"*. Dans ce cas, c_i et c_j sont négativement corrélés, i.e., la satisfaction d'un critère unique doit produire un impact très faible par rapport à la satisfaction des deux critères ensemble. Intuitivement, dans un contexte de RI, cette propriété favorise les documents qui sont satisfaits équitablement par tous les ensembles de critères, plutôt que les documents sur-estimés selon un seul critère de pertinence. Dans ce cas, les critères peuvent être également considérés comme présentant un degré de complémentarité ou d'opposition

– *Interaction négative* (*synergie négative*), quand le poids global de deux critères est plus petit que leurs poids individuels : $\mu_{\{c_i,c_j\}} < \mu_{c_i} + \mu_{c_j}$. Dans ce cas, on peut dire que l'union des critères n'a pas de valeur ajouté sur l'évaluation globale des documents, i.e., la contribution marginale de c_j à chaque combinaison de critères contenant c_i est strictement inférieure à la contribution marginale de c_j à cette même combinaison mais où c_i est exclu. Ces deux critères présentent alors une sorte de redondance. Cette spécificité est parmi les points clés de l'intégrale de Choquet, vu qu'elle permet d'absorber le biais qui pourrait être introduit par l'implication des critères de pertinence redondants dans l'évaluation globale des documents. Ceci est effectué par l'association d'un degré d'importance μ_{c_i,c_j} relativement faible au sous ensemble des critères positivement corrélés.

– *Indépendance*, quand il n'existe aucune corrélation entre l'ensemble des critères. Dans ce cas, on dit que la mesure floue est additive : $\mu_{\{c_i,c_j\}} = \mu_{c_i} + \mu_{c_j}$. La moyenne arithmétique pondérée est exemple de ce type de fonction qui permet l'indépendance des critères. Le poids de chaque critère indique son importance relative.

Pour faciliter la tâche d'interprétation du modèle résultat de l'intégrale de Choquet, nous allons exploiter deux paramètres appelés, "indice d'importance" et "indice d'interaction" (Grabisch *et al*., 2000) qui permettent de traduire les relations ainsi que l'importance des critères. L'indice d'importance, appelé également indice de Shapley, permet d'estimer la contribution moyenne qu'un critère (c_i) apporte à toutes les autres combinaisons de critères possibles. L'indice d'interaction permet de donner des informations sur le phénomène d'interaction pouvant exister entre un ensemble de critères.

DÉFINITION 3 (Indice de Shapley). — *Soit μ_{c_i} le poids du critère c_i et $\mu_{Cr\cup c_i}$ sa contribution marginale à chaque sous ensemble de critères $Cr \in \mathcal{C}$. L'indice d'importance de c_i selon la mesure floue μ est défini comme la moyenne de toutes ces contributions :*

$$\phi_\mu(c_i) = \sum_{Cr\subseteq\mathcal{C}\setminus\{c_i\}} \frac{(N-|Cr|-1)!.|Cr|!}{N!}\left[\mu_{Cr}.\mu_{(Cr\bigcup c_i)}\right]$$

$\phi_\mu(c_i)$ mesure la contribution moyenne que (c_i) fournit à toutes les combinaisons de critères possibles.

L'indice d'importance ne donne aucune information sur le phénomène d'interaction pouvant exister entre les critères. L'importance globale de c_i ne peut pas être uniquement déterminés par son poids μ_{c_i}, mais aussi avec sa contribution marginale à tous les autres sous ensembles de critères. Alors, pour quantifier le degré d'interaction entre ces derniers, nous introduisons dans ce qui suit, le concept d'indice d'interaction.

DÉFINITION 4 (Indice d'interaction). —

Soit $(\Delta_{c_ic_j}\mu_{Cr})$, avec $Cr = \mathcal{C}\setminus\{c_i, c_j\}$, est la différence entre la contribution marginale du critère c_j à toute combinaison de critère contenant c_i, et une combinaison dans laquelle c_i est exclu.

$$(\Delta_{c_ic_j}\mu_{Cr}) = \left[\mu_{(\{c_ic_j\}\cup Cr)} - \mu_{(c_i\cup Cr)}\right] - \left[\mu_{(c_i\cup Cr)} - \mu_{Cr}\right]$$

Cette expression est définie pour estimer l'opposition entre deux critère c_i et c_j. Quand cette expression est positive (*resp.* négative) pour tout $Cr \in \mathcal{C}\setminus\{c_i, c_j\}$, on dit que les deux critères interagissent positivement (*resp.* négativement) (i.e., la contribution du critère c_j est plus significative avec la présence de c_i). L'interaction entre les deux mesures est alors définie comme suit :

$$I_\mu(c_i, c_j) = \sum_{Cr\subseteq\mathcal{C}\setminus\{c_i,c_j\}} \frac{(N-|Cr|-2)!.|Cr|!}{(N-1)!}(\Delta_{c_ic_j}\mu_{Cr})$$

Quand les deux critères sont indépendants, la valeur d'interaction, qui appartient à l'intervalle $[-1, 1]$, est nulle. Dans le cas où les deux critères interagissent positivement (*resp.* négativement), la valeur est positive (*resp.* négative).

Il est à noter que certaines méthodes issues du domaine d'apprentissage automatiques permettent aussi de donner une idée sur le poids d'importance d'un critère, comme par exemple les poids de connexion dans les réseaux de neurones. Toutefois, comme précédemment indiqué dans plusieurs travaux de l'état de l'art, ces méthodes ne permettent pas d'expliquer pourquoi un critère doit être plus important qu'un autre (Eickhoff *et al*., 2013). De plus, les poids d'importance ne sont définis que sur des critères individuels et non pas des sous critères tel que c'est le cas avec l'intégrale de Choquet.

3.2. Apprentissage des préférences des utilisateurs

Tableau 1. Synthèse des notations utilisées avec l'algorithme 1

Notation	**Description**
Q^u_{app}	L'ensemble des requêtes utilisées pour apprendre les valeurs de capacités de l'utilisateur u
N	Nombre de critères de pertinence
$\mathcal{D}$	La collection de documents
K	Nombre de documents utilisés pour l'apprentissage pour chaque requête
$\gamma^{i,r}$	Liste ordonnée de documents en réponse à la requête q_r suivant la combinaison de capacité $\mu^{(i)}$. Soit $P@X(\gamma^{r,i})$ la $P@X$ de $\gamma^{r,i}$ et $AVP@X(\gamma^i)$ soit sa moyenne de $P@X$ sur toutes les requêtes $\in Q_{app}$ suivant $\mu^{(i)}$
I_{Cr}	Tous les sous ensembles de critères possibles de Cr
$\mathcal{S}_\mu$	Ensemble de combinaisons de capacité expérimentées. Chaque combinaison $\mu^{(i)} \in \mathcal{S}_\mu$ contient les valeurs de capacités de tous les ensembles et sous ensemble de critères

L'objectif de la phase d'apprentissage est d'optimiser les mesures floues selon une mesure objective de RI (e.g. $P@X$) en identifiant les valeurs de capacité. La méthode proposée doit permettre aussi de personnaliser les résultats de recherche d'un utilisateur en particulier, tout en considérant ses préférences individuelles sur les critères de pertinence.

