Toward a Personalized Approach for Combining Document Relevance Estimates

Bilel Moulahi[1,2], Lynda Tamine[1], and Sadok Ben Yahia[2]

[1] IRIT Laboratory - 118 route de Narbonne, 31062 Toulouse Cedex 9, France
{moulahi,tamine}@irit.fr
[2] Faculty of Science of Tunisia, LIPAH, 2092 Tunis, Tunisa
sadok.benyahia@fst.rnu.tn

Abstract. A large body of work in the information retrieval area has highlighted that relevance is a complex and a challenging concept. The underlying complexity appears mainly from the fact that relevance is estimated by considering multiple dimensions and that most of them are subjective since they are user-dependent. While the most used dimension is topicality, recent works risen particularly from personalized information retrieval have shown that personal preferences and contextual factors such as interests, location and task peculiarities have to be jointly considered in order to enhance the computation of document relevance. To answer this challenge, the commonly used approaches are based on linear combination schemes that rely basically on the non-realistic independency property of the relevance dimensions. In this paper, we propose a novel fuzzy-based document relevance aggregation operator able to capture the user's importance of relevance dimensions as well as information about their interaction. Our approach is empirically evaluated and relies on the standard TREC[3] contextual suggestion dataset involving 635 users and 50 contexts. The results highlight that accounting jointly for individual differences toward relevance dimension importance as well as their interaction introduces a significant improvement in the retrieval performance.

Keywords: Relevance, Aggregation, Personalization, Choquet Integral, Fuzzy Measure

1 Introduction

Relevance has long been already a complex subject and a challenge which has received a steady attention in information retrieval (IR) studies over the last two decades [1][16]. In fact, an extensive body of works in IR have attempted to revisit this concept which yields to a shift from topical to multidimensional relevance, involving other document relevance criteria coming mostly from the user's perspective such as cognitive, situational and affective relevance. In practice, the key

[3] Text REtrieval Conference

problem is how to design a document relevance scoring model able to involve individual relevance estimates linked to both user-dependent and user-independent relevance criteria and consider their overall interaction. This problem is faced in many IR settings including for instance personalized IR [17,6,5] mobile IR [10], social IR [15] and geographic IR [14]. For instance, in a social search activity involving Twitter, the task is driven by a variety of criteria such as *authority*, *topicality* and *recency* of *tweets* [15]. For the sake of addressing this challenge, previous approaches are mostly based on classical aggregation functions such as weighted means or linear combination schemes in the form of products and sums. However, these aggregation operators assume that relevance dimensions act independently [10] whereas other works have shown that they are not independent of each other and they interact in relevance judgments [18][2][16]. Although advanced aggregation operators were recently proposed [4][9][8], we are aware of only a few works that considered specifically the aggregation of relevance estimates in IR [8]. Most of state-of-the-art approaches tackle the IR problem without exploring the relevance dimension level within the IR task at hand and thus they do not consider the aggregation problem as a core in the ranking process. Aggregating methods proposed in previous work are generally built up considering the task relevance dimensions, ignoring the differences in the user's personal ratings of the different dimensions; For instance, in a Tweet search task, a user may prefer results that are relevant *w.r.t* both relevance criteria: recency and topicality or only authority and topicality. Thus, we need the aggregating operator to be able to offer insight to humans about why some relevance criteria were weighted more highly than other ones and to personalize the majority preference regarding the IR task specificity as well as the user preferences.

In this paper, we assume a more general scenario where different dependent or independent relevance dimensions are considered within a document retrieval task. More specifically, we address the following research questions:

RQ1. How to aggregate several interacting relevance criteria considering the user's IR task at hand?

RQ2. How to personalize relevance criteria importance regarding the user preferences in order to tailor the search results for each individual user?

