Manuscript JASIST-2013-XX-XXXX

iAGGREGATOR: Multidimensional Relevance Aggregation Based on a Fuzzy Operator

Bilel Moulahi · Lynda Tamine · Sadok Ben Yahia

Received: Mars 22, 2013 / Revised: July 31, 2013 / Accepted: July 31, 2013

Abstract Recently, an increasing number of IR studies has triggered a resurgence of interest in redefining the algorithmic estimation of relevance, which implies a shift from topical to multidimensional relevance assessment. A key underlying aspect which emerged when addressing this fundamental concept in information sciences, is the aggregation of the relevance assessments related to each of the considered dimensions. In order to tackle this issue, the most commonly adopted forms of aggregation are based on classical weighted means and linear combination schemes. Although some initiatives were recently proposed, none was concerned with considering

B. Moulahi

Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France,

Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH, 2092, Tunis, Tunisie, E-mail: bilel.moulahi@irit.fr.

L. Tamine

Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France, E-mail: lynda.tamine@irit.fr.

S. Ben Yahia

Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH, 2092, Tunis, Tunisie,

Institut Mines-TELECOM, TELECOM SudParis, UMR CNRS Samovar, 91011 Evry Cedex, France, E-mail: sadok.benyahia@fst.rnu.tn.

the inherent dependencies and interactions existing among the relevance criteria as it is the case in many real-life applications. In this paper, we present a new fuzzy based operator, called iAggregator, for multidimensional relevance aggregation. Its main originality, beyond its capability of modelling interactions between the different relevance criteria, stands in its generalization of many classical aggregation functions. To validate our proposal, we apply our operator within a *tweet* search task. Carried out experiments using a standard benchmark, namely TREC Microblog[1], emphasize the relevance of our contribution and its superiority to traditional aggregation schemes such as linear combination mechanisms, weighted means and the Owa operator. In addition, it outperforms state-of-the-art aggregation operators such as the Scoring and the And prioritized operators as well as some representative learning to rank algorithms.

Keywords Information retrieval, multiple criteria, multidimensional relevance, aggregation, fuzzy measure, Choquet integral, microblogging, *tweet* search.

1 Introduction

Multi-criteria aggregation is an issue that has been thoroughly addressed in social choice (Condorcet, 1785; Fishburn, 1972; Arrow, 1974), engineering design (Neumann & Morgenstern, 1953; Keeney & Raiffa, 1993) and computer vision applications (Dubois & Prade, 2004; Torra, 2005), to cite but a few. The multi-criteria aggregation arises when for a given task, there are several alternatives that have to be ordered with respect to different criteria, and we are faced with the problem of combining them in order to figure out a ranking over the set of alternatives. The need of aggregating several inputs into a single representative output allowed successful applications of aggregation functions to fields as diverse as information retrieval (IR) (Farah & Vanderpooten, 2007) to Multiple Criteria Decision Analysis (MCDA) (Steuer, 1986; Grabisch, Kojadinovic, & Meyer, 2008), data fusion (Vogt & Cottrell, 1999; Ah-Pine, 2008) to database retrieval (Le Calvé & Savoy, 2000). In this paper, we are more particularly interested in addressing this issue within the IR field. Actually, as ranking

[1] https://sites.google.com/site/microblogtrack/

and relevance are the main cores of IR systems (Hawking, Craswell, Bailey, & Griffiths, 2001), a great deal of research has triggered a resurgence of interest in revisiting the concept of relevance considering several criteria. In fact, many of the proposed state-of-the-art early IR models rank documents by computing single scores separately with respect to one single objective criterion, rather than considering other relevance dimensions encompassing contextual features with respect to users or documents (Borlund, 2003). This most commonly used criterion, that in some applications even becomes a synonym of relevance, is the topical one, also namely *subject* relevance (Vickery, 1959). It expresses the document's topical overlap with the user's information need, which is basically and solely based on the topicality matching. However, several studies showed that relevance is a *multidimensional* concept (Borlund, 2003; A. R. Taylor, Cool, Belkin, & Amadio, 2007; Saracevic, 2007; A. R. Taylor, 2012) that goes beyond simple topical relevance. Authors in (A. R. Taylor et al., 2007) conducted an experimental study and reported that: *"IR systems must provide a richer set of search criteria beyond topicality"*.

Furthermore, this multidimensional property is witnessed in many IR applications such as mobile IR (Cong, Jensen, & Wu, 2009; Bouidghaghen et al., 2011; Church & Smyth, 2008; Göker & Myrhaug, 2008), social IR (Becker, Naaman, & Gravano, 2011; Ounis, Macdonald, & Soboroff, 2011; Chen et al., 2012; Berardi, Esuli, Marcheggiani, & Sebastiani, 2011; Damak et al., 2011) and personalized IR (Gauch, Chaffee, & Pretschner, 2003; Costa Pereira, Dragoni, & Pasi, 2009; Daoud, Tamine, & Boughanem, 2010; Costa Pereira, Dragoni, & Pasi, 2012; F. Liu, Yu, & Meng, 2004; Ma, Pant, & Sheng, 2007). In a mobile IR setting, users usually search for information while moving. The goal of any IR system addressing this issue is to tailor the search results to the user's needs according to several contextual criteria such as *location*, *time* and user's interests features, in order to deliver the information that better addresses the user's situation in spatio-temporal applications. While personalized IR approaches consider *user preferences* as the main relevance criteria, social IR ones consider the user's community rather than just the individual as the basic clue for relevance computation. The latter problem is faced in many settings by involving some significant features regarding the search task at hand. For instance, the *tweet* search task is driven by a variety of criteria such as *authority*, *topicality* and *recency* of *tweets* (Chen et al., 2012; Duan, Jiang, Qin, Zhou, & Shum, 2010).

Thus, the main challenge that actually arises, is to find the suitable aggregation scheme to combine the single scores related to single criteria evaluations into a global score of documents representing the overall relevance estimate. We notice that despite the overwhelming number of publications that highlighted the multidimensional nature of relevance and the wide range of aggregation operators that have been proposed in the literature, the multidimensional relevance aggregation problem in IR has not sufficiently grasped the attention it deserves (Costa Pereira et al., 2009, 2012). The most widely used forms of aggregation are the weighted sum and its variations as well as linear combination mechanisms due to their simplicity (Vogt & Cottrell, 1999; Larkey, Connell, & Callan, 2000; Si & Callan, 2002; Damak et al., 2011; Z. Wei et al., 2011). However, as stated in (Costa Pereira et al., 2009), the major inconvenience of these works is that the criteria are combined in a linear model independently of the user's preferences on the relevance dimensions. Furthermore, in addition to their insufficiency to model several user's preferences, these operators are not suitable for the aggregation of interacting criteria, since it requires them to act independently.

In this paper, we are concerned with the application of a more sophisticated operator, already of use in other fields, to handle the multidimensional relevance aggregation problem in IR. This operator named the Choquet integral (Choquet, 1953; Grabisch, 1995), is still a very successful paradigm in multi-criteria decision making problems (Grabisch & Labreuche, 2010). The Choquet integral generalizes many other aggregation operators (Grabisch, 1995) such as the weighted mean (WAM) and the ordered weighted averaging (OWA) operator (Yager, 1988). While the Choquet integral is not widely known in IR so far, the exploitation of this operator in combining multidimensional relevance estimates in the IR area is the first work into this insight. From a theoretical perspective, the Choquet operator exhibits a number of properties that appear to be appealing from an IR point of view. It allows for modeling interactions between several criteria, which is prominent among relevance criteria and are sometimes undesirable phenomena in some IR applications. Interestingly enough, the proposed aggregation model is general and it may be applied to any set of criteria. The main contributions of this paper are twofold:

1. We introduce a general multi-criteria aggregation approach, namely, iAggregator, based on a well studied and theoretically justified mathematically aggregation operator, for multidimensional relevance aggregation in the IR domain. Thus, we survey a problem that has not get enough attention in the literature. We particularly model the multi-criteria relevance aggregation within dependent and interacting criteria.
2. We apply and experiment iAggregator to evaluating multi-criteria relevance aggregation in a social IR setting. More particularly, on a *tweet* search task (Ounis et al., 2011; Ounis, Macdonald, & Soboroff, 2012), where the jointly considered criteria are *topicality*, *recency* and *authority*.

The remainder of the paper is organized as follows. Section 2 reviews related work on multidimensional relevance aggregation, gives an overview of the learning to rank problem for IR and emphasizes our motivations. We provide, later in Section 3, a critical overview of the aggregation problem in multi-criteria decision making area and specify the problem within an IR task. Our proposal for a multidimensional relevance estimation with the discrete Choquet integral is presented in Section 4. Section 5 describes the experimental setting within a *tweet* search task. In Section 6, we present and discuss the obtained results. Section 7 concludes the paper and outlines future work.

2 Multidimensional relevance aggregation in IR

In this section, we present a review of related work on multidimensional relevance followed by a synthesis of works dealing with aggregation operators used for that purpose as well as learning to rank approaches.

2.1 Relevance in IR: a multidimensional concept

As pointed out by many key papers in the literature (Saracevic, Rothenberg, & Stephan, 1974; Schamber, 1991; Barry, 1994; Mizzaro, 1998; Cosijn & Ingwersen, 2000; Saracevic, 2000), the relevance has been already a complex subject and a challenge which has received a steady attention in IR studies over the last two decades. While early research put the focus on the relevance concept from a topical perspective, more recent research paid attention on it from various points of view (Borlund, 2003; A. R. Taylor et al., 2007; A. R. Taylor, 2012) which implies a shift from topical relevance to multidimensional relevance. The great number of contributions devoted to analyzing the multidimensional concept of relevance has led to identifying many types and facets of relevance, such as cognitive and situational relevance, in addition to algorithmic and topical ones. Table 1 gives an overview of these studies.

Main references	**Studied relevance criteria**
(Cuadra & Katter, 1967; Rees & Schultz, 1967)	40 criteria including style and level of difficulty of the document.
(Cooper, 1973)	Novelty, informativeness, credibility, importance, clarity, positive/negative factors
(R. S. Taylor, 1986)	Ease of use, noise reduction, quality, adaptability, time saving, cost saving
(Schamber, 1991)	10 criteria (3 categories; *information, source, presentation*)
(Su, 1992, 1994)	20 measures (groups: *success,efficiency, utility, user satisfaction*)
(Barry, 1994)	24 criteria grouped into 7 broad groups
(Saracevic, 1996)	(Topical, algorithmic, cognitive, situational, motivational/affective) relevance
(Mizzaro, 1998)	Information resources, user problem, time, components
(Cosijn & Ingwersen, 2000)	Topical, cognitive/pertinence, situational, socio-cognitive
(Borlund, 2003)	Topical, cognitive, situational

Table 1: A Synthetic overview of empirical studies emphasizing the multidimensional aspect of relevance concept in IR.