Nous proposons dans ce qui suit un algorithme générique permettant d'apprendre ces capacités indépendamment du nombre de critères de pertinence, et de la tâche de RI considérée.

Les données d'apprentissage nécessaires pour identifier les mesures floues de l'intégrale de Choquet comprennent un ensemble de requêtes d'apprentissage, et pour chaque requête, un ensemble trié de documents représentés par des vecteurs contenant des scores partiels selon chaque critère ; chaque document est annoté avec une

Algorithme 1 : Apprentissage des mesures floues

Entrées: Q_{learn}, N, K.
Sortie: Combinaison de capacité optimale $\mu^{(**)}$.

Étape 1 : Initialisation des valeurs de capacités

$m \leftarrow (2^N - 1) \times N$;

1. **Pour** $i = 1$ à m {*Identification des combinaisons de capacités*} **Faire**
2. $\quad \mu^{(i)} = (\bigcup_{j:1..N} \{\mu_{c_j}\}) \bigcup (\bigcup_{Cr \in \mathcal{C}, |Cr|>1} \{\mu_{I_{Cr}}\})$; $\mu_{I_{Cr}} = \sum_{c_i \in Cr, |c_i=1|} \mu_{c_i}$
3. **Fin Pour**
4. **Si** $N \geq 4$ {*Supposer la* 2*-additivité*} **Alors**
5. $\quad$ **Pour** chaque $I_{Cr} \in \mu^{(i)}$ tel que $|Cr| > 2$ **Faire**
6. $\quad\quad \mu_{I_{Cr}} = 0$
7. $\quad$ **Fin Pour**
8. **Fin Si**
9. $\mathcal{S}_\mu = \bigcup_{i:1..m} \{\mu^{(i)}\}$
10. **Pour** chaque $\mu^{(i)} \in \mathcal{S}_\mu$ {*paramétrage des capacités*} **Faire**
11. $\quad$ Calculer $AVP@X(\gamma^i)$
12. **Fin Pour**
13. $Cmax$ = $\underset{1...|\mathcal{S}_\mu|}{\text{Argmax}}$ $(AVP@X(\gamma^i))$; $\mu^{(*)} = \mu^{(cmax)}$

Étape 2 : Optimiser les valeurs de capacités

14. $D^{app} = \emptyset$
15. **Pour** $r = 1$ à $|Q^{app}|$ {*Interpoler les scores globaux*} **Faire**
16. $\quad D^{app} = D^{app} \cup \gamma^{*,r}$
17. $\quad$ **Pour** $j = 1$ à K **Faire**
18. $\quad\quad RSV^{int}_{\mathcal{C}}(q_r, d_j) = \underset{1...d'_j \in \gamma^{*,r}, d'_j \succ_{\mathcal{C}} d_j}{\text{Max}} (RSV_{\mathcal{C}}(q_r, d'_j))$; $\gamma^{*,r} = \gamma^{*,r} \setminus \{d_j\}$
19. $\quad$ **Fin Pour**
20. **Fin Pour**

{*Optimisation basée sur la méthode des moindres carrés.*}

21. **Répéter**
22. $\quad \mathcal{F}_{LS}(\mu) = \sum_{d_j \in D^{app}} [Ch_\mu(\text{RSV}_{c_1}(d_j), \ldots, \text{RSV}_{c_N}(d_j)) - \text{RSV}^{int}_{\mathcal{C}}(d_j)]^2$
23. **Jusqu'à** convergence
24. **Retourner** le résultat $\mu^{(**)}$

étiquette (*e.g.,* pertinent ou non pertinent). La méthodologie adoptée est détaillée dans l'algorithme 1. Le Tableau 1 décrit les notations utilisées dans cet algorithme. Ce dernier comprend deux étapes principales :

– *Initialisation des valeurs initiales des combinaisons de capacités.* Une combinaison de capacités $\mu^{(.)}$ désigne l'ensemble des valeurs de capacités associées à chaque critère et à chaque sous-ensemble de critères. Par exemple, dans le cas de trois critères de pertinence, une combinaison de capacités comprend

($\{\mu_{c_1}; \mu_{c_2}; \mu_{c_3}; \mu_{c_1,c_2}; \mu_{c_1,c_3}; \mu_{c_2,c_3}\}$). Afin de paramétrer ces valeurs, nous utilisons une mesure de RI telle que la $P@X$ sur les requêtes d'apprentissage Q^u_{app}. Le paramétrage est concevable étant donné que le nombre de critères de pertinence est généralement petit (Saracevic, 2007). Cependant, lorsque le nombre de critères est supérieur ou égal à 4, nous pouvons réduire la complexité du paramétrage (i.e., initialement 2^{N-2} valeurs de capacité à identifier) en nous basant sur la famille des capacités 2-additive (Grabisch *et al.*, 2000) nécessitant $N-2$ coefficients à définir.

– *Optimisation des valeurs de capacités.* En partant d'une combinaison de capacités $\mu^{(*)}$ obtenue dans l'étape précédente, on extrait les K premiers documents retournés en réponse à chaque requête $q \in Q^u_{app}$. Les scores de ces documents (D^u_{app}) sont interpolés pour placer les documents non pertinents à la fin de l'ordonnancement. Après avoir obtenu les scores de pertinence globaux désirés $RSV^{int}_{\mathcal{C}}(q, d_j)$ pour chaque document $d_j \in D^u_{app}$, et étant donné que nous disposons des étiquettes $RSV^u_{c_i}(q, d_j)$, nous procédons à l'application de la méthode des moindres carrées pour l'identification des valeurs de capacités des critères et des sous-ensembles de critères considérés.