Our specific contribution in this paper includes:

- A novel personalized multi-criteria aggregation approach for document relevance estimation: the core of the approach is based on the Choquet Integral [3][11], a fuzzy operator that allows (1) computing an aggregated document relevance score; (2) considering the interaction between relevance criteria; (3) personalizing the relevance score considering the user's importance rating of each relevance criterion.
- A large-scale experimental evaluation using a standard evaluation benchmark, namely the TREC contextual suggestion IR task that shows significant improvements in retrieval effectiveness from using our relevance aggregation operator.

The remainder of the paper is organized as follows. In Section 2, we briefly survey related work to put our contribution in context. Section 3 describes our multidimensional relevance aggregation operator. In Sections 4 and 5, we describe the experimental setup and then present the experiments and discuss the obtained results. Section 6 concludes the paper and outlines future work.

2 Related Work

The relevance concept has gathered a great attention in IR during the last decade [1][16]. The main outcome concerns the multidimensional nature of user relevance assessments that treats the IR process from a user-centered cognitive approach. Although there is no wide consensus, at a general level, this includes mainly content, object, validity, situational, affective and belief dimensions [16]. Each dimension refers to a group of criteria considered by the users to make relevance inferences. Besides, the main finding is that relevance dimensions are not independent of each other and generally those related to content, which include topical relevance, are rated as the highest ones in importance, but interact with other dimensions [16][8]. Considering this finding, several works in many recent IR tasks such as mobile IR [10], social IR [15] and personalized IR [6,17], attempt to go beyond the classical content dimension to cover as much as possible the dimensions related to user's context such as location and interests. However, the proposed approaches tackled the problem of relevance aggregation using simple linear combination strategies relying basically on the unrealistic assumption of both relevance dimension independency and additivity. Despite the fact that recent research has continued to exploit the relevance concept as a multi-faceted one, only a few have investigated how to accurately combine the individual document relevance estimates or scores related to the different relevance dimensions, regarding a given user and IR task [4][9][8]. Celia et *al.* [4] proposed a multidimensional representation of relevance and made use of 4 criteria: aboutness, coverage, appropriateness, and reliability through a general prioritized aggregation scheme involving two operators namely, "And" and "Scoring". These aggregation operators model a priority order over the set of relevance criteria which makes the weights associated to each criterion dependent upon the satisfaction of the higher preferred criterion. Gerani et *al.* [9] have proposed a multi-criteria aggregation model allowing to generate a global score that does not necessarily require the comparability of the combinable individual scores. The authors rely on the Alternating Conditional Expectation Algorithm and the BoxCox model to analyze the incomparability problem and perform a score transformation whenever necessary. More recently, Eickhoff et *al.* [8] introduced a statistical framework based on *copulas* to address the multidimensional relevance assessment and showed its performance in modeling complex dependencies between correlated relevance criteria.

From another side, learning to rank methods have been widely used in IR to combine multiple evidence with the goal of improving the overall search result quality [13]. Given a training set of queries and the associated ground truth containing document labels (relevant, irrelevant), the objective is to optimize a loss function that maps the document feature-based vector to the most accurate ranking score. However, these methods tend to offer only limited insight on how to consider importance and interaction between groups of features mapped to different relevance dimensions [8]. By contrast, we propose to investigate the combination of general level relevance dimensions using a fuzzy-based aggregation operator addressing: (1) the interaction between criteria through the Choquet integral, (2) the personalization of user's preferences regarding each relevance dimension.

3 Combining Relevance Estimates With the Choquet Integral

3.1 Viewing Relevance Aggregation by Means of the Choquet Integral

We address here the multidimensional relevance aggregation problem as a multi-criteria decision making (MCDM) problem. In fact, the difficulty in the aggregation problem is twofold: *(i) Criteria importance estimation*: correctly identifying which individual criterion and/or subset of criteria need to be enhanced *vs.* weakened regarding the IR task at hand and the user's preferences on the relevance criteria; *(ii) aggregation*: accurately combining the relevance criteria by taking into account their dependency.