The respective studies of (Cuadra & Katter, 1967) and (Rees & Schultz, 1967) investigated the factors that may affect relevance, and identified about 40 possible variables that could influence the users' relevance judgments. Cooper pointed out on its informal work in (Cooper, 1973), that many factual features based on documents' properties may be included. Cooper distinguishes between "logical relevance" or "topicality" (relevance concerning the topical component) and "utility" (relevance concerning the three components), among which are: accuracy, credibility, recency, etc., and assumes that these criteria could impact the relevance judgments. In the same context, (Barry, 1994) claimed that the relevance is a multidimensional concept and cannot be derived from a single relevance criterion. She performed an exploratory study in which she identified 23 categories of relevance. These categories embody numerous criteria that may be applied to documents' content as well as to any aspect of documents like contextual factors (*e.g.,* the user situation and environmental effects) or quality of the document source (*e.g.,* authority and reputation). Cosijn and Ingwersen (2000) developed a table of manifestations and attributes for relevance where manifestations consist in topical, cognitive, situational and socio-cognitive ones. In (Borlund, 2003), the authors emphasize 3 relevance dimensions: *topical*, *cognitive* and *situational*. More specifically, it has been shown that the "multidimensionality" of relevance can be viewed with reference to different conceptions of relevance such as "the classes, types, criteria, degrees, and levels of relevance". The author outlines the different conceptions of the "multidimensionality" of relevance as well as the inherent aspect of dynamic relevance. Accordingly, in (Saracevic, 2007), it has been demonstrated that "*topicality plays an important, but not at all an exclusive, role in relevance inferences by people. A number of other relevance clues or attributes, enter into relevance inferences"*, these criteria affect the user's perception of relevance and interact with topicality as judgments are made.

Roughly speaking, regarding the research focus of early studies on the use of relevance, we can distinguish between two main categories of approaches. In the first category (Vickery, 1959; Cooper, 1971; Harter, 1992), authors consider topicality as the basic part of relevance and assume that all the other criteria are topical-dependent. In contrast, other approaches in the same category, mainly involved in IR applications, adopt the idea that there are many different criteria beyond topicality, that may influence the user's perception of relevance. However, they didn't investigate the design of aggregation functions and so used basic aggregation operators such as the arithmetic mean and the weighted sum. Unlike previous cited studies, the second category of contributions (Costa Pereira et al., 2009, 2012; Gerani, Zhai, & Crestani, 2012) aims at designing general theoretical frameworks of relevance aggregation regardless of the application at hand. This line of research did not grasp the attention that it deserves, especially in the IR field. Our contribution attempts to fill in this gap, by the proposal of a general flexible aggregation mechanism based on the well studied and mathematically justified Choquet integral function.

2.2 Relevance aggregation in IR

In the following, we review the research contributions dealing with IR applications such as mobile IR, personalized IR, social IR and geographic IR that make use of aggregation operators to compute a global relevance score. In fact, most of the proposed approaches deal with classical aggregation mechanisms, without having a research focus on the modelling of general multi-criteria aggregation functions to combine all of the considered criteria. Second, we synthesize works that, in contrast, have specifically a research focus towards the design of appropriate combination operators, in order to support ranking functions in IR, regardless of any application.

2.2.1 Applying basic relevance aggregation operators in IR applications

The application of relevance aggregation is crucially important in many recent IR applications. It has been experienced without being the research focus in mobile IR (Cong et al., 2009; Church & Smyth, 2008; Göker & Myrhaug, 2008), personalized IR (Daoud et al., 2010; Gauch et al., 2003), social IR (Becker et al., 2011; Ounis et al., 2011; Chen et al., 2012; Berardi et al., 2011; Damak et al., 2011) and gographic IR (Mata & Claramunt, 2011; Kishida, 2010; Daoud & Huang, 2013). The approaches that have been proposed with this respect, are mostly based on linear combination mechanisms. Indeed, the main research subject of these works is the simple combination of individual relevance scores in one given IR setting. We outline in Table 2 a synthetic overview of the main IR tasks involving multidimensional relevance aggregation. For each of these tasks, we cite the main research contributions, we give the used relevance criteria and then mention the exploited aggregation operator.

IR task	Main references	Used relevance criteria	Aggregation operators
Mobile IR	(Göker & Myrhaug, 2008; Church & Smyth, 2008; Cong et al., 2009; Hattori et al., 2007; Cheverst et al., 2000; Schilit et al., 2003; Yau et al., 2003; Cantera et al., 2008)	*topicality, user interests, user's location, time, social* features	*Linear combination mechanism*
Personalized IR	(Gauch et al., 2003; Daoud et al., 2010; F. Liu et al., 2004; Ma et al., 2007; Sieg et al., 2007)	*Aboutness, coverage, appropriateness, reliability, user interests*	*Linear combination mechanism: summation of partial relevance scores, factor product*
Social IR	(Becker et al., 2011; Metzler & Cai, 2011; Damak et al., 2011; Berardi et al., 2011; Ben Jabeur et al., 2010; Chen et al., 2012; Smith et al., 2008; Leung et al., 2006)	*content* features, *Twitter* features, *author* features; *time*	*Linear combination mechanism: summation of partial relevance scores, factor product*
Geographic IR	(Mata & Claramunt, 2011; Kishida, 2010; Daoud & Huang, 2013)	*content, time, geographic location, proximity*	*Linear combination mechanism: summation of partial relevance scores*

Table 2: A synthetic overview of works involving relevance aggregation in IR tasks.

In mobile IR settings, Göker and Myrhaug (2008) use both time and location criteria through a linear combination operator to compute the global documents' scores. Cong et al. (2009) have proposed an IR model based on user's location and a topical relevance dimension in which the documents are ranked through a simple linear combination mechanism of both considered criteria.

In (Yau et al., 2003), the authors combine situation-based adaptation and profile-based personalization into the IR model. A situation is a set of past context attributes and/or actions such as: *location, time, light, device, etc.* A user profile includes a usage history and general interests that have been automatically learned using a modified Naive Bayesian classifier. (Cantera et al., 2008) propose to use the *MCI* model (Multiplicative Competitive Interaction) for combining *topical* scores of documents, the *geographic location* and the *user's interests* in a mobile context. The general expression of utility of a document in the *MCI* model is given by a linear combination of the individual scores. The considered relevance contextual criteria are mainly the *location*, the context of the used mobile device, combined with *text documents* scores.

In personalized IR settings, several works such as (Gauch et al., 2003; Sieg et al., 2007; Daoud et al., 2010) proposed a combination model of original scores and personalized scores of documents, computed according to their similarity to the user's profile represented through his interests. The exploited aggregation method is the linear combination of the considered criteria. More precisely, the authors compute the overall relevance score of a document as a linear combination of the personalized score obtained and the original one computed with respect to the topical relevance criterion.

In social IR settings, it is worth noting that a wide range of research papers have been recently proposed in the context of the TREC 2011 and 2012 Microblog Tracks. The majority of the proposed approaches in this area, are based on linear combination strategies of relevance criteria. Berardi et al. (2011) have focused on the problem of retrieval and ranking in *Twitter* and proposed an IR system called *CipCipPy* for that purpose. The authors explored the use of text quality ranking measures to filter out of vocabulary tweets, as well as the use of information contained in *hashtags* and linked content. The individual scores are then combined through a simple linear combination mechanism. Damak et al. (2011) proposed two tweet search models integrating several features. The first one is based on content features (*e.g., tweet popularity, tweet length*), *Twitter* features (*e.g., URL presence/frequency, hashtag*) and author features (*e.g., number of tweets/mentions*). For the computation of the final score involving these criteria, the authors adopted a linear combination strategy. Metzler and Cai (2011) proposed a learning to rank approach taking into account a textual similarity to the query, a time difference between a *tweet* and a query, as well as some *tweet* content features such as the *URL* presence, the *hashtag* existence, the *tweet length* and the percentage of words out of vocabulary.

Exploring the combination of *geospatial* and *temporal* criteria into geographic IR has been shown to have significant improvements in traditional search engines (Daoud & Huang, 2013; Mata & Claramunt, 2011; Kishida, 2010). For instance, in (Daoud & Huang, 2013), the authors propose a geotemporal retrieval strategy that models and exploits geotemporal context-dependent evidence extracted from pseudo-relevant feedback documents. The final score of the document is based on combining the *content-based* score, the *temporal score*, the *geographic* score, and the *proximity* score using a linear combination operator.

2.2.2 Designing specific relevance aggregation operators

To the best of our knowledge, despite the important attention paid to the multidimensional property of relevance, as highlighted above (see Section 2.1), only few recent works focused on the design of appropriate combination operators in order to support multidimensional relevance based ranking functions in IR. Among these studies, the recent work of (Costa Pereira et al., 2009, 2012), in which the authors proposed a multidimensional representation of relevance and suggested a prioritized aggregation scheme based on two prioritized aggregation operators namely, AND and SCORING. This prioritization makes the weights associated with each criterion dependent upon the satisfaction of the higher preferred criterion. The authors made use of 4 criteria in a personalized IR setting: *aboutness, coverage, appropriateness*, and *reliability*. In the same trend, Bouidghaghen et al. (2011) suggested a multi-criteria relevance model, but on a mobile IR setting, based on three dimensions of relevance: *topic*, *interest* and *location*. To aggregate these relevance criteria, the authors made use of the two previously cited "prioritized operators" (Costa Pereira et al., 2009), defining a priority order over the set of relevance dimensions. Palacio, Cabanac, Sallaberry, and Hubert (2010) have considered a geographic IR system involving three relevance dimensions: *spatial*, *temporal*, and *topical* information. The proposed system combines the results of three criteria with COMB* (Fox & Shaw, 1993) aggregation functions. Later, Gerani et al. (2012) have proposed a multicriteria relevance based method that allows generating a global score that does not necessarily require that the individual scores, that have to be combined, should be comparable. The authors rely on the Alternating Conditional Expectation (*ACE*) Algorithm (Breiman & Friedman, 1985) and the BOXCOX (Box, G. E. P. & Cox, D. R., 1964) model to analyse the incomparability problem and perform a score transformation whenever it is necessary. As an IR application, the authors consider a blog opinion IR setting. More recently, Eickhoff, Vries, and Collins-Thompson (2013) have introduced a copula-based method for combining multidimensional relevance estimates. The authors model multivariate document relevance scores based on a number of a document quality criteria and shown that *copulas* are able to model complex multidimensional dependencies between these relevance criteria. Their approach has been evaluated within three IR tasks for multidimensional relevance aggregation namely, opinionated blogs retrieval, personalized social bookmarking and child-friendly web search. The authors have tested the proposed copula-based approach against the product and sum baselines as well as the linear combinations scheme and shown that it outperforms these three baselines. Thereafter, they have investigated the usefulness of the approach in the score fusion problem relying on copula-based extensions of the two popular score fusion schemes *CombSUM* and *CombMNZ* (Fox & Shaw, 1993).