Notons que dans le cas d'un contexte de RI personnalisée, cet algorithme est appliqué pour chaque utilisateur permettant ainsi d'ajuster les degrés d'importance de chaque critère de pertinence selon les préférences de l'utilisateur en question. En revanche dans le cas d'un contexte de RI non personnalisée, cet algorithme est appliqué indifféremment pour toutes les requêtes issues de la tâche, permettant de quantifier des scores de pertinences génériques liées à la tâche et non spécifiquement aux utilisateurs. On calcule alors pratiquement $RSVc_i(q, d_j)$ au lieu de $RSV^u_{c_i}(q, d_j)$.

4. Cadre expérimental

Notre évaluation expérimentale est basée sur trois tâches de RI. La première tâche correspond à une recherche de tweets au sein d'une collection de microblogs, proposée dans le cadre de la tâche Microblog de TREC 2011 et 2012 (Ounis *et al.*, 2011 ; Soboroff *et al.*, 2012). Le deuxième cadre d'évaluation concerne une tâche de recherche personnalisée dans les folksonomies. Nous nous basons sur une collection de signets collectés à partir du système d'annotations sociales *Del.icio.us* [3]. Enfin, nous évaluons notre approche dans une tâche de RI contextuelle en nous basant sur la collection de test standard fournie par la tâche "*Contextual Suggestion*" de TREC [4] 2013 (Dean-Hall *et al.*, 2013). Nous utilisons la plateforme Terrier [5] pour l'indexation et la recherche.

3. http://www.delicious.com
4. Text REtrieval Conference (http://trec.nist.gov/)
5. http://terrier.org

4.1. Cadre de recherche d'information sociale

Dans cette section, nous décrivons le cadre expérimental de la première tâche d'évaluation. Cette dernière est une tâche en temps réel dans laquelle les utilisateurs s'intéressent à l'information pertinente et récente, à la fois.

4.1.1. Données expérimentales

Nous exploitons ici la collection de tweets fournie par la tâche Microblog de TREC 2011 et TREC 2012 (Ounis *et al*., 2011 ; Soboroff *et al*., 2012). La collection inclut environ 16 millions de tweets publiés sur 16 jours. Les statistiques sont données dans le Tableau 2.

Tableau 2. Statistiques de la collection fournie par la tâche Microblog de TREC 2011 et 2012

Tweets	$16,141,812$
Tweets null	$1,204,053$
Termes uniques	$7,781,775$
Nombre de twitters	$5,356,432$
Nombre de Topics de TREC Microblog 2011	49
Nombre de Topics de TREC Microblog 2012	60

4.1.2. Protocole d'évaluation

Nous avons exploité 49 requêtes de la tâche Microblog de TREC 2011 pour l'apprentissage des capacités et nous avons utilisé les 60 requêtes de TREC 2012 pour le test. Trois critères de pertinence liés à la tâche ont été utilisés pour le calcul des scores des documents (Nagmoti *et al*., 2010) : sujet et fraîcheur de l'information et autorité du *twitterer*. Le calcul de ces différents critères est décrit dans (Moulahi, Tamine, Yahia, 2014). Pour évaluer les performances de notre approche dans ce cadre, nous avons comparé les résultats issus de l'application de l'opérateur proposé à ceux issus de l'application d'opérateurs d'agrégation classiques tels que la méthode de combinaison linéaire (MCL), ainsi que quelques algorithmes d'apprentissage d'ordonnancement (*learning to rank*) tels que RANDOM FOREST (RF) (Breiman, 2001) et $\lambda-$MART (Burges *et al*., 2005). Les mesures d'évaluation utilisées sont la précision ($P@5$, $P@10$, $P@20$, $P@30$) et la précision moyenne (MAP).

4.2. Cadre de recherche personnalisée dans les folksonomies

Dans cette section, nous décrivons le cadre d'évaluation proposé dans le contexte d'une RI personnalisée. Dans ce qui suit, nous décrivons la collection de test ainsi que le protocole d'évaluation utilisés.

4.2.1. Données expérimentales

Nous avons exploité une collection de 33k signets [6] collectés à partir du système d'annotation *Del.icio.us*. Le corpus inclut des informations d'évaluation données pour 35 utilisateurs en réponse à 177 requêtes (Vallet, Castells, 2012), suivant deux dimensions de pertinence : thématique (Th) des signets étant donnée une requête et leurs pertinence personnelle (Us) étant donné un utilisateur.

4.2.2. Protocole d'évaluation

Pour cette deuxième tâche de recherche, nous utilisons 75% des requêtes pour l'apprentissage et nous exploitons le reste des requêtes pour le test. Étant donné que pour cette collection de documents, nous disposons uniquement des 5 premiers résultats pertinents pour chaque requête, nous exploitons la mesure $P@5$ pour l'évaluation des résultats de recherche comme recommandé dans (Vallet, Castells, 2012).

4.3. Cadre de recherche d'information contextuelle

La tâche "*Contextual Suggestion*" de TREC a pour objectif d'évaluer les techniques de recherche répondant à des besoins en information, qui sont fortement tributaires du contexte et des centres d'intérêts des utilisateurs. Étant donné un utilisateur, cette tâche a pour objectif de chercher des places d'attractions (e.g., restaurants, parcs d'attractions, zoo, etc.) pouvant l'intéresser suivant deux critères de pertinence : (1) les centres d'intérêt de l'utilisateur, *i.e.,* ses préférences personnelles sur un historique de recherche de places ; (2) sa localisation géographique.

4.3.1. Données expérimentales

La collection de test présente les caractéristiques suivantes :

– **Utilisateurs :** le nombre total d'utilisateurs est égal à 635. Chaque utilisateur est représenté par un profil reflétant ses préférences sur des lieux d'une liste de 50 exemples de suggestions. Un exemple de suggestion est un lieu d'attraction qui est susceptible d'intéresser l'utilisateur. Chaque exemple est représenté par le titre du lieu, une brève description et une URL du site web correspondant. Les préférences des utilisateurs sont données sur une échelle de 5 points et sont attribuées aux descriptions et aux URLs des exemples de suggestions. Les préférences positives (*resp.,* négatives) sont celles ayant un degré de pertinence égal à 3 ou à 4 (*resp.,* 0 ou 1) selon la description du site et la correspondance par rapport à l'URL.

– **Contextes (requêtes) :** le nombre de contextes fournis est égal à 50 ; chaque contexte correspond à une position géographique dans une ville donnée. La position géographique est décrite par une longitude et une latitude. Étant donnée un utilisateur et un contexte représentant la requête, l'objectif principal de la tâche est de fournir une

6. http://ir.ii.uam.es/ david/webdivers/

liste de 50 suggestions triée par ordre de pertinence selon les critères centres d'intérêt de l'utilisateur et géolocalisation.