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ be a set of documents, $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ a set of relevance criteria and q a given query. The task of combining performance scores denoted by $RSV^u_{c_i}(q,d_j)$, of document $d_j \in \mathcal{D}$, obtained *w.r.t* each relevance criterion $c_i \in \mathcal{C}$, is called *aggregation*. The function $\mathcal{F}$ that computes the personalized relevance score of document d_j in response to query q, considering user u, has the following general form:

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

Where $RSV^u_{c_i}(q,d_j)$ is the performance score of d_j *w.r.t* an individual criterion c_i, considering user u.

In the sequel, we rely on the Choquet operator as a multidimensional relevance aggregation. This mathematical function is built on a fuzzy measure (or *capacity*) μ, defined below.

Definition 1. *Let $I_\mathcal{C}$ be the set of all possible subsets of criteria from $\mathcal{C}$. A fuzzy measure is a normalized monotone function μ from $I_\mathcal{C}$ to $[0 \ldots 1]$ such that: $\forall I_{C_1}, I_{C_2} \in I_\mathcal{C}$, if $(I_{C_1} \subseteq I_{C_2})$ then $\mu(I_{C_1}) \leq \mu(I_{C_2})$, with $\mu(I_\varnothing) = 0$ and $\mu(I_\mathcal{C}) = 1$.*

For the sake of notational simplicity, $\mu(I_{C_i})$ will be denoted by μ_{C_i}. The value of μ_{C_1} can be interpreted as the importance degree of the interaction between the criteria involved in the subset C_1. The personalized Choquet integral based-relevance aggregation function is defined as follows:

Definition 2. *$RSV^u_\mathcal{C}(q,d_j)$ is the d_j document personalized relevance score for user u w.r.t the set of relevance criteria $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ defined as follows:*

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

Where Ch_μ is the Choquet aggregation function, $rsv^u_{(i)j}$ is the i^{th} element of the permutation of $RSV(q,d_j)$ on criterion c_i, such that $(0 \le rsv^u_{(1)j} \le ... \le rsv^u_{(N)j})$, $\mu^u_{\{c_i,...,c_N\}}$ is the importance degree of the set of criteria $\{c_i,...,c_N\}$ for user u. In this way, we are able to automatically adjust the ranking model's parameters for each user and make results dependent on its preferences over the considered criteria. Note that if μ is an additive measure, the Choquet integral corresponds to the weighted mean. Otherwise, it requires fewer than 2^N capacity measures in the case where the fuzzy measure is k−order additive, *i.e.,* $\mu_A = 0$ for all criteria subsets $A \subseteq \mathcal{C}$ with $|A| > k$. From a theoretical perspective, the Choquet operator exhibits a number of properties that appear to be appealing from an IR point of view; since it is built on the concept of fuzzy measures, it allows modeling flexible interactions and considering complex dependencies among criteria [12]. To facilitate the task of interpreting the Choquet integral behavior, we exploit two parameters namely, the "importance indice" and the "interaction indice" [12] that offer readable interpretations and qualitative understanding of the resulting aggregation model. While the former assesses the average contribution that a criterion (c_i) brings to all possible combinations of criteria, the latter gives information on the phenomena of dependency existing among the criteria. Indeed, this is a key point of the Choquet operator, as it may give insight to humans about why some criteria were weighted highly (resp. low) for relevance or to see if the criteria are really correlated. For further details on the computation of these indices, the reader can refer to the original paper [12].

3.2 Training the Fuzzy Measures Within an IR Task

Notation	**Description**
Q^u_{learn}	The set of queries used to train the capacity values belonging to user u
N	Number of relevance criteria
$\mathcal{D}$	The document collection
K	Number of top retrieved documents for each query used for learning
$\gamma^{i,r}$	List of ranked documents in response to query q_r *w.r.t* a capacity combination $\mu^{(i)}$. Let $P@X(\gamma^{r,i})$ be the $P@X$ of $\gamma^{r,i}$ and $AVP@X(\gamma^i)$ be its $P@X$ average over all queries $\in Q_{learn}$ *w.r.t* $\mu^{(i)}$
I_{Cr}	Subset of all possible criteria from Cr
$\mathcal{S}_\mu$	Set of the experimented capacity combinations values. Each combination $\mu^{(i)} \in \mathcal{S}_\mu$ contains the capacities values of all the set and subsets of criteria

Table 1: A summary of notations used within Algorithm 1.