It is worth noting that many other studies that dealt with rank aggregation have been also proposed (Dwork, Kumar, Naor, & Sivakumar, 2001; F. Wei, Li, & Liu, 2010). The rank aggregation task which is encountered in many prominent situations such as metasearch (Aslam & Montague, 2001; Akritidis, Katsaros, & Bozanis, 2011), consists in computing a consensus ranking given individual ranking preferences of several judges (Renda & Straccia, 2003). Given the ranked lists of documents returned by multiple search engines in response to a given query, the problem of meta-search is to combine these lists in a way which optimizes the performance of the combination (Aslam & Montague, 2001). These ranking fusion methods can be classified based whether they rely on the scores or the ranks. In fact, the difference between multidimensional relevance aggregation and rank aggregation, is that aggregation occurs without dealing with the multidimensional nature of relevance or the criteria used for searching. These ranking functions use different methods in querying, but in most of the cases, they are based on the topical criterion or topical-dependent factors, despite the different used sources of evidence. For instance, Farah and Vanderpooten (2007, 2008) proposed a multi-criteria framework for rank aggregation using a decision rule based mechanism operating with the multidimensional property of the topical criterion. Among these dimensions, we cite for example, the *frequency*, *document length*, *prominence, etc.*

2.2.3 Learning to rank for IR

Based on machine learning algorithms, learning to rank methods have been widely used in IR to combine multiple document features with the purpose of optimizing document rankings. The features commonly include query-dependent measures such as BM25 score or query-independent ones such as PageRank importance. 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. Learning to rank approaches fall into three categories namely, the pointwise, pairwise and listwise (T.-Y. Liu, 2009). In the pointwise-approach, regression-based algorithms, classification-based algorithms and ordinal regression-based algorithms are used to predict relevance scores. The main idea behind the well known learning to rank algorithms that fall into the pairwaise-approach, such as RankSVM (Joachims, 2006) and RankNet (Burges et al., 2005), is the optimization of document pairs preference orderings based on a loss function. Listwise learning to rank methods straightforwardly represent the ranking task for IR as they minimize a loss function corresponding to standard IR evaluation measures, considering a ranked list of documents as input.

Intuitively speaking, the multidimensional relevance aggregation problem can be tackled by learning to rank methods where the features belong to different relevance dimensions. However, while this community have significant expertise in estimating topical document relevance and other additional criteria, the commonly applied combination schemes ignore the problem of modeling complex, multi-dimension dependencies. In practice, the sophisticated deployed learning to rank techniques tend to offer only limited insight to humans about why they were weighted highly for relevance. Indeed, these methods do not explore the relevance dimension level within an IR task and thus do not allow giving insight on how to consider importance and interaction between groups of features mapped to different relevance dimensions as stated by the aforementioned studies (Borlund, 2003; Saracevic, 2007). Through the fuzzy measure, our Choquet based aggregation approach is able to model many interactions between criteria and leads to results that are human- interpretable. As we previously stated, thanks to the interaction and importance indices our method offer qualitative understanding of the resulting model.

2.3 Contribution and motivations

As we previously stated, although many of the proposed approaches effectively perform in some IR applications, they are not effective in real-life applications since the user's needs involve preferences that lead to several relevance criteria which usually interact with each other. In practice, this problem is usually avoided by considering independent criteria (Cong et al., 2009; Göker & Myrhaug, 2008). Nevertheless, other works (Saracevic, 2007; Wolfe & Zhang, 2010; Carterette, Kumar, Rao, & Zhu, 2011; Eickhoff et al., 2013) have shown that relevance criteria usually interact. For instance, authors in (Carterette et al., 2011) have proved through an empirical study in a *tweet* search task, the existence of a positive correlation between the recency and the topical relevance criteria.

Moreover, classical aggregation operators are assumed to hold the additive property which can be effective and convenient in some applications, but can also be somewhat inadequate in many real-life IR tasks. For example, consider the relevance assessment of two documents D_1 and D_2 with respect to two relevance criteria. Then, assume that D_1, equivalently satisfied $w.r.t$ both criteria, is preferred to D_2 for which the global score is biased by one criterion. Actually, this problem can be dealt by using an averaging operator such as the weighted sum, but this does not give any way of preferring D_1 over D_2 if we consider that the latter have apparently the same global relevance scores. Clearly, this preference needs to tradeoff both relevance criteria appropriately. This becomes particularly challenging if we consider that a low score obtained on a given criterion can be a serious reason for discounting a document. Although some initiatives were recently proposed (Costa Pereira et al., 2012; Gerani et al., 2012), none were concerned with considering the interactions existing among the relevance criteria as it is the case in many real-life applications. The following example (inspired from (Grabisch, 1995)) sketches the impact of dependencies between correlated criteria on the global aggregated score.

Example Consider the problem of estimating the relevance scores of a subset of documents with respect to three relevance criteria: *topicality*, *authority* and *popularity*. Suppose that an averaging aggregation operator is used to evaluate these scores and assume that the first criterion is more important than the two others, *i.e.*, the weights could be 0.4, 0.3, and 0.3, respectively. Clearly, *authority* and *popularity* criteria may interact since, usually, documents published by influential authors are potentially popular and vice versa. Therefore, since these two criteria may present some degree of redundancy, the global evaluation will be overestimated (*resp.* underestimated) for popular (*resp.* non popular) documents published by influential authors (*resp.* uninfluential). Moreover, if we deal with a classical aggregation method such as the WAM, the documents scores *w.r.t* these redundant relevance criteria will be double counted. This undesirable phenomenon can be easily tackled by using a suitable fuzzy measure, where a *negative interaction* between the criteria *authority* and *popularity* is modeled to absorb the bias effect of these redundant criteria.
Consider again the three relevance criteria and suppose that one requires that the satisfaction of only one criterion produces almost the same effect than the satisfaction of both. For example, it is important that documents should be either popular or published by potential users. Of course, it is better that they would be relevant *w.r.t* both criteria. Clearly, such a behavior cannot be expressed by a classical aggregation method. Here, the importance of the pair is close to that of the single criteria, even within the presence of the other criteria. This condition could be easily expressed using a fuzzy measure, by modeling again a *negative interaction*. Alternatively, one can require that the satisfaction of only one criterion produces a very weak effect compared with

the satisfaction of both. Then, we speak about a *positive interaction*, where documents which are equivalently satisfied by all the set of criteria, should be preferred to those which are overestimated by one single relevance criterion.

To tackle these challenges, we propose to investigate the combination of general level relevance dimensions using a fuzzy-based aggregation operator. More oriented to the specific problem of relevance aggregation, our method is able to address the property of interaction between dimensions by modeling an integral aggregation function, namely the Choquet integral, with respect to a fuzzy measure expressing both their individual and joint importance. This aggregation method has the advantage of facilitating the task of interpreting the interaction phenomena between the relevance criteria with readily available interpretations via the Shapley and interaction indices. This mathematical facet of calculation makes the Choquet integral model flexible and robust (Grabisch, 1996). To the best of our knowledge, this kind of aggregation has not been previously used for such IR purpose. In this paper, we particularly explore the following thriving issues:

1. *How to model multidimensional relevance aggregation within dependent criteria?*

   As stated above, many pioneering works on multidimensional relevance argued that relevance dimensions usually interact with each other. Likely, we assume, in our context, that relevance dimensions, that will be used for aggregation, interact in real IR settings. To do so, we will use the Choquet integral to model interactions between the relevance criteria. One of its main benefits stands in its capability in representing many kinds of interaction among any set of criteria. This is done thanks to a fuzzy measure μ (or capacity), defined on each criterion and each subset of criteria I_i, which enables to avoid the overestimation (*resp.* underestimation) caused by possible dependencies between some criteria.

2. *How is effective the aggregation proposed within a social search task, namely the tweet search task?*

   To show the effectiveness of our aggregation approach on real IR world situations, we propose to instantiate our model on a social (microblogging) IR setting, more particularly, on a *tweet* search task. We will consider jointly three relevant criteria: *topicality*, *recency* and *authority*, formally described in previous works (Duan et al., 2010; Nagmoti, Teredesai, & De Cock, 2010). We experimentally show the dependency among these criteria and then show the appropriateness of the Choquet integral in aggregating them.

3 Background: aggregation in decision making problems

In this section, we present an overview of the aggregation problem in multi-criteria decision making (MCDM) problem. Then, we introduce a formalization of the MCDM problem and present formal definitions and related notions of aggregation operators.

3.1 Aggregation operators: an overview

Aggregation functions involve the ordering of a group of alternatives based upon their satisfaction to a collection of criteria. The research effort concerning aggregation functions has been disseminated throughout various fields including decision making, knowledge based systems and many other areas (James, 2010). The most widely applied aggregation functions are those in the averaging class. The arithmetic mean and its variations were often prominent in most of the cases. Aggregation operators or functions can be roughly classified into several categories compensatory, non-compensatory, conjunctive, disjunctive, and weighted aggregation approaches (Hwang & Yoon, 1981).

Compensatory operators are based on the assumption that a low score of a given alternative with respect to a high-preference criterion may be compensated by a high score on another high preference criterion. Compensative operators are comprised between minimum and maximum, *i.e.,* they are neither conjunctive nor disjunctive. The weighted sum is the most representative aggregation function of this class. The global score of each alternative is computed by multiplying the criterion weight by the alternative's performance score obtained on this criterion. Weighted quasi-arithmetic means (QAM) are also particularly interesting aggregation functions of this family. These functions are used prominently throughout the literature since they generalize a group of common means, *e.g.,* the harmonic mean, the quadratic mean and the power mean, which in turn includes as special cases other classical means like the arithmetic and geometric mean (Kolmogorov, 1930; Aczel, 1948). Another family extensively studied in the literature, is that of ordered weighted averaging functions (OWA) (Yager, 1988). The fundamental aspect of the OWA operator is the re-ordering step. More specifically, a given performance score is not associated with a particular weight, but rather a weight is associated with a particular ordered position of the score, which introduces a non-linearity into the aggregation process.