– **Collection de documents :** pour chercher des suggestions de lieux à partir du web, nous avons exploité l'API Google Place [7]. Comme pour la plupart des groupes participant à la tâche "*Contextual Suggestion*" (Dean-Hall *et al*., 2013), nous commençons par interroger l'API Google Place avec les requêtes appropriées en se basant sur la localisation géographique des lieux. Étant donné que l'*API Google Place* renvoie jusqu'à 60 suggestions par requête, nous avons effectué une nouvelle recherche avec des paramètres différents tels que les types de lieux qui sont pertinents par rapport à la tâche (*e.g.,* restaurant, pizzeria, musée, etc.). Nous avons collecté, en moyenne, environ 157 suggestions par requête et 3 925 suggestions au total. Pour obtenir les scores des documents collectés selon le critère de géolocalisation, nous avons calculé la distance entre les lieux collectés et le contexte. Les scores des documents selon le critère centres d'intérêts est calculé en se basant sur le cosinus de similarité entre la description des suggestions et le profil de l'utilisateur. Les profils des utilisateurs sont représentés par des vecteurs de termes construits à partir de leurs préférences personnelles sur les exemples de suggestions. La description des lieux est construite à partir des "*snippets*" des résultats renvoyés par le moteur de recherche Google [8] lorsque l'URL du lieu est soumise sous forme d'une requête.

– **Jugements de pertinence :** les jugements de pertinence de cette tâche sont effectués par les utilisateurs et mandatés par TREC à la fois (Dean-Hall *et al*., 2013). Chaque utilisateur représenté par un profil, juge les lieux qui lui sont suggérés de la même façon que les exemples de suggestions. Ainsi, l'utilisateur affecte un jugement de $0 - 4$ à chaque titre/description et à chaque URL, tandis que les assesseurs de TREC jugent les suggestions uniquement en termes de correspondance au critère géolocalisation avec une évaluation de (2, 1 et 0). Une suggestion est considérée comme pertinente si elle a un degré de pertinence égal à 3 ou 4 selon le critère centre d'intérêts (profil) et une évaluation égale à 1 ou 2 selon le critère géolocalisation. Dans ce qui suit, ces jugements de pertinence constituent notre réalité de terrain utilisée pour l'apprentissage et le test.

4.3.2. Protocole d'évaluation

Nous avons adopté une méthodologie entièrement automatisée basée sur une validation croisée afin d'identifier les valeurs de capacité des utilisateurs et tester les performances du modèle d'agrégation. À cette fin, nous avons procédé à une partition aléatoire de l'ensemble des 50 contextes en deux ensembles de même taille, noté Q^u_{app} et Q^u_{test} utilisés respectivement pour l'apprentissage et le test. En outre, pour éviter le problème de sur-apprentissage, l'ensemble des contextes est divisé aléatoirement dans un second tour en deux ensembles différents d'apprentissage et de test.

7. https://developers.google.com/places
8. https://www.google.com

L'objectif principal de la phase d'apprentissage est d'apprendre les capacités ($\mu^u_{\{centre\_interet\}}$, $\mu^u_{\{localisation\}}$) qui correspondent à l'importance des critères de pertinence. Nous commençons d'abord par une mesure floue initiale donnant le même poids d'importance pour les deux critères de pertinence. Ensuite, nous calculons la mesure de précision $P@5$ de tous les contextes de l'ensemble d'apprentissage Q^u_{app}. En utilisant la vérité de terrain fournie avec la tâche "*Contextual Suggestion*" de TREC 2013, et en se basant sur l'algorithme 1, nous identifions pour chaque utilisateur ses préférences personnelles sur les deux critères : centres d'intérêts et localisation géographique. Enfin, pour tester l'efficacité de notre approche, nous nous sommes appuyés sur l'ensemble de contextes restants Q^u_{test} et nous avons utilisé la mesure officielle de la tâche $P@5$ pour le calcul de performances. Cette mesure de précision est équivalente à la proportion des suggestions de lieux pertinents retournés parmi les 5 premiers.

5. Résultats expérimentaux

5.1. Analyse de l'importance des critères de pertinence

Notre premier objectif consiste ici à analyser les valeurs de capacité issues de l'algorithme 1, représentant le degré d'importance des critères de pertinence utilisés dans chaque cadre de RI.

5.1.1. Importance des critères dans la tâche de RI sociale

L'analyse d'importance des critères en utilisant l'indice d'importance montre une importance du critère thématique avec une valeur de $0,631$. Le critère fraîcheur d'information est aussi donné peu d'importance $(0,25)$ comparé à l'autorité des utilisateur $(0,12)$. Ceci peut s'expliquer par le fait que les *twitterer* cherchent généralement des documents thématiquement pertinents plutôt que récents ou autoritaires. Ainsi, lors de l'évaluation des tweets de la collection par les assesseurs de TREC [9], ces derniers accordent plus d'importance à la correspondance thématique entre les requêtes et les documents.

La Figure 1 montre les valeurs des indices d'interactions des trois critères thématique *To*, fraîcheur d'informations *Re* et autorité *Au* suivant les requêtes des deux tâches Microblog 2011 et 2012. On remarque que le critère autorité n'est pas important et ne donne aucune contribution quand il est combiné avec les deux autres critères. Cependant, nous notons une interaction positive entre le critère thématique et fraîcheur d'information, ce qui implique une grande contribution aux scores globaux lorsqu'ils sont combinés ensemble.