The objective of the training step is to optimize the fuzzy measures *w.r.t* a target IR measure (e.g. $P@X$) by identifying the values of the Choquet capacities allowing to personalize the search results toward a particular user considering his individual preferences over the relevance criteria.

Considering a user, the typical training data required for learning the Choquet

Algorithm 1 Training the Fuzzy Measures

```
Data: $Q^u_{learn}$, $N$, $K$.
Result: Optimal capacity combination $\mu^{(**)}$.
   Step 1: Initialize the capacity values
   $m \leftarrow (1-N)\times N$;
   For $i = 1$ to $m$ {Capacity combinations identification} do
      $\mu^{(i)} = (\bigcup_{j:1..N}\{\mu_{c_j}\})\bigcup(\bigcup_{Cr\in\mathcal{C},|Cr|>1}\{\mu_{I_{Cr}}\})$; $\mu_{I_{Cr}} = \sum_{c_i\in Cr,|c_i=1|}\mu_{c_i}$
   End for
   If $N \geq 4$ {Assume 2-additivity} then
      For each $I_{Cr}\in\mu^{(i)}$ such that $|Cr|>2$ do
         $\mu_{I_{Cr}} = 0$
      End for
   End if
   $\mathcal{S}_\mu = \bigcup_{i:1..m}\{\mu^{(i)}\}$
   For each $\mu^{(i)}\in\mathcal{S}_\mu$ {Capacity tuning} do
      Compute $AVP@X(\gamma^i)$
   End for
   $Cmax = \text{Argmax}_{1\ldots|\mathcal{S}_\mu|}\ (AVP@X(\gamma^i))$; $\mu^{(*)} = \mu^{(cmax)}$
   Step 2: Optimize the capacity values
   $D^u_{learn} = \emptyset$
   For $r = 1$ to $|Q^u_{learn}|$ {Interpolate the global scores} do
      $D^u_{learn} = D^u_{learn}\cup\gamma^{*,r}$
      For $j = 1$ to $K$ do
         $RSV^{int}_{\mathcal{C}}(q_r,d_j) = \text{Max}_{1\ldots d'_j\in\gamma^{*,r},d'_j\succ_C d_j}\ (RSV^u_{\mathcal{C}}(q_r,d'_j))$ ; $\gamma^{*,r} = \gamma^{*,r}\setminus\{d_j\}$
      End for
   End for
   {Least-square based optimization}
   Repeat
      $\mathcal{F}_{LS}(\mu) = \sum_{d_j\in D^u_{learn}}[Ch_\mu(RSV^u_{c_1}(d_j),\ldots,RSV^u_{c_N}(d_j)) - RSV^{int}_{\mathcal{C}}(d_j)]^2$
   Until convergence
   Return the outcome $\mu^{(**)}$
```

fuzzy measures includes a set of training queries and for each query, a list of ranked documents represented by pre-computed vectors containing performance scores; each document is annotated with a rank label (*e.g.,* relevant or irrelevant). The adopted methodology for that purpose is detailed in Algorithm 1. Table 1 describes the notations used within the Algorithm. The latter runs in two main steps:

1- Setting the initial values of the capacity combinations. For simplicity, we call capacity combination $\mu^{(.)}$ the set of capacity values assigned to each criterion and subset of criteria. For instance, in the case of three relevance criteria, a possible capacity combination involves $(\{\mu_{c_1};\mu_{c_2};\mu_{c_3};\mu_{c_1,c_2};\mu_{c_1,c_3};\ \mu_{c_2,c_3}\})$. In order to tune these values, we make use of a target IR measure such as $P@X$ over the training queries Q^u_{learn}. The tuning is conceivable since there is generally only a few relevance dimensions [16]. However, when the number of criteria is strictly higher than 3, we can avoid the tuning complexity by relying on subfamilies of capacities namely 2-additive measures [12], requiring less coefficients to be defined and assuming that there is no interaction among subsets of more than 2 criteria. This assumption is made only in the initialization step.