This provides a means for aggregating scores associated with the satisfaction of multiple criteria, which unifies in one operator the conjunctive and disjunctive behavior. In addition to these families, another form of operators investigated in the context of multi-criteria aggregation under uncertainty, is the concept of triangular norm (Menger, 1942). The current notion of *t-norm* and its dual operator *t-conorm* are introduced by (Schweizer & Sklar, 1960) and (Schweizer & Sklar, 1983). These operators may be seen as a generalization of the conjunctive "AND" (t-norms) and disjunctive "OR" (t-conorms) logical aggregation functions. Compensatory operators often require the user or the decision maker to specify priorities or preference relations expressed by means of cardinal weights or priority functions over the set of criteria. On the other side, non-compensatory functions, such as the MIN or the MAX aggregation schemes (Fox & Shaw, 1993) are generally dominated by just one criterion value, *i.e.*, the worst or the best score. The main limitation of these families is the fact that a large part of scores are ignored in the final aggregation process.

	Choquet Integral
OWA	$\mu_C = \sum_{j=0}^{i-1} w_{n-j}$, $\forall$ C such that $\|C\|= i$, where $\|C\|$ denotes the cardinal of the subset of criteria C.
WAM	The weight w_i of each criterion c_i is equal to (μ_{c_i}) and for every subset of criteria $C_1 \in \mathcal{C}$, $\mu_{C_1} = \sum_{c_i \in C_1} \mu_{c_i}$
AM	$\mu_{C_1} = \frac{\|C_1\|}{\|\mathcal{C}\|}$

Table 3: Particular cases of the Choquet integral.

Fuzzy integrals such as the Choquet integral and the Sugeno integral (Choquet, 1953) may be considered as a meta-class of aggregation functions. These aggregation operators which are defined with respect to a fuzzy measure are useful for modeling interactions between criteria, such as redundancies among the inputs or complementarity between some criteria. Special cases of the Choquet integral, depending on the fuzzy measure μ include weighted arithmetic mean, OWA operator and arithmetic mean. In Table 3, we present the corresponding measures in order to get a particular operator. The Choquet integral has grasped a lot of attention in fuzzy sets as well as decision making communities. However, research into its real-world use in the IR field is still in its seed.

3.2 Multidimensional relevance aggregation as a MCDM problem

A Multiple Criteria Decision Making (MCDM) method deals with the process of making decisions in the presence of multiple objectives or alternatives (Triantaphyllou, 2000). The main goal of MCDM methods is to assist a decision maker in selecting the *best* alternative(s) from a number of given ones $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$, under the presence of multiple criteria $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ and diverse criterion preferences.

The point of departure for any MCDM technique is the generation of the discrete set of alternatives, the formulation of the set of criteria and then the evaluation of the impact of each alternative on every criterion (Jankowski, 1995). The estimated impacts of alternatives a_j $(1 \leq j \leq M)$ on every criterion c_i $(1 \leq i \leq N)$ are called performance scores (or evaluations), that we denote C_{ij}, defined with respect to a partial preference order $\preceq_{c_i}$. Thereafter, preferences on the set of criteria may be formulated as it is the case for the weighted averaging operator, in a cardinal vector of normalized criterion preference weights $W = (w_l, w_2, \ldots, w_n)$ (with $0 \leq w_i \leq 1$ and n is then number of criteria). In the final step, performance scores are aggregated into a single one, for each alternative a_j $(1 \leq j \leq M)$, using an appropriate aggregation function $\mathcal{F}(C_{1j}, C_{2j}, \ldots, C_{Nj})$. The result is then an ordered set of alternatives with respect to the defined preferences.

Based on the multidimensional property of relevance, detailed in Section 2, we suggest in the following to model the multidimensional relevance aggregation as a MCDM problem. Thus, the set of alternatives is represented by the document collection and criteria are the possible relevance dimensions. Our research considers the following retrieval setting: a user U interacts with a document space $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ with a typical search engine through an information need stated by means of query q_k. In this setting, the user' relevance judgment is affected by a set of criteria $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, each of which having a given importance degree or preference. The aggregation problem consists in combining the performance scores C_{ij} of each document *w.r.t* all the relevance criteria. As we deal with an IR setting, we denote C_{ij} by $RSV_{c_i}(q_k, d_j)$ (*i.e.,* Retrieval Status Value), obtained for each document d_j, on each single criterion c_j in response to a given query q_k. Then, the result consists in the global score denoted by $RSV_{\{c_1,\ldots,c_N\}}(q_k, d_j)$, *w.r.t* a global preference relation $\preceq_{\mathcal{C}}$ upon the set of all criteria. More formally, an aggregation operator is expressed as follows:

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

$RSV_{c_i}(q_k, d_j)$ may also be interpreted as the satisfaction degree of document d_j *w.r.t* criterion c_i. To avoid the overestimation (*resp.,* underestimation) of the global relevance scores by those having high (*resp.*, low) values *w.r.t* some criteria, we normalized the performance scores before aggregation by scaling them into the range $[0 \ldots 1]$. The aggregation function that will be used within our approach, is the discrete Choquet integral (Choquet, 1953; Grabisch, 1996). This function allows to define a weight not only on each criterion, but also on each subset of criteria which gives rise to a more flexible representation of interaction among criteria (Grabisch, 1996).

4 iAggregator: a multidimensional relevance aggregation operator using the Choquet integral

In the remainder, we will rely on the Choquet integral fuzzy-based function to solve the multidimensional relevance aggregation problem. The choice of this operator is mainly motivated by its flexible representation of complex interactions among criteria, especially in situations involving redundant or complementary information. A first step that should be performed before proceeding to the multi-criteria aggregation with the Choquet integral, is the definition of the fuzzy measure values or capacities $\mu_{\{c_i\}}$ on each criterion and each subset of relevance criteria.

4.1 Definition of the fuzzy measure on the set of relevance dimensions

Let $\mathcal{C}$ be the set of criteria (*i.e.*, the relevance dimensions) and $I_{\mathcal{C}}$ be the set of all possible subsets of criteria from $\mathcal{C}$. A fuzzy measure is a 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 $\mu(I_{\mathcal{C}}) = 1$.

$\mu(I_{C_i})$ can be interpreted as the importance degree of the combination of the subset of criteria I_{C_i}, or similarly, its power to make decisions alone without the remaining relevance criteria. For the sake of notational simplicity, $\mu(I_{C_i})$ will be denoted in the remainder by μ_{C_i}.

Assume now that we have a document collection $\mathcal{D}$, and $d_j \in \mathcal{D}$. The global document's score of d_j given by the Choquet integral with respect to the fuzzy measure μ *w.r.t* a set of N relevance criteria $\mathcal{C}$ is defined by:

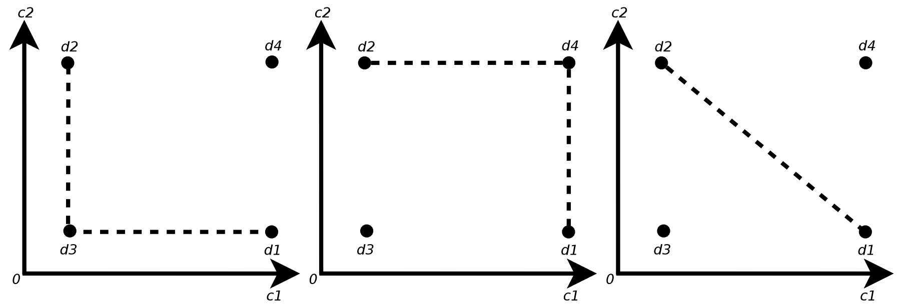
$$Ch_{\mu}(C_{1j}, ..., C_{Nj}) = \sum_{i=1,\ldots,N} (c_{(i)j} - c_{(i-1)j}).\mu_{C_{(i)}} \tag{1}$$

where $c_{(i)j}$ is the score obtained on a given criterion. The notation $c_{(\cdot)j}$ indicates that the indices have been permuted such that $0 \leq c_{(1)j} \leq ... \leq c_{(N)j}$. $C_{(i)} = \{c_i, ..., c_N\}$ is the set of relevance criteria with $C_{(0)} = 0$ and $\mu_{C_{(1)}} = 1$ and C_{ij} is the performance score[2] of d_j with respect to criterion c_i.

Obviously, the crucial part of using the Choquet integral is the modelling of interactions between criteria via the fuzzy measure μ. As the latter can model correlations or dependencies among criteria, which are relevance dimensions in our case, it is worth mentioning that there are three possible kinds of interactions represented in figure 1. The $x-$axis and $y-$axis represent the performance scores of the four documents d_1, d_2, d_3 and d_4 *w.r.t* the criteria c_1 and c_2 respectively. The documents connected by dashed lines have the same importance degree.

- ***Positive interaction:*** which can be called also *complementarity*; when the global weight of two relevance criteria is greater than their individual weights: $\mu_{c_i,c_j} > \mu_{c_i} + \mu_{c_j}$. This inequality can also be expressed as

[2] The difference between $c_{(i)j}$ and C_{ij} is that the performance scores $c_{(i)j}$ have been permuted before computing the overall scores.



(a) A postive interaction (b) A negative interaction (c) An independence

Fig. 1: Possible interactions between the set of criteria

follows: the contribution of criterion c_j to every combination of criteria that contains c_i is strictly greater than that of c_j to the same combination when c_i is excluded. In this case, criteria c_i and c_j are said to be negatively correlated. In other words, we say that the satisfaction of only one single relevance criterion should produce a very weak effect compared with the satisfaction of both criteria. Intuitively, in an IR setting this kind of preference favours documents which are satisfied equivalently by all the set of criteria, rather than those which are overestimated by one single relevance criterion. For instance, in Figure (1a), document d_4 should be preferred to documents d_2 and d_3, as they don't satisfy equivalently the two criteria c_1 and c_2.

- ***Negative interaction:*** when the global weight of two relevance criteria is smaller than their individual weights: $\mu_{c_i,c_j} < \mu_{c_i} + \mu_{c_j}$. We say that the union of criteria does not bring anything and the criteria are considered to act disjunctively. Thus, they are said to be *redundant*. This is indeed a key point about the Choquet integral, as it smooths the bias effect of redundant relevance criteria in the global documents evaluation. This is done by associating a small importance degree μ_{c_i,c_j} to the subset of the two redundant relevance criteria, compared to their single importance weights μ_{c_i} and μ_{c_j}. From figure (1b), we remark that document d_4 has the same importance as documents d_2 and d_3, as the satisfaction of one criterion from c_1 or c_2, which are in turn redundant, is sufficient to judge a document as relevant one.