9. http://trec.nist.gov

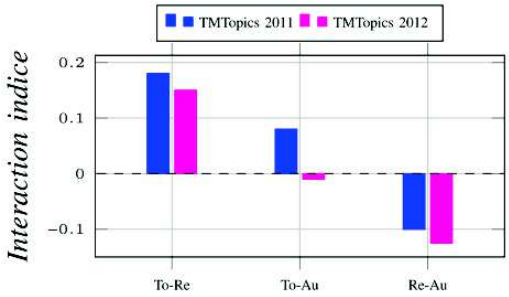


Figure 1. Indice d'interaction des critères

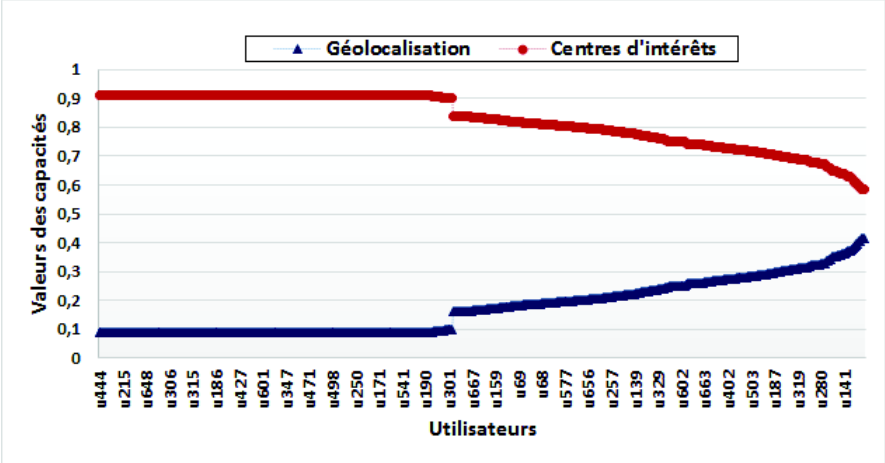
5.1.2. Importance des critères dans la tâche de RI personnalisée

Après le calcul des indices d'importance des deux critères utilisés dans cette tâche, nous avons trouvé que tous les deux ont presque le même degré d'importance avec une valeur de $0,48$ pour la pertinence thématique et $0,51$ pour la pertinence utilisateur. Par ailleurs, la valeur d'interaction entre ces deux derniers est égale à $0,028$, ce qui est un peu faible pour supposer qu'ils sont réellement dépendants. Ce résultat a eu un effet sur les performances de notre approche (cf. section 5.2.2).

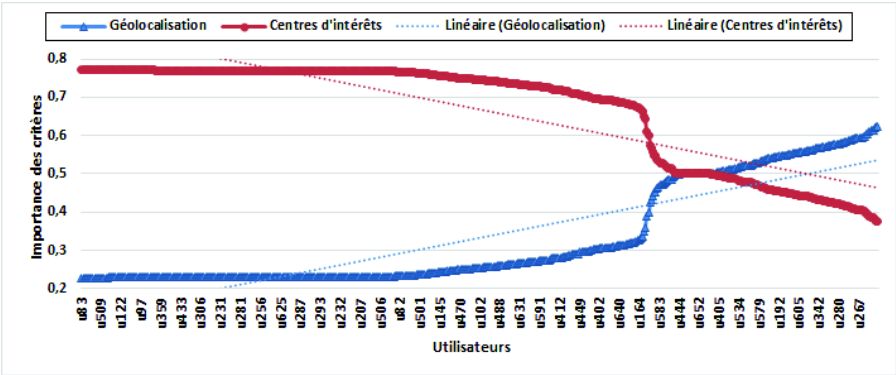
5.1.3. Importance des critères dans la tâche de RI contextuelle

Dans cette section, nous analysons l'importance des critères pour les utilisateurs ($\mu^u_{\{centre\_interet\}}$, $\mu^u_{\{geolocalisation\}}$). A cet effet, nous commençons par analyser l'importance intrinsèque de chaque critère indépendamment des autres critères. La Figure 3(a) montre la variation des valeurs de capacité pour chaque utilisateur selon les deux critères de pertinence sur l'ensemble Q^u_{app} d'apprentissage. L'axe des abscisses représente l'ensemble des utilisateurs (35-669) et l'axe des ordonnées représente les valeurs de capacité correspondantes selon les critères centres d'intérêt (Ci) et géolocalisation (Geo).

En se référant à la Figure 2(a), nous constatons que le critère Ci se voit accorder une capacité plus importante que le critère Geo. Par exemple, l'utilisateur 285 a une valeur de capacité de l'ordre de $0,23$ pour le premier critère alors qu'il a une mesure de l'ordre de $0,76$ pour le critère Geo. Ceci est prévisible étant donné que les utilisateurs de cette tâche s'intéressent généralement aux lieux qui correspondent à leurs préférences personnelles, même si elles ne sont pas géographiquement pertinentes. Cependant, la Figure 2(a) montre que la distribution des valeurs de capacité est loin d'être la même pour tous les utilisateurs et met en exergue des valeurs qui vont de $0,09$ à $0,414$ pour le critère Geo et d'autres qui vont de $0,585$ à $0,909$ pour le critère Ci. Pour mieux comprendre ce constat, nous traçons sur la Figure 2(b), les valeurs des indices d'importance reflétant, pour chaque utilisateur, le degré de préférence globale selon les deux critères de pertinence Ci et Geo. A la différence de la Figure 2(a), la Figure 2(b) met en évidence l'importance moyenne de chaque critère de pertinence quand il est associé à l'autre critère. On peut observer sur la Figure 2(b) que les préfé-



(a) Valeurs de capacités des utilisateurs suivant les deux critères de pertinence centres d'intérêt et géolocalisation



(b) Importance des critères centres d'intérêt (Ci) et géolocalisation (Geo)

Figure 2. Valeurs de capacités des utilisateurs et importance des critères de la tâche "Contextual Suggestion" de TREC 2013

rences des utilisateurs sur les deux critères sont totalement différentes. Le lissage des valeurs d'importance obtenues selon ces critères donne deux courbes linéaires avec des valeurs tout à fait constantes et différentes, corroborant ainsi les résultats obtenus sur la Figure 2(a). Le critère "centre d'intérêt" est encore pondéré par une importance relativement élevée pour la plupart des utilisateurs. Néanmoins, on peut également remarquer au milieu de la figure (valeurs comprises entre $0,4$ et $0,7$) que certains utilisateurs ont une préférence élevée sur le critère géolocalisation et inversement.

Dans une seconde étape, nous analysons à travers la Figure 3, la dépendance entre les critères pour chaque utilisateur par le biais de l'indice d'interaction (Grabisch, 1995). Plus les valeurs de cet indice sont proches de 1 (*resp.,* -1) plus les deux cri-

tères sont dépendants et l'interaction est positive (*resp.,* négative). Si la valeur de l'indice d'interaction est égale 0, les deux critères sont considérés comme indépendants et par conséquent, il n'existe aucune interaction entre ces derniers. On peut constater que les valeurs obtenues sur tous les utilisateurs sont toutes positives et varient entre $0,28$ et $0,99$. La valeur moyenne est de l'ordre de $0,56$ ce qui implique une interaction positive entre les deux critères de pertinence considérés lorsqu'ils sont combinés ensemble.