2- Optimizing the capacity values. Starting from the initial capacity combination $\mu^{(*)}$ obtained in the previous step, we pull the top K documents returned by each training query $q \in Q^u_{learn}$. The scores of these documents, referred to as D^u_{learn}, are first interpolated to boil down the non relevant ones. After we obtain the desired overall relevance scores $RSV^{int}_{\mathcal{C}}(q, d_j)$ for each document $d_j \in D^u_{learn}$, and since we are given the labels $RSV^u_{ci}(q, d_j)$, we proceed to the application of the Least-squares based optimization, which is a generalization of classical multiple linear regression.

4 Experimental Design

Our experimental evaluation is based on TREC[4] 2013 Contextual Suggestion Track [7]. This IR track examines search techniques that aim to answer complex information needs that are highly dependent on context and user interests. Roughly speaking, given a user, the track focuses on travel suggestions (e.g., attraction places) based on two dependent relevance criteria: (1) users' interests which consist of his personal preferences and past history; (2) his geographical location. This section describes the used data sets and the evaluation protocol.

4.1 Datasets

We use the TREC 2013 Contextual suggestion data set [7] which includes the following characteristics:

- **Users**: The total number of users is 635. Each user is represented by a profile reflecting his preferences for places in a list of 50 example suggestions. An example suggestion is an attraction place expected to be interesting for the user. The preferences, given on a 5-point scale, are attributed for each place description including a title, a brief narrative description and a URL website. Positive preferences are those having a relevance judgment degree of about 3 or 4 *w.r.t* the above features. Ratings of 0 and 1 on example suggestions are viewed as non relevant and those of 2 are considered as neutral.
- **Contexts and queries**: A list of 50 contexts is provided, where each context corresponds to a particular city location, described with longitude and latitude parameters. Given a pair of user and context which represents a query, the aim of the task is to provide a list of 50 ranked suggestions satisfying as much as possible the considered relevance criteria.
- **Document collection**: To fetch for the candidate suggestion places, we crawl the open web through the Google Place API[5]. As for most of the

[4] `http://trec.nist.gov`

[5] `https://developers.google.com/places`

TREC Contextual Suggestion track participants [7], we start by querying the Google Place API with the appropriate queries corresponding to every context based on the location. This API returns up to 60 suggestions, thus, we search again with different parameters, like place types that are relevant to the track. Approximately 157 resulting candidate suggestions are crawled on average per context and 3925 suggestions in total. To obtain the document scores *w.r.t* the geolocalisation criterion, we compute the distance between the retrieved places and the context, whereas we exploit the cosine similarity between the candidate suggestions description and the user profile to compute the user interest score. User profiles are represented by vectors of terms constructed from his personal preferences on the example suggestions. The description of a place is the result snippet returned by the search engine Google[6] when the URL of the place is issued as a query.

- **Relevance assessments**: Relevance assessments of this task are made by both users and NIST assessors [7]. The user corresponding to each profile, judged suggestions in the same way as examples, assigning a rating of 0−4 for each title/description and URL, whereas NIST assessors judged suggestions in term of geographical appropriateness on a 3-point scale (2, 1 and 0). A suggestion is relevant if it has a relevance degree of about 3 or 4 *w.r.t* user interests (profile) and a rating of about 1 or 2 for geolocalisation criterion. Those relevance assessments will constitute our ground truth used for both training and testing, in the remainder.