- ***Independence:*** when there is no correlation between the set of criteria, the fuzzy measure is said to be *additive*: $\mu_{c_i,c_j} = \mu_{c_i} + \mu_{c_j}$. The WAM is an example of such functions that allow this independence between criteria. Accordingly, the importance of the inputs is taken into account and the weight of each criterion indicates its importance.

To facilitate the task of interpreting the behaviour of the Choquet integral and the interaction phenomena between the relevance criteria, we introduce below the *Importance index* (or *Shapley value*) (Shapley, 1953) and the *Interaction index* modelled by the underlying fuzzy measure.

Definition 1 *Importance index:* Let μ_{c_i} be the weight of relevance criterion c_i and $\mu_{Cr \cup c_i}$ its marginal contribution to each subset $Cr \in \mathcal{C}$ of other criteria. The importance index (Shapley, 1953) of c_i *w.r.t* a fuzzy measure μ is then defined as the mean of all these contributions:

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

$\phi_\mu(c_i)$ measures the average contribution that criterion (c_i) brings to all the possible combinations of criteria.

This *Importance index* gives no information on the phenomena of interaction existing among the relevance criteria. The overall importance of criterion c_i is not solely determined by its weight μ_{c_i} but also by its contribution to each subset of other criteria. Then, to quantify the degree of interaction between a subset of criteria, we introduce in the following the concept of *Interaction index* (Murofushi & Soneda, 1993).

Definition 2 *Interaction index:* Let $(\Delta_{c_i c_j} \mu_{Cr})$, with $Cr = \mathcal{C} \setminus \{c_i, c_j\}$, be the difference between the marginal contribution of criterion c_j to every combination of criteria that contains criterion c_i, and a combination from which criterion c_i is excluded:

$$(\Delta_{c_i c_j} \mu_{Cr}) = [\mu_{(\{c_i c_j\} \cup Cr)} - \mu_{(c_i \cup Cr)}] - [\mu_{(c_i \cup Cr)} - \mu_{Cr}]$$

This expression is defined to appraise the strength among two criteria c_i and c_j. When this latter expression is positive (*resp.* negative) for any $Cr \in \mathcal{C} \setminus \{c_i, c_j\}$, we say that both criteria c_i and c_j positively (*resp.* negatively) interact (*i.e.*, the contribution of criterion c_j is higher with the presence of criterion c_i).

The interaction index among two measures is thus defined as follows:

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

The interaction value, which falls into the interval $[-1..1]$, is zero when both criteria are independent and it is positive (*resp.* negative) whenever the interaction between them is positive (*resp.* negative).

4.2 Design of a multidimensional relevance function

The overall relevance score of document d_j, given by the Choquet integral *w.r.t* a fuzzy measure μ and according to the set $\mathcal{C}$ of N relevance criteria, is defined by:

$$\begin{aligned} RSV_{(c_1, c_2, \ldots, c_N)}(q_k, d_j) &= Ch_\mu(RSV_{c_1}(q_k, d_j), \ldots, RSV_{c_N}(q_k, d_j)) \\ &= \sum_{i=1}^{N} (rsv_{(i)j} - rsv_{(i-1)j}).\mu_{C_{(i)}} \end{aligned} \tag{2}$$

Where Ch_μ is the Choquet aggregation function, $rsv_{(i)j}$ is the permutation of $RSV(q_k, d_j)$ on criterion c_i such that $(0 \leq rsv_{(1)j} \leq ... \leq rsv_{(N)j})$, $C_{(i)} = \{c_i, ..., c_N\}$ is a set of relevance criteria with $\mu_{C_{(0)}} = 0$ and $\mu_{C_{(1)}} = 1$.

Once the Choquet operator as well as the interactions between criteria are defined, we present in the following the mechanism used for the identification of the fuzzy measures. In fact, the proposed methods in the literature for capacity identification differ according to the preferential information they require as input. Most of them are classified as optimization problems (Grabisch et al., 2008). In this paper, we will rely on the Least-squares based approach for the identification of capacities representing preferences on the relevance dimensions. This method is the most extensively used approach in the literature (Grabisch, 2002). First, we suppose to have initially a small selected subset of documents $\mathcal{D}$, that can be seen as a learning set. A ground truth is built with respect to a set of relevance criteria. Suppose now that we know the performance scores $RSV_{c_i}(q, d_j)$ assigned to each document d_j (*w.r.t* c_i) from the chosen subset of documents. In addition, we also suppose that we know the desired overall relevance scores $RSV^*_{\{c_1,\ldots,c_N\}}(q,d_j)$ for each document. The initial preferences can be formalized as follows:

- Given the partial order relation $\preccurlyeq_{c_i}$ (ranking of documents with respect to criterion c_i), the relation $d_1 \preccurlyeq_{c_i} d_2$ can be interpreted as "d_1 *is more relevant than* d_2 *according to the relevance criteria* c_i". In the context of the Choquet integral, this relation is translated as $Ch_\mu\ (RSV_{c_i}(q,d_1)) \leq Ch_\mu\ (RSV_{c_i}(q,d_2))$.
- $Ch_\mu\ (RSV_{c_i}(q,d_1)) \simeq (RSV_{c_i}(q,d_2))$ can be interpreted as "*the degree of satisfaction of* d_2 *w.r.t the relevance criterion* c_i *is the same as that of* d_1" ($d_1 \simeq_{c_i} d_2$).
- A partial preference order on the set of criteria $\preccurlyeq_{\mathcal{C}}$, *i.e.,* $c_1 \preccurlyeq_{\mathcal{C}} c_2$ is interpreted as c_1 is more important than c_2.
- A partial preference order on the subset of criteria $\preccurlyeq_I$, *i.e.,* $I_{C_1} \preccurlyeq_I I_{C_2}$ is interpreted as the combination of criteria I_{C_2} is more important than the combination of the subset of criteria I_{C_1}.

Suppose now that we know the performance scores $RSV_{c_i}(q,d_j)$ that should be assigned to each document $d_j \in \mathcal{D}$ (*w.r.t* criterion c_i). Then, the main objective of the Least-squares based approach is to minimize the total squared error E^2 between the desired global relevance score, given on each document d_j, and the global scores calculated by the Choquet integral as follows:

$$E^2 = \sum_{k=1}^{l}(Ch_\mu(RSV_{c_1}(q,d_j),\ldots,RSV_{c_N}(q,d_j)) - RSV^*_{\{c_1,\ldots,c_N\}}(q,d_j))^2 \tag{3}$$

This optimization process is discussed in detail in Section 6.3.

5 Experimental evaluation setting

The proposed multidimensional relevance operator is evaluated within a social IR setting, namely tweet search task. In this section, we present the experimental evaluation setup, the dataset used as well as the evaluation protocol.

5.1 Tweet search task

Seeking for information over microblogging spaces becomes a challenging task due to the increasing amount of published information. One of the most well known microblogging networking services that enables users to broadcast informations is Twitter[3]. The TREC 2011 Microblog Track (Ounis et al., 2011) defines *tweet* search as a real-time adhoc task where the users are interested in most recent and relevant informations. Recent works addressing the *tweet* search integrate a number of interesting features that were identified with potential implications in the final ranking of documents (Duan et al., 2010; Nagmoti et al., 2010). A number of proposed interesting criteria include, for instance, textual features, user's preferences, microblogging and social network features. In this work, we evaluate our Choquet integral based operator in a *tweet* search setting considering three relevance criteria: *topicality*, *recency* and *authority*. The aggregation of these criteria with the Choquet integral with respect to a fuzzy measure μ, in response to a user' query q, is defined as:

$$Ch_{\mu}(RSV_{To}(q,T_j), RSV_{Au}(q,T_j), RSV_{Re}(q,T_j)) = \sum_{i=1}^{3}(rsv_{(i)j} - rsv_{(i-1)j}).\mu_{C_{(i)}} \tag{4}$$

where T_j is a *tweet* (or microblog), $rsv_{(i)j}$ indicates that the performance score[4] considering criterion c_i on query q has been permuted such that: $0 \leq rsv_{(1)j} \leq rsv_{(2)j} \leq rsv_{(3)j}$ ($i.e.$, $rsv_{(i)j}$ is the $i{-}th$ smallest d_j score obtained on criterion $c_i \in \{To, Au, Re\}$). Note that $C_{(i)} = \{c_i, ..., c_3\}$ and Ch_{μ} is the global score that defines the final ranking of each tweet with respect to the three considered criteria.

In the following, we present a formal description of these relevance criteria in our evaluation setting.

- ***Topicality:*** is a content relevance criterion which describes the relevance between queries and *tweets*. To deal with this criterion, we propose to use the Okapi BM25 ranking function to rank *tweets* according to their relevance to a given search query. The standard BM25 weighting function is defined as follows:

$$BM25(T,Q) = \sum_{q_i \in Q} \frac{Idf(q_i)\cdot tf(q_i,T)\cdot (k_1+1)}{tf(q_i,T) + k_1(1 - b + b\frac{Length(T)}{avg_{length}})} \tag{5}$$

[3] *http://www.twitter.com*

[4] All the performance scores are normalised so that they belong to $[0 \ldots 1]$.

where $Idf(q_i)$ is the inverse document frequency, $Length(T)$ denotes the length of *tweet* T and avg_{length} represents the average length of tweets in the collection.

- ***Authority:*** represents the influence of *tweet*'s authors in Twitter. We define it as it was presented in (Nagmoti et al., 2010): $Au(T) = Au_{nb}(T) + Au_{me}(T)$, where:
  - $Au_{nb}(T)$ is the *total number of tweets*, to favor *tweets* published by influential users. $Au_{nb}(T) = N(a_i(T))$, with $a_i(T)$ represents the author of *tweet* T, and $N(a_i(T))$ denotes the number of *tweets* published by a_i.
  - $Au_{me}(T)$ is *number of mentions*, *i.e.,* more an author has been cited (or mentioned), more popular he is. It is defined as: $Au_{me}(T) = N_{me}(a_i(T))$, with $N_{me}(a_i(T))$ denotes the number of times the author of *tweet* T has been mentioned in the collection.
- ***Recency:*** is the difference between the time a *tweet* was published $Tp(T)$ and the query submission's timestamp $Ts(Q)$. $Re(T) = T_s(Q) - T_p(T)$. As we are interested in attempting the real time ad-hoc search task, all the tweets that occurred after the query time are excluded from the scoring.