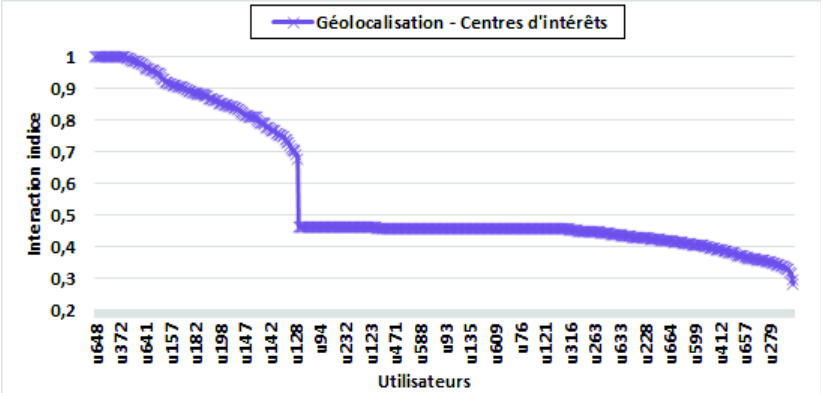


Figure 3. Indices d'interaction entre les critères de pertinence centres d'intérêt et géolocalisation pour chaque utilisateur

5.2. Analyse des performances de recherche

Dans cette section nous analysons les résultats des performances de recherche de notre approche selon les trois scénarios d'évaluation déjà énoncés. Pour chaque scénario d'évaluation, nous avons utilisé les métriques ainsi que les scénarios de référence (*Baseline*) appropriés. La significativité des éventuels accroissements obtenus ont été estimés en appliquant le test statistique de *student*.

5.2.1. Estimation de la pertinence dans le cadre de RI sociale

Le Tableau 3 montre les performances de recherche obtenues par l'opérateur de Choquet avec les référentiels de comparaison. Nous avons commencé par le paramétrage de l'algorithme $\lambda-$MART à l'aide d'une validation croisée. Nous avons trouvé que le meilleur nombre d'itérations lié à l'algorithme est 1000 alors que le taux d'apprentissage est égal à $0,1$.

Le Tableau 3 montre que notre approche est plus performante que la méthode de combinaison linéaire ainsi que les algorithmes d'apprentissage d'ordonnancements. Le taux d'amélioration est égal à $20,73\%$ pour la méthode MCL alors qu'il est plus important pour l'algorithme RF. Cette amélioration peut être expliquée par la considération de la dépendance éventuelle entre les critères, ce qui a permis de réduire le

*Tableau 3. Évaluation comparative des performances recherche. "% Amélioration" indique l'amélioration de notre approche en terme de $P@30$. Les symboles § et ⋆ dénotent le test t-*student *: "§": $0,05 < t \leqslant 0,1$; "⋆": $t \leqslant 0,01$*

	Précision						
APPROCHE	**P@5**	**P@10**	**P@20**	**P@30**	**MAP**	**% Amélioration**	
MCL	$0,1965$	$0,1860$	$0,1833$	$0,1854$	$0,1309$	**+20,73 %**	§
RF	$0,1000$	$0,0810$	$0,0681$	$0,0687$	$0,0628$	**+70,68 %**	⋆
$\lambda-$**MART**	$0,2931$	$0,2276$	$0,2092$	$0,2043$	$0,1856$	**+11,67 %**	⋆
Notre approche	**0,2379**	**0,2362**	**0,2422**	**0,2313**	**0,1295**	–	

biais à l'aide de l'introduction d'un degré d'importance aux sous ensembles de critères. Dans ce contexte d'évaluation, nous notons que nous avons appliqué l'opérateur de Choquet standard non personnalisé, étant donné que nous ne disposons pas d'une vérité de terrain liée aux préférences individuelles des utilisateurs sur les différentes dimensions de pertinence, tel est le cas pour les deux cadres d'évaluations qui suivent.

5.2.2. Estimation de la pertinence dans le cadre de RI personnalisée dans les folksonomies

Dans cette section, nous évaluons notre approche d'agrégation personnalisée dans un contexte de RI dans les folksonomies. Dans le Tableau 4, nous remarquons que les résultats obtenus avec l'opérateur de Choquet sont très proches des résultats obtenus avec les modèles de référence. Ceci peut être expliqué par le fait que le nombre de critères est réduit d'une part et qu'ils sont en plus indépendants d'autre part, comme montré dans la section 5.1.2.

Tableau 4. Évaluation comparative des performances de recherche dans le contexte de RI personnalisée. Le symbole ⋆ dénote le test t-student : "⋆⋆⋆": $t \leqslant 0,01$

	MCL	**OWA**	**AND**	**SCORING**	**RANKSVM**	**Notre approche**
P@5	**0,6310**	**0,6310**	$0,6286$	**0,6310**	$0,6286$	**0,6310**
% ↗	0%	0%	**+0,003 %**	0%	**+0,003 %**	–
	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	

5.2.3. Estimation de la pertinence dans le cadre de RI contextuelle

Notre second objectif est d'évaluer les performances de notre approche en termes : (i) d'agrégation de pertinence multidimensionnelle ; et (ii) de personnalisation des préférences des utilisateurs sur les critères de pertinence. Pour ce faire, nous comparons les résultats obtenus sur l'ensemble de contextes de test Q^u_{test} aux méthodes d'agrégation de référence (*baseline)* : la moyenne arithmétique pondérée (MAP) largement utilisée dans la plupart des approches impliquant la combinaison des scores de pertinence et les deux opérateurs d'agrégation prioritaires SCORING et AND, précé-

demment utilisés pour l'agrégation de pertinence dans un cadre de RI personnalisée. Il convient de préciser que nous avons effectué une série d'expérimentations avec une validation croisée pour identifier les meilleurs scénarios de priorisation devant être utilisés avec les deux opérateurs SCORING et AND sur le même ensemble d'apprentissage utilisé pour trouver les valeurs de capacité de Choquet. Comme pour les résultats obtenus dans la phase d'analyse des indices d'importance, nous avons également constaté que le meilleur scénario est celui donnant une priorité au critère "centres d'intérêt" des utilisateurs. Cependant, les opérateurs d'agrégation ne sont pas en mesure de quantifier le degré d'importance des critères comme c'est le cas pour l'intégrale de Choquet.