4.2 Evaluation Protocol

Similar to a previous work [19], we adopt a fully-automated methodology through a 2-fold cross validation in order to train the users' capacity values and test the aggregation model effectiveness. For this purpose, we randomly split the 50 contexts into two equivalent sets, noted Q^u_{learn} and Q^u_{test} used respectively for training and testing. In addition, the set of contexts is randomly split into two different other training and testing sets in another round in order to avoid the learning overfitting. The objective of training is to learn the capacities ($\mu^u_{\{user\_interest\}}$, $\mu^u_{\{geolocalisation\}}$) viewed as the relevance criteria importance. We first start by an initial fuzzy measure giving the same importance weight for both relevance criteria and issued the P@5 measure for all contexts from Q^u_{learn}. Then, using the ground truth provided within the TREC 2013 Contextual suggestion track, and based on Algorithm 1, we learn for each user the personal preferences for both criteria: user interest and geographical location. We use the same data to train the best users' criteria priority scenarios for both prioritized aggregation operators baselines detailed in section 5.2. Finally, we use Q^u_{test} set to test the effectiveness of our approach based on the remaining queries, relying on the official measure of the track, namely the precision at rank 5 (P@5). This latter is a high precision measure computing the proportion of relevant suggestions ranked at the top 5 of the output list of suggestions.

[6] `https://www.google.com`

5 Results and Discussion

5.1 Analyzing the Users' Relevance Criteria Importance

Here, we aim to analyze the learned capacity values issued from Algorithm 1, reflecting the users' relevance criteria importance degrees ($\mu^u_{\{user\_interest\}}$, $\mu^u_{\{geolocalisation\}}$). First we analyze the intrinsic importance of each criterion independently of each other. Figure 1 depicts the analysis of the variation of the capacity values on the relevance criteria over the learning set of contexts for each user. The x-axis represents each user (id's from 35 to 669) and the y-axis represents the capacity values of criterion user interest or geolocalisation for each user. Figure 1 shows that the user interest criterion is accorded a higher capacity than the location for all users. For instance, user 285 has a capacity value of about 0.23 for the first criteria and a measure of about 0.76 for the second one. This is natural given that users generally seek first for places that match their personal preferences even if they are not geographically relevant. However, we can see from Figure 1 that the capacity values distribution is far from being the same for all users and reveals values from 0.09 to 0.414 for geolocalisation and from 0.585 to 0.909 for user interest.

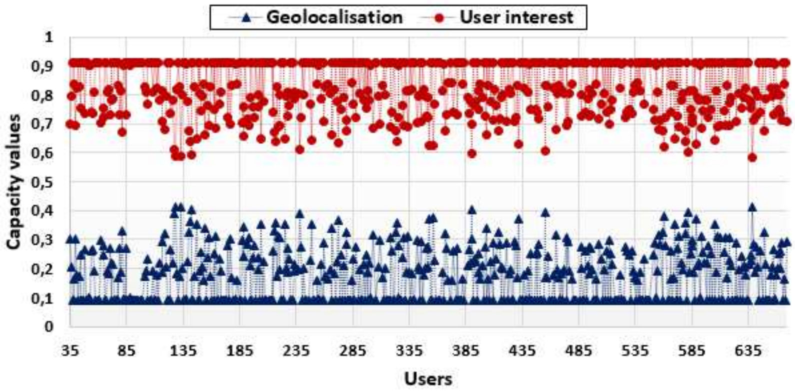

Fig. 1: Capacity values for TREC 2013 Contextual suggestion Track users.

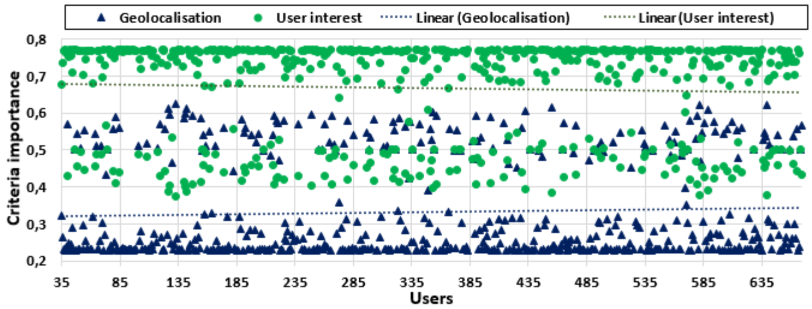

Fig. 2: Relevance criteria importance for Contextual suggestion Track users.