5.2 Experimental datasets

We exploit the datasets distributed by TREC 2011 and 2012 Microblog tracks (Ounis et al., 2011, 2012). The Microblog Track is a focus area within TREC to examine search issues in Twitter. The *Tweets2011* corpus includes approximately 16 million *tweets* published over 16 days. The real-time ad-hoc task of the TREC Microblog 2011 track includes 49 time stamped topics which serve as queries. Each topic represents an information need at a specific point of time. Actually, we exploit the 49 topics of the TREC Microblog 2011 track for the capacities learning and we used the 60 TREC Microblog 2012 track for testing (Ounis et al., 2012). The general dataset statistics are reported in Table 4.

Tweets	16,141,812
Null tweets	1,204,053
Unique terms	7,781,775
Microbloggers	5,356,432
TREC Microblog 2011 Topics	49
TREC Microblog 2012 Topics	60

Table 4: Statistics of the TREC 2011 and 2012 Microblog tracks dataset.

We use the Terrier[5] search engine for indexing and retrieval. Since the task focuses on English tweets only, we eliminated the non English tweets using a simple language identifier tool. We also used some regular expressions to filter out some common types of tokens known in Twitter, but we did not filter the terms starting with the @ or # symbols. Although spam tweets are included, we did not perform any further processing as the main concern of our work is the multi-criteria relevance assessment.

5.3 Evaluation protocol

We adopt an evaluation protocol, consisting in two steps, as described in the following.

- ***Training step:*** This step consists in learning the Choquet capacities that be of use within each relevance dimension and each subset of relevance criteria in the aggregation process. Thus, we propose to exploit the TREC Microblog 2011 track (49) topics to experiment different combinations of capacities. As the relevance assessments relative to the track are available, we select the best capacities which optimize our aggregation model effectiveness in a such IR task.
- ***Testing step:*** This step consists in testing the iAggregator effectiveness based on the TREC Microblog 2012 track (60) topics. To assess the effectiveness of our approach, we rely on the precisions P@10, P@20, P@30 and MAP. We notice that P@30 is used officially to evaluate the retrieval performances of the

5 *http://terrier.org*

participating groups in the Microblog Tracks. These evaluation measures are computed with the standard *trec_eval*[6] tool.

Moreover, as it is case for learning to rank methods (T.-Y. Liu, 2009; Macdonald, Santos, & Ounis, 2013), our Choquet-based approach involves the use of a sample of top-ranked documents returned in response to a given query, initially based on the $BM25$ standard weighting model. Then, these documents are re-ranked with respect to the other considered criteria and the aggregation is done on the three relevance dimensions. This manner in which the approach is deployed, is also used by most of the TREC Microblog participants (Miyanishi, Seki, & Uehara, 2012; Liang, Qiang, Hong, Fei, & Yang, 2012) who used instead of $BM25$ a language model ranking. The participants are required to return top-ranked tweets prior to a query time per document according to their relevance score.

6 Results and discussion

In this section, we evaluate the effectiveness of iAggregator. We start first by introducing the evaluation objectives as well as the method used to tune the Choquet capacities and then discuss the obtained retrieval results.

6.1 Evaluation objectives

The aim of the experiments presented in the remainder is twofold:

- *Evaluate the impact of criteria interactions:* we show the ability of the Choquet integral in combining the relevance of dependent dimensions. The dependency property is estimated using a ranking correlation analysis. We also exploit the interaction and importance indices given through the fuzzy measure (*Cf.,* Section 4.1) to estimate the interactions between the considered criteria. The impact of the criteria dependency on the retrieval performances is also discussed.

[6] *http://trec.nist.gov/trec_eval*

- *Compare* iAggregator *to state-of-the-art aggregation operators:* we compare our approach *vs.* the Arithmetic Mean (Am), the Weighted Arithmetic Mean (Wam) and the linear combination mechanism as well as the Min, Max, Owa (Yager, 1988), OWmin (Dubois & Prade, 1996), And and Scoring aggregation operators (Costa Pereira et al., 2009). Afterward, we evaluate iAggregator with three conventional state-of-the-art learning to rank algorithms namely, RankNet (Burges et al., 2005), RankSVM (Joachims, 2006) and ListNet (Cao, Qin, Liu, Tsai, & Li, 2007).

6.2 Correlation analysis of the relevance dimensions

One of the main advantages in using the Choquet integral is its capability in aggregating interacting or correlated criteria. We present in this section a correlation analysis of the relevance dimensions through the Kendall's tau (τ) coefficient (Kendall, 1938). Our objective behind this is to show the interaction that could exist among the set of the considered criteria and to justify the use of the Choquet integral in such problems although the wide range of works proving this fact (Carterette et al., 2011; F. Wei et al., 2010).

Criterion	*Topicality (T)*	*Recency (R)*	*Authority (A)*
Topicality	1	**0.1580**	0.0013
Recency	–	1	0.0010
Authority	–	–	1

(a) Rank correlation coefficient for the single criteria rankings.

Criterion	$\{T, R\}$	$\{T, A\}$	$\{R, A\}$
$\{T, R\}$	1	**0.2290**	**0.1210**
$\{T, A\}$	–	1	−0.1030
$\{R, A\}$	–	–	1

(b) Rank correlation coefficient for the subset of criteria.

Table 5: Rank correlation analysis of the relevance criteria in the tweet search task.

The Kendall's tau (τ) correlation coefficient analyses the agreement between two rankings considering concordant and discordant pairs. In our context, we analyse the agreement between tweet rankings returned by each considered criterion solely on one side and subsets of criteria in the other side. The more similar (*resp.* reversed) the rankings are, the closer to 1 the correlation coefficient *tau* is (*resp.* -1). If the rankings are independent, then we would expect the coefficient to be approximately equal to zero. Table (5a) and Table (5b) show respectively the rank correlation coefficient for the individual criteria rankings and for the subset of relevance criteria rankings. Each coefficient is computed over the TREC Microblog 2012 track topics rankings. The global results are averaged over the resulted documents from each ranking. At a glance, Table (5a) highlights that *recency* and *topicality* are significantly correlated whereas authority seems to be independent and less important. From Table (5b), we notice, unlikely, that authority impacts ranking in presence of both *topicality* and *recency*. One can see that the impact is more important in presence of *topicality*, which is quite expected.

To present an in depth understanding of this interaction phenomena, we show in the following the Shapley values as well as the interaction indices obtained through the fuzzy measure within the TREC Microblog 2011 dataset. These parameters provide meaningful information that can be used to interpret the resulting model behavior.

Criterion	*Topicality*	*Recency*	*Authority*
Importance indice	**0.63**	**0.25**	**0.12**

(a) Criteria Importance indice.

Criterion	*Topicality*	*Recency*	*Authority*
Topicality	–	**+0.18**	+0.01
Recency	–	–	−0.10
Authority	–	–	–

(b) Criteria Interaction indice.

Table 6: Criteria importance and interaction indices

As we can see in Table (6a), given the marginal contribution of the content matching criterion in this IR task, we notice a high importance index of *topicality* with a value of 0.631. The *recency* relevance criterion is also given a quite high importance compared to the authority relevance dimension. This is not surprising as we deal with a real-time adhoc task as far as we are interested in the most relevant and recent *tweets* (Ounis et al., 2011). To analyse the *Interaction* phenomena existing among these relevance criteria and quantify its degree, we report in Table (6b), the values of the *Interaction* index between the three relevance criteria *Topicality*, *Recency* and *Authority*. From this Table, we can also remark that the authority criterion is not important and it does not bring any contribution when it is combined with topical relevance criteria.

It is also worth to mention notice a positive interaction between *topicality* and *recency* relevance criteria. This explains the higher contribution of these two criteria on the overall global scoring when they are present together and this concords with the aim of the considered IR setting. As it can be seen, these results are in concordance with those obtained by the Kendall's tau (τ) correlation coefficient, which prove the dependencies between the relevance criteria and motivate the use of the Choquet integral to aggregate them.

6.3 Tuning the Choquet capacities

In this section, we study the tuning of the capacity values that should be assigned to each criterion and each subset of criteria before computing the global Choquet scores. As we have the relevance assessments corresponding to the TREC Microblog 2011 track topics, we used the Least-squares based approach (*Cf.* Section 4.2) to tune the best combination of capacities that should be attributed to the relevance dimensions. Actually, each combination $\mu^{(i)}$ is composed by the following subsets of criteria:

$\mu^{(i)} = \{\mu_{\{topicality\}}, \mu_{\{authority\}}, \mu_{\{recency\}}, \mu_{\{topicality,authority\}}, \mu_{\{topicality,recency\}}, \mu_{\{recency,authority\}}\}$.

The different experimented capacity combinations $\mu_{\{.\}}$ used within each criterion and each combination of criteria, fall into $[0 \dots 1]$ and are computed with a step equal to 0.1. The adopted methodology for assigning capacity values for these relevance criteria is described below:

- **Step** 1: We start by assigning higher capacity values to the *topical* criterion and we start by 0.8. The capacity values of the *recency* and *authority* criteria are, respectively, equal to 0.1 and 0.1, *i.e.,* the sum of the three relevance criteria capacities is 1. The capacity values of each subset of criteria is the sum of its single capacity criteria. Then, we decrement the *topical* capacity value by 0.1 and we increment the *recency* capacity value, with the same step. This process is repeated until the *topical* capacity reaches 0.1 and the *recency* criterion capacity reaches 0.8.
- **Step** 2: We assign the *recency* criterion a high capacity, equal to 0.8. We decrement the *recency* capacity and we increment the *authority* criterion capacity until it reaches 0.8 (the step is 0.1).
- **Step** 3: We assign the *authority* criterion a high capacity, equal to 0.8. We decrement the *authority* capacity and we increment the *topicality* criterion capacity until it reaches 0.8 (the step is 0.1).

The adopted methodology for that purpose is detailed in Algorithm 1 while Table 7 describes the notations used within the Algorithm.

Notation	Description
Q_{learn}	The set of queries used to train the capacity values
$\mathcal{D}$	The document collection
$qrels$	The set of user's relevance assessments including relevant documents for each query $q \in Q_{learn}$. $qrels(q)$: relevant documents of query q.
$\mathcal{S}_{\mu^{(l)}}$	The set of the experimented capacity combination values. Each combination $\mu^{(i)} \in \mathcal{S}_{\mu^{(l)}}$ contains the capacities values of all the set and subsets of criteria. For instance, in the case of three criteria, each $\mu^{(i)}$ involves $(\{\mu_{c_1}; \mu_{c_2}; \mu_{c_3}; \mu_{c_1,c_2}; \mu_{c_1,c_3};\ \mu_{c_2,c_3}\})$.

Table 7: Notations used within Algorithm 1.