Afin de montrer l'efficacité de l'approche de personnalisation, nous comparons notre opérateur d'agrégation personnalisé Choquet, notée CHOPER *versus* l'opérateur d'agrégation Choquet classique non personnalisé. Les capacités utilisées avec l'opérateur de Choquet classique sont obtenus en appliquant l'algorithme 1 une seule fois (et non pas pour chaque utilisateur), donnant ainsi en sortie des valeurs d'importance sur les critères indépendamment des préférences individuelles de chaque utilisateur. Ceci donne lieu à une valeur de $0,86$ pour le critère centre d'intérêt et une valeur de l'ordre de $0,14$ pour le critère géolocalisation. Les mesures de précision obtenues sont moyennées sur toutes les séries de tests et pour l'ensemble des requêtes de test.

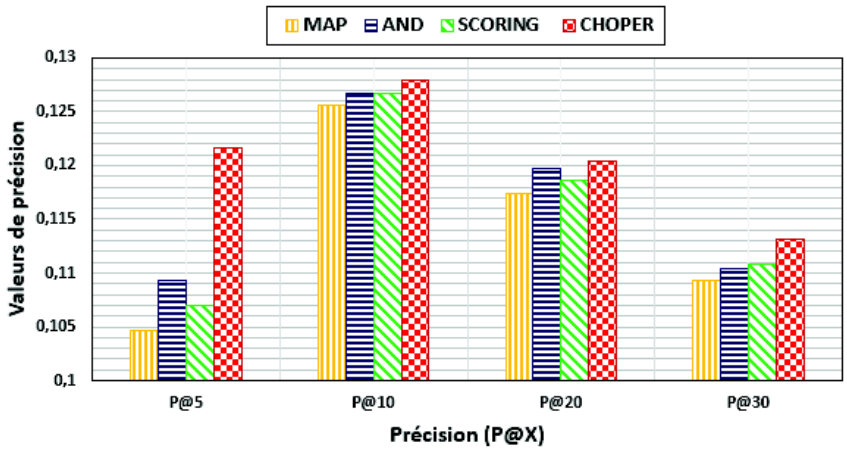


*Figure 4. Efficacité de notre approche d'agrégation de pertinence dans la tâche "*Contextual Suggestion*" de TREC* 2013 *en comparaison avec les méthodes de référence*

La Figure 4 présente les résultats obtenus par notre approche CHOPER, en comparaison avec les méthodes de référence. La Figure 4 montre que les performances de l'opérateur CHOPER sont significativement plus élevées que toutes les autres méthodes suivant la mesure officielle $P@5$, mais également suivant les autres mesures.

La meilleure amélioration obtenue par notre approche suivant $P@5$ est marquée avec la méthode MAP (13.98%). En comparaison avec la meilleure méthode de

référence (*i.e.,* AND), les améliorations sont significatives mais moins importantes $(10,11\%)$ en termes de $P@5$. Ces résultats sont probablement dus au fait que l'opérateur d'agrégation prioritaire AND est principalement basé sur l'opérateur MIN, ceci pourrait pénaliser les lieux pertinents selon le critère le moins important à savoir, le critère géolocalisation. Vu que la plupart des utilisateurs ont une préférence moins importante selon ce critère, la pénalisation de ce dernier permet d'améliorer les performances de recherche. La différence obtenue dans la performance, en faveur de CHOPER, s'explique par la prise en compte des différents niveaux de préférence suivant les deux critères de pertinence ainsi que la prise en compte de l'interaction qui existe entre ces derniers.

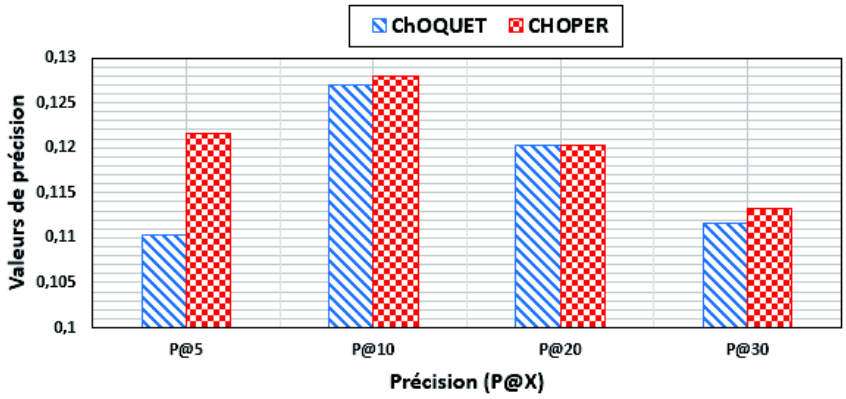


Figure 5. Efficacité de notre approche en terme de personnalisation en comparaison avec l'opérateur d'agrégation de Choquet classique

En termes de personnalisation, la Figure 5 présente les résultats obtenus en termes de précisions ($P@5$, $P@10$, $P@20$ et $P@30$) entre l'opérateur classique Choquet et sa version personnalisée CHOPER. Ces résultats montrent que le dernier est plus performant sur toutes les mesures de précision. La meilleure amélioration est de l'ordre de $9,29\%$ en termes de $P@5$. Ces résultats confirment ceux obtenus dans la phase d'identification des capacités (Cf. section 5.1.3) où nous avons montré que les degrés d'importance des critères dépend des préférences de l'utilisateur et ne sont pas les mêmes pour tous. La prise en compte des poids d'importance appropriés pour chaque critère et chaque utilisateur permet de donner ainsi des résultats à la fois pertinents et adaptés aux préférences personnelles des utilisateurs.

6. Conclusion et perspectives

Dans ce papier, nous avons présenté une approche se reposant sur une méthode d'agrégation floue qui permet d'estimer la pertinence globale des documents dans des cadres des recherche d'information différents. En se basant sur les indices d'importance et d'interaction, notre modèle permet de mesurer et d'interpréter les poids d'importance associés avec chaque critère et sous ensemble de critères. Dans une première

étape, nous avons évalué l'opérateur d'agrégation dans un contexte de RI non personnalisée et nous avons montré que les dimensions de pertinence utilisés présentent une sorte de dépendance. Ainsi, l'utilisation d'une approche tenant compte de ces corrélations, et ne se basant pas sur l'hypothèse d'additivité des critères, peut améliorer les performances de recherche. Ensuite, nous avons testé l'opérateur personnalisé dans une tâche de recherche de signets au sein des folksonomies. Les résultats ont montré que le nombre de critères moins élevé a un effet sur les performances de recherche en comparaison avec des méthodes d'agrégation standards. L'évaluation de notre approche dans une tâche de recherche de lieux d'attraction au sein d'une collection de test fournie par la tâche "*Contextual Suggestion*" de TREC, montre des résultats encourageants grâce à la personnalisation des préférences des utilisateurs.
En perspective, nous envisageons d'étendre l'approche pour qu'elle soit adaptée dans le cas d'absence de scores sur les critères, permettant ainsi de pallier au problème d'insuffisance des exemples d'apprentissage. Une autre direction de recherche intéressante consiste à estimer les valeurs de pertinence dans le cas des collections de documents dynamiques comme le flux de documents continus tel dans le cas des systèmes de RI temps-réel comme Twitter. Ainsi, intégrer le temps dans le processus d'agrégation et de personnalisation des préférences pourrait avoir un impact majeur sur les performances des systèmes de recherche d'information.