Second, we analyze the dependency between criteria through the computation of the interaction indice [12]. The obtained values are found to be positive and vary between 0.28 and 0.99. The average value for all users is 0.56 which implies a positive interaction between the considered relevance criteria when they are combined together. To get a better understanding of this phenomenon, we plot in Figure 2, the importance indice values [12] reflecting the overall relevance criteria importance degree for each user. Unlike Figure 1, Figure 2 highlights the average importance of each relevance criterion when it is mainly combined with the other one. One can observe from Figure 2 that users' preferences among the criteria are totally different. The smoothing of the obtained importance values *w.r.t* both relevance criteria values gives the two linear curves with quite constant values which bear out the results obtained in Figure 1. The user interest relevance criterion is still given a quite high importance for the majority of users, but we can also see interestingly in the middle of the figure (values between 0.4 and 0.7), that some users have higher importance degree for the geolocalisation criterion and vice versa. We argue that this difference in preferring some criteria than others unveil the need to personalize the users' relevance importance degrees.

5.2 Evaluating the Retrieval Effectiveness

The objective here is to evaluate the effectiveness of our approach based on aggregation and personalization properties. For this aim, we compare the retrieval results using the testing set to the following baselines: the weighted arithmetic mean (WAM), the Scoring and And aggregation operators [4] widely used in most approaches involving combination of relevance estimates. Note that we ran preliminary series of experiments through cross validation to find the best prioritized scenario for the Scoring and And aggregation operators for each user, within the same learning set used to find the Choquet capacity values. Similar to the results obtained through the importance indice analysis, we also found that the best scenario is that giving a priority to the user interest relevance criterion, except that a prioritized operator is not able to quantify the importance degree of criteria. Furthermore, in order to show the personalization effectiveness, we compare our Choquet PERsonalised aggregation operator, denoted ChoPER with the classical non personalized aggregation Choquet operator. This latter is performed using Algorithm 1 once (not for each user), involving a learning method of criteria capacity values regardless of the users. This gives rise to a value of about 0.86 for the user interest criterion and a value of about 0.14 for the geolocalisation relevance criterion. Precision measures are averaged through the testing rounds, for all the different baselines over all the testing queries.
Table 2 shows the retrieval performances obtained using our operator, namely CHoPER, in comparison with the baselines described above. From Table 2, we can see that the performance of ChoPER is significantly higher than the overall baselines for the official measure P@5 but less important for the other measures. Interestingly, we also figure out that ChoPER performance is stable *w.r.t* all the evaluated measures. Compared to the best baseline And, the performance improvements for ChoPER reach 10.11% *w.r.t* official measure P@5.

Those results are likely due to the fact that the And aggregation operator is mainly based on the Min operator, which could penalize places highly satisfied by the least important criterion. The obtained difference of performance, in favor of ChoPER, is explained by the consideration of the different preference levels toward the relevance criteria and the interaction that exist between both of them. In terms of personalization, the retrieval effectiveness results *w.r.t* precisions at (5, 10, 20 and 30) between the classical Choquet operator and ChoPER show that the latter performs significantly for all the precision measures. The best improvement for Choquet is up to 9.29 for P@5 measure. These results confirm those obtained in the capacity training phase (Cf. Setion 5.1) where we show that the importance degree of criteria depends on the users' preferences and are not the same for all of them.