Algorithm 1 Identification of the Fuzzy Measures

Data: The set of queries Q_{learn}, document collection $\mathcal{D}$, the set $qrels$ of relevance assessments, capacity combinations $\mathcal{S}_{\mu^{(i)}}$.

Result: Capacity values $\mu_{\{i\}}$ of all the criteria and the subset of criteria.

1. **For** each query $q_k \in Q_{learn}$ **do**
2. **For** each capacity combination value $\mu^{(i)} \in \mathcal{S}_{\mu^{(l)}}$ **do**
3. Compute the $P@X$ of the returned documents in response to query q_k.
4. **End for**
5. **End for**
6. Select the combination of capacities $\mu^{(*)}$ that gives the best average $P@X$ on the training set Q_{learn}.
7. Select a subset of returned relevant documents $d_j \in R(q_k)$ such as $R(q_k) \subset qrels(q_k)$ with their given partial and global scores $RSV_{c_i}(q_k, d_j)$ based on combination $\mu^{(*)}$.
8. Select a subset of returned non relevant documents $d_{nr} \in NR(q_k)$ such as $NR(q_k) \subset \mathcal{D}$ and $d_{nr} \notin qrels(q_k)$ with their given partial and global scores $RSV_{c_i}(q_k, d_{nr})$ based on combination $\mu^{(*)}$.
9. Assign to each document $d_j \in R(q_k)$ higher (partial and global) scores than each document $d_{nr} \in NR(q_k)$ (even if they are ranked on the bottom).
10. Apply the Least-squares based approach on the set of assigned scores, return the outcome $\mu^{(**)}$.

We denote by $(\mu^{(1)})$ as the best combination obtained during the learning phase, which gives the higher average value of $P@30$ on the set of the TREC Microblog 2011 learning topics. This combination includes the following values: ($\mu_T = 0.8$, $\mu_A = 0.1$, $\mu_R = 0.1$, $\mu_{T,A} = 0.9$, $\mu_{T,R} = 0.9$, $\mu_{A,R} = 0.2$) where T, A and R stands respectively for *topicality*,*authority* and *recency*.

Figure 2 plots the performance of our approach within the TREC Microblog 2011 track topics, using the experimented combinations of capacities, which are obtained as described above. The x-axis represents the 21 trained capacities combinations $\mu^{(i)} \in \mathcal{S}_{\mu^{(l)}}$, which correspond to the fuzzy measures values of each criterion and each subset of criteria, as previously illustrated. The y-axis represents the results obtained in terms of $P@30$ after application of the Choquet integral within the aforementioned relevance criteria (To, Au, Re). The highlighted value in Figure 2 $(\mu^{(1)})$ indicates the best combination obtained during the learning phase as it gives the higher average value of $P@30$ on the set of the TREC Microblog 2011 learning topics (Q_{learn}).

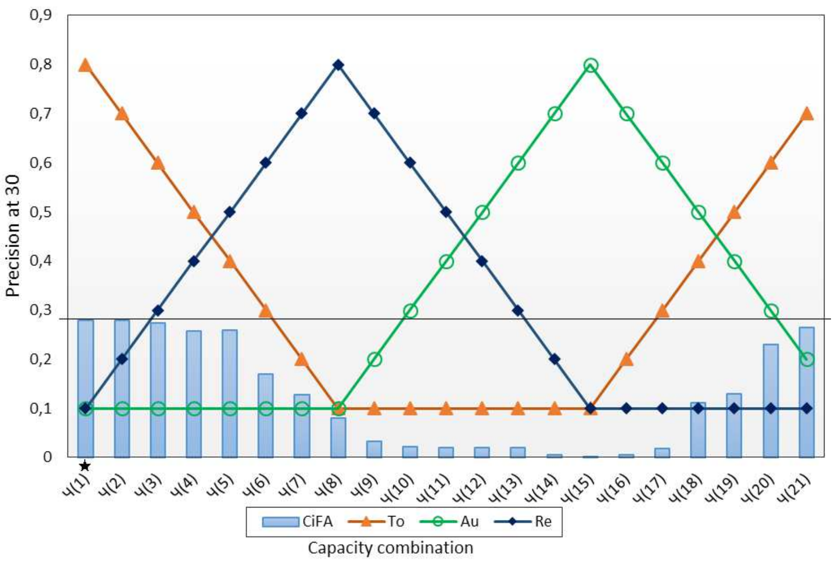


Fig. 2: iAggregator effectiveness within different capacity combination values on the learning phase.

As it may be seen from the returned capacity combination values of $(\mu^{(1)})$ and from the other experimented combination values in Figure 2, iAggregator is likely to be penalized for showing any preference for *tweets*, for which the *topical* and *authority* criteria are important. In fact, this latter is underperformed for topics for which *authoritative* tweets' scores are important, those tweets occur deeper in the ranking. Nevertheless, more recent topically scored tweets are more likely to be relevant, and this explains the positive interaction for both criteria. Therefore, as the system performs well when the *topical* and *recency* criteria are important, we consider it a "success" at dealing with the real-time TREC Microblog task.

Furthermore, the capacity combination returned by the Least squares based approach $\mu^{(*)}$ is composed by: (μ_{To} = 0.633, μ_{Re} = 0.204, μ_{Au} = 0.153, $\mu_{\{To,Re\}}$ = 0.961, $\mu_{\{To,Au\}}$ = −0.210, $\mu_{\{Re,Au\}}$ = −0.5). Our approach gives more importance to the *topical* and *recency* criteria. This fits well the Microblog track aim, as users are generally interested in *tweets* arriving at a specific time and concerning something happening now. We notice that the capacity values on the subsets $\{To, Au\}$ and $\{Re, Au\}$ are negatives. Thus, the contribution of the *topicality* relevance criterion to every combination of criteria that does not contain *authority*, is greater than its contribution when the criterion *authority* is highly scored. The same fact holds for the relevance dimensions *recency* and *authority*. The *authority* relevance dimension interacts negatively with both other criteria. Furthermore, despite its importance as a relevance criteria in *Twitter* (Chen et al., 2012), the *authority* criterion does not appear to be a factor for the topic, whence, the negative capacities assigned to $\mu_{\{To,Au\}}$ and $\mu_{\{Re,Au\}}$. However, the higher fuzzy measure associated to $\{To, Re\}$ indicates a positive interaction between both criteria. Interestingly enough, all the capacities obtained on the combination of relevance dimensions support the assumption that these criteria usually interact and this fact should be considered whenever it comes to aggregating them. All these results are consistent with those obtained from the correlation analysis presented in Section 6.2.

6.4 Effectiveness evaluation

In the following, we report the comparative effectiveness of iAggregator with state-of-the-art aggregation approaches and learning to rank methods.

6.4.1 Comparative evaluation with state-of-the-art aggregation operators

Here, we compare our approach to some traditional and state-of-the-art aggregation operators. More particularly, with the Arithmetic Mean (AM), the Weighted Arithmetic Mean (WAM) and the linear combination strategy (LCS) as well as the MIN, MAX, OWA (Yager, 1988), OWMIN (Dubois & Prade, 1996), AND and SCORING aggregation operators (Costa Pereira et al., 2012). The final scoring function for linear combination is computed as follows: $LCS(T) = \sum_{i=1}^{3}(\alpha_i lcs_i(T))$, where $lcs_i(T)$ is the performance score of *tweet* T on the criterion c_i, with $i \in$ *{topicality, authority, recency}*. The criteria weights used within WAM and LCS are tuned during the capacities learning phase within the TREC Microblog 2011 topics. We attributed them the optimal weights, *i.e.,* those giving the best average on $P@30$ during this phase: $\alpha_{recency} = 0.23$, $\alpha_{authority} = 0.16$ and $\alpha_{topicality} = 0.61$, where α_i is the weight of the criterion c_i.

Table 8 reports the results, by means of $P@10$, $P@30$ and MAP obtained by IAGGREGATOR against the aforementioned aggregation baseline operators. As it may be seen in Table 8, our aggregation model outperforms the whole baselines in both high precisions and MAP. In order to evaluate the significance of IAGGREGATOR' improvement, we conducted a paired two-tailed t-test. Significance testing based on the student t-test statistic is computed on the basis of all the tested precision levels. Considering the obtained p-values, we have marked with symbols †, ‡ and $\star$ statistically significant differences. The positive improvements obtained in favour to our approach were found to be statistically significant with p-values between 0.01 and 0.05 for LCS, and with p-values < 0.01 for the others aggregation operators. From this table, we also remark that the performances' improvements are important for the classical aggregation operators. We found performance improvement up to $P@30$ values of about 60.26% for the WAM and of 63.23% for the MAX operator, then the AM had similar performance, even enough there is a slight improvement drop. For the SCORING operator, the significant improvement is less important. As we considered the prioritization scenario Sc_1: $\{topicality\} \succ \{recency\} \succ \{authority\}$, giving the best $P@30$ average, we can conclude that the obtained difference of performance, in favour of IAGGREGATOR, is explained by the consideration of the interactions existing among the set of criteria, that we

Operator	Precision				% change
	P@10	P@20	P@30	MAP	
Am	0.1140 ‡	0.0991 ⋆	0.0936 ⋆	0.0535	**+59.89%**
Wam	0.1161 ⋆	0.0991 ⋆	0.0929 ⋆	0.0539	**+60.28%**
Lcs	0.1860 ‡	0.1833 ‡	0.1854 ‡	0.0928	**+20,73%**
Max	0.1088 ⋆	0.0895 ⋆	0.0860 ⋆	0.0604	**+63,23%**
Min	0.1793 ‡	0.1767 ‡	0.1764 ⋆	0.0879	**+24.58%**
Owa	0.1879 †	0.1776 ‡	0.1764 ⋆	0.0882	**+24.58%**
OWMin	0.1897 †	0.1776 ‡	0.1833 ⋆	0.0902	**+21.63%**
And	0.1793 ‡	0.1767 ‡	0.1764 ⋆	0.0882	**+24.58%**
Scoring	0.2018 ‡	0.1982 ‡	0.1977 ⋆	0.1091	**+15.47%**
iAggregator	**0.2345**	**0.2293**	**0.2339**	**0.1252**	–
	+13.94%	**+13.56%**	**+15.47%**	**+12.85%**	

Table 8: Comparative evaluation of retrieval effectiveness with state-of-the-art aggregation operators. % change indicates the iAggregator improvements in terms of $P@30$. The symbols †, ‡ and ⋆ denote the student test significance: "†": $0.05 < t \leqslant 0.1$; "‡": $0.01 < t \leqslant 0.05$; " ⋆ ": $t \leqslant 0.01$. The last row shows the iAggregator improvement in terms of $P@X$ and MAP with the best baseline (*i.e.*, Scoring).

involved by means of the fuzzy measures. Thus, the global scores can no longer be biased by dependent criteria. Compared to the And operator, the improvement difference is significantly better. We also notice, that although being a prioritized aggregation method, the And operator has low performances compared to those of the Lcs. The same holds for the Owa operator. This can be explained by the tuning performed for Lcs over the criteria weights, during the learning phase in order to get the best coefficients for each relevance criterion, against the Owa operator which primarily focuses on the weights with high values and gives low importance to the smallest weights in the evaluation. As the idea underlying this type of aggregation is to minimize the impact of small documents scores *w.r.t* a given criterion, a low weight can be a serious reason for discounting a document, which leads to a biased global evaluation. Regarding the OWMIN operator, the performance improvement is about 20% which is the same obtained for the LCS. This method uses a vector of levels of importance in order to minimize the impact of low weighted terms on the final documents scoring. Unlike OWA, the OWMIN operator uses the minimum instead of the average to compute the global documents' scores. This may explain the low performances of the classical averaging aggregation functions as sketched by Table 8. From this analysis, we can conclude that the major reason for the performance drop of the aforementioned aggregation operators, is the bias introduced by documents *w.r.t* to some criteria, especially those which are dependent (*Cf.* Section 6.2).