Bibliographie

Baeza-Yates R. A., Ribeiro-Neto B. A. (1999). *Modern information retrieval*. ACM Press / Addison-Wesley.

Borlund P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, vol. 54, n° 10, p. 913–925.

Bouidghaghen O., Tamine L., Boughanem M. (2011). Personalizing mobile web search for location sensitive queries. In *International conference on mobile data management*, p. 110–118. IEEE Computer Society.

Bouidghaghen O., Tamine L., Pasi G., Cabanac G., Boughanem M., Costa Pereira C. da. (2011). Prioritized aggregation of multiple context dimensions in mobile IR. In *In proceedings of the 7th asia conference on information retrieval technology*, vol. 7097, p. 169-180. Berlin, Heidelberg, Springer.

Breiman L. (2001). Random forests. *Mach. Learn.*, vol. 45, n° 1, p. 5–32.

Burges C., Shaked T., Renshaw E., Lazier A., Deeds M., Hamilton N. *et al*. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on machine learning*, p. 89–96. New York, NY, USA, ACM.

Cao Z., Qin T., Liu T.-Y., Tsai M.-F., Li H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on machine learning*, p. 129–136. New York, NY, USA, ACM.

Choquet G. (1953). Theory of capacities. *Annales de l'Institut Fourier*, vol. 5, p. 131–295.

Costa Pereira C. da, Dragoni M., Pasi G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing and Management*, vol. 48, nº 2, p. 340–357.

Daoud M., Huang J. X. (2013). Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science*, vol. 64, nº 1, p. 190-212.

Daoud M., Tamine L., Mohand B. (2011). A personalized search using a semantic distance measure in a graph-based ranking model. *Journal of Information Science (JIS)*, vol. 37, nº 6, p. 614–636.

Dean-Hall A., Clarke C., Kamps J., Thomas P., Simone N., Voorhes E. (2013). Overview of the trec 2013 contextual suggestion track. In *Text retrieval conference (trec)*. National Institute of Standards and Technology (NIST).

Duan Y., Jiang L., Qin T., Zhou M., Shum H.-Y. (2010). An empirical study on learning to rank of tweets. In *In proceedings of the 23rd international conference on computational linguistics*, p. 295–303. Stroudsburg, PA, USA, Association for Computational Linguistics.

Eickhoff C., Vries A. P. de, Collins-Thompson K. (2013). Copulas for information retrieval. In *In proceedings of the 36th annual international* ACM SIGIR *conference on research and development in information retrieval.* Dublin, Ireland, ACM.

Gerani S., Zhai C., Crestani F. (2012). Score transformation in linear combination for multi-criteria relevance ranking. In *In proceedings of the 34th european conference on advances in information retrieval*, p. 256–267. Berlin, Heidelberg, Springer-Verlag.

Göker A., Myrhaug H. (2008). Evaluation of a mobile information system in context. *Inf. Process. Manage.*, vol. 44, nº 1, p. 39–65.

Grabisch M. (1995). Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, vol. 69, nº 3, p. 279–298.

Grabisch M., Murofushi T., Sugeno M., Kacprzyk J. (2000). *Fuzzy measures and integrals. theory and applications*. Physica Verlag, Berlin.

Li H. (2011). *Learning to rank for information retrieval and natural language processing*. Morgan & Claypool Publishers.

Mata F., Claramunt C. (2011). Geost: geographic, thematic and temporal information retrieval from heterogeneous web data sources. In *In proceedings of the 10th international conference on web and wireless geographical information systems*, p. 5–20. Berlin, Heidelberg, Springer-Verlag.

Metzler D., Cai C. (2011). USC/ISI at TREC 2011: Microblog track. In *Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Moulahi B., Tamine L., Ben Yahia S. (2014). Prise en compte des préférences des utilisateurs pour l'estimation de la pertinence multidimensionnelle d'un document. In *Inforsid*, p. 295-310.

Moulahi B., Tamine L., Yahia S. B. (2014). iAggregator: Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, vol. 65, nº 10, p. 2062–2083.

Nagmoti R., Teredesai A., De Cock M. (2010). Ranking approaches for microblog search. In *In proceedings of the 2010* IEEE/WIC/ACM *international conference on web intelligence and intelligent agent technology*, vol. 01, p. 153–157. Washington, DC, USA, IEEE Computer Society.

Ounis I., Macdonald C., Lin J., Soboroff I. (2011). Overview of the TREC-2011 microblog track. In *In proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011).*

Palacio D., Cabanac G., Sallaberry C., Hubert G. (2010). On the evaluation of geographic information retrieval systems: Evaluation framework and case study. *Int. J. Digit. Libr.*, vol. 11, n° 2, p. 91–109.

Saracevic T. (1976). Relevance: A review of the literature and a framework for thinking on the notion in information science. In *Advances in librarianship*, p. 79–138. Academic Press.

Saracevic T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the American Society for Information Science*, vol. 58, n° 13, p. 2126–2144.

Sieg A., Mobasher B., Burke R. (2007). Web search personalization with ontological user profiles. In *In proceedings of the sixteenth acm conference on conference on information and knowledge management*, p. 525–534. New York, NY, USA, ACM.

Soboroff I., Ounis I., Macdonald C., Lin J. (2012). Overview of the TREC-2012 microblog track. In *Proceedings of the 21th* T*ext* RE*trieval* C*onference (*TREC *2012).* National Institute of Standards and Technology (NIST).

Taylor A. R., Cool C., Belkin N. J., Amadio W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing and Management*, vol. 43, n° 4, p. 1071–1084.

Vallet D., Castells P. (2012). Personalized diversification of search results. In *Proceedings of the 35th annual international* ACM SIGIR *conference on research and development in information retrieval*, p. 841-850. ACM.