	Precision				
Operator	**P@5**	**P@10**	**P@20**	**P@30**	**% change**
WAM	0.1046	0.1255	0.1174	0.1093	**+13.98%** §
AND	0.1093	0.1267	0.1197	0.1104	**+10.11%** §
SCORING	0.1069	0.1267	0.1186	0.1108	**+12.08%** §
ChOQUET	0.1103	0.1269	0.1203	0.1116	**+9.29%** §
CHoPER	**0.1216**	**0.1279**§	**0.1203**	**0.1131**	–
	+10.11%§	**+0.93%**	**+0.49%**	**+2.38%**§	

Table 2: Comparative evaluation of retrieval effectiveness. % change indicates the CHoPER improvements in terms of *P*@5. The last row shows the performance improvements against the best aggregation baseline, And. The symbols § denotes the student test significance: "§": $t < 0.05$.

6 Conclusion and Future Work

We presented a novel general multi-criteria framework for multidimensional relevance aggregation. Our approach relies on a fuzzy method based on the well studied and theoretically justified Choquet mathematical operator. The proposed operator supports the observation that relevance dimensions, measurable through criteria, may interact and have different weights (importance) according to the task at hand. The resulting model criteria behavior regarding the different user preferences is analyzed with readable interpretations through the importance and interaction indices. Empirical evaluation using a standard appropriate dataset shows that our approach is effective. In future, we plan to investigate how to extend the personalization toward groups of users rather than individual users. This would offer opportunities to learning relevance criteria importance from similar users and thus, tackling the lack of training user' examples of preferences.

References

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

2. B. Carterette, N. Kumar, A. Rao, and D. Zhu. Simple rank-based filtering for microblog retrieval: Implications for evaluation and test collections. In *Proceedings of the 20th Text REtrieval Conference*, 2011.
3. G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953.
4. C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Inf. Process. Manage.*, 48(2):340–357, 2012.
5. M. Daoud, L. Tamine, and M. Boughanem. A personalized graph-based document ranking model using a semantic user profile. In *UMAP*, pages 171–182, 2010.
6. M. Daoud, L. Tamine, M. Boughanem, and B. Chebaro. Learning implicit user interests using ontology and search history for personalization. In *Proceedings of the 2007 International Conference on Web Information Systems Engineering*, WISE'07, pages 325–336, Berlin, Heidelberg, 2007. Springer-Verlag.
7. A. Dean-Hall, C. Clarke, J. Kamps, P. Thomas, N. Simone, and E. Voorhes. Overview of the trec 2013 contextual suggestion track. In *Text REtrieval Conference (TREC)*. National Institute of Standards and Technology (NIST), 2013.
8. C. Eickhoff, A. P. de Vries, and K. Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 2013.
9. S. Gerani, C. Zhai, and F. Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, pages 256–267. Springer-Verlag, 2012.
10. A. Göker and H. Myrhaug. Evaluation of a mobile information system in context. *Inf. Process. Manage.*, 44(1):39–65, 2008.
11. M. Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69(3):279–298, 1995.
12. M. Grabisch, T. Murofushi, M. Sugeno, and J. Kacprzyk. *Fuzzy Measures and Integrals. Theory and Applications.* Physica Verlag, Berlin, 2000.
13. T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
14. F. Mata and C. Claramunt. Geost: geographic, thematic and temporal information retrieval from heterogeneous web data sources. In *Proceedings of the 10th international conference on Web and wireless geographical information systems*, pages 5–20, Berlin, Heidelberg, 2011. Springer-Verlag.
15. R. Nagmoti, A. Teredesai, and M. De Cock. Ranking approaches for microblog search. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 153–157, Washington, DC, USA, 2010. IEEE Computer Society.
16. T. Saracevic. 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*, 58(13):2126–2144, 2007.
17. A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM Conference on Information and Knowledge Management*, CIKM '07, pages 525–534, New York, NY, USA, 2007.
18. S. R. Wolfe and Y. Zhang. Interaction and personalization of criteria in recommender systems. In *Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 183–194, Berlin, Heidelberg, 2010. Springer-Verlag.
19. P. Yang and H. Fang. Opinion-based user profile modeling for contextual suggestions. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 80–18. ACM, 2013.