In order to get a more detailed understanding of the effectiveness of IAGGREGATOR with respect to the other aggregation approaches, we show in the overall curves, plot in Figure 3, a comparison with the aforementioned aggregation methods. The difference of $P@30$ values between our approach, OWA, OWMIN and prioritized

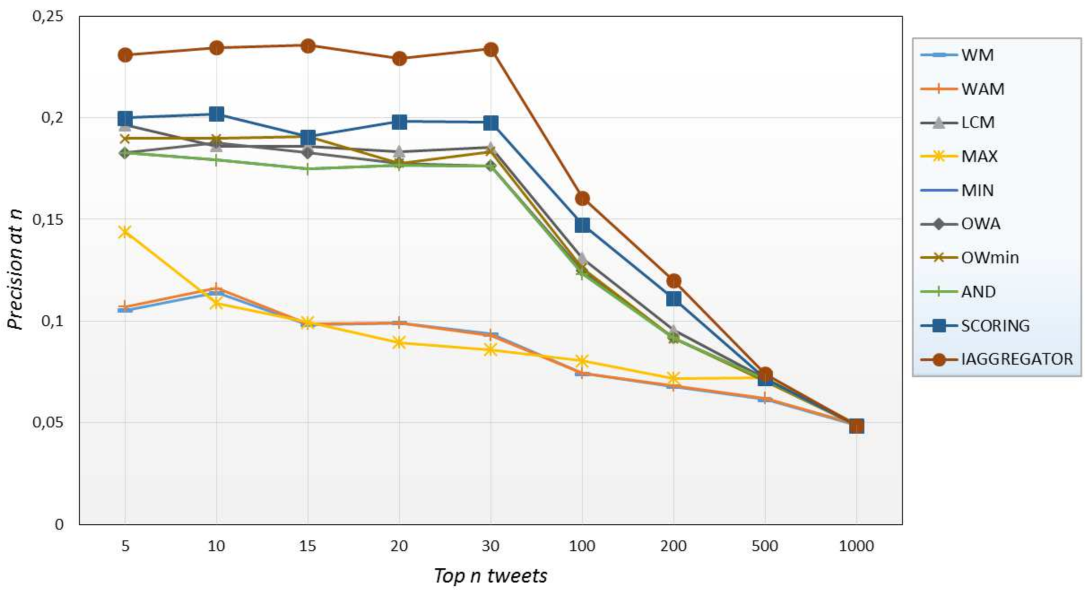


Fig. 3: Average precision at n comparison between IAGGREGATOR and standard aggregation mechanisms as well as some state-of-the-art aggregation operators

aggregation operators is more important comparing to standard aggregation schemes. As previously discussed, the lowest $P@30$ values are noticed for the AM and the WAM operators as well as the MAX aggregation method. For the latter, this is likely due to the fact that the global scores are dominated by the best single scores. For instance, the lack of satisfaction of a given relevance dimension such that *topicality*, which is an important criterion in *Twitter*, can be compensated by the surplus satisfaction of another criterion such as *authority*, which could be unrealistic. For the MIN and AND aggregation operators, the similar obtained results are not predictable, since the former is generally dominated by the worst score, whereas the latter, mainly based on the MIN operator, penalizes *tweets* highly satisfied by the least important criteria. However, if there are many *tweets* highly scored with respect to the *authority* criterion (which is likely the case), its overall satisfaction degrees would be biased by this relevance criterion.

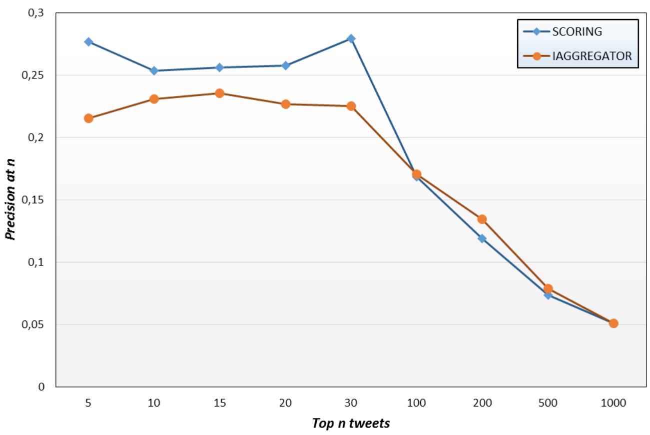
In order to further the effectiveness analysis, we present in the following a gain and failure analysis of the IAGGREGATOR approach. Table 9 presents the percentage of queries $\mathcal{R}^+$, $\mathcal{R}^-$ and $\mathcal{R}$ for which IAGGREGATOR performs better (*resp.* lower, equal to) than the different baseline operators, in terms of $P@30$, with an improvement higher (*resp.* lower, equal to) than 5% in comparison with the 5 best baseline operators. From Table 9, we

Query set	**AM**	**MIN**	**OWA**	**OWMIN**	**SCORING**
$\mathcal{R}^+$	56, 89%	43, 10%	43, 10%	36, 20%	36, 20%
$\mathcal{R}$	22, 41%	37, 93%	37, 93%	43, 10%	41, 37%
$\mathcal{R}^-$	20, 68%	18, 96%	18, 96%	18, 96%	22, 41%

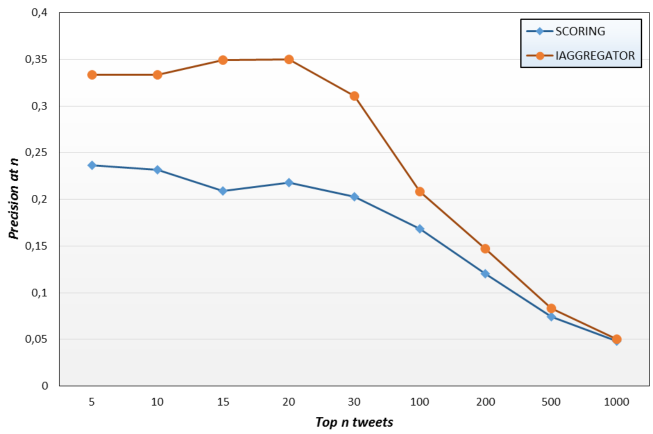
Table 9: Percentage of queries $\mathcal{R}^+$, $\mathcal{R}^-$ and $\mathcal{R}$ for which IAGGREGATOR performs better (resp. lower, equal to) than the different baseline operators, in terms of $P@30$.

can see that the percentage of queries for which IAGGREGATOR is underperformed by the baseline operators is almost the same, with an average of about 20, 34%. A manual analysis of these queries revealed that they are practically the same for all the aggregation baselines, with a quite difference for the AM aggregation method.

The high percentage for $\mathcal{R}^+$ queries, is attempted, as expected, for the same aggregation operator, *i.e.*, the Am. The difference of percentages is also well-nigh similar for the three sets of queries and these latter are almost the same for these three sets *w.r.t* the aforementioned baselines. We note that the lower percentage for $\mathcal{R}^+$ is marked for the Scoring and OWmin aggregation operators with a 36,2% of queries, whereas for $\mathcal{R}^-$ queries, the difference is noticeable for the Scoring operator with a percentage of about 22,41%. For the set of $\mathcal{R}$ queries, as the behavior of iAggregator and the Am aggregation mechanism are totally different, the percentage of queries, for which the performance in terms of $P@30$, is equal for both operators and is too low compared to the other baseline operators.



(a) Comparison for queries $\mathcal{R}^-$



(b) Comparison for queries $\mathcal{R}^+$

Fig. 4: Average precision at n comparison between iAggregator and the Scoring aggregation operator for both queries $\mathcal{R}^-$ and $\mathcal{R}^+$.

In Figure (4), we plot the difference performances in terms of $P@5 \ldots P@1000$ between iAggregator and the best baseline operator namely the Scoring operator for both $\mathcal{R}^+$, $\mathcal{R}^-$. As it may be seen from figure (4a), the difference of performance between both aggregation operators is not very significant for queries $\mathcal{R}^-$. Despite the fact that the Scoring operator performs well for these queries, our approach is shown to have quite good results. It is worth to mention that our operator gives a null $P@30$ score for 4 queries from $\mathcal{R}^-$. The average of performance difference is about $5,43\%$ and the high outperformance is marked for $n = 5$ with a difference equal to $22,21\%$. The worst $P@30$ difference performance values are observed for queries $T63$ and $T65$ from the set of the TREC Microblog 2012 track topics with the values of $75,01\%$ and $28,54\%$ respectively. The first topic namely: *"Bieber and Stewart trading places"* is a time sensitive query. Our model failed in retrieving the most relevant results first. This is likely due to the quite low capacity value assigned to the *recency* criterion ($\mu_{Re} = 0.204$) compared with the *topicality* one ($\mu_{To} = 0.633$). Though the high capacity assigned to the combination of both relevance dimensions, the Choquet operator failed in retrieving the most relevant *tweets* on the top of the ranking. The same holds for topic 65: *"Michelle Obama's obesity campaign"* and this is also likely due to the hypothesis that *tweets* which are recently published are considered as more important. This assumption is not suitable for every topic, as queries may have different hot time. If we assume that recent *tweets* have a higher score, this will affect relevant *tweet's* ranking position if more relevant documents are not published at the most recent time. For these topics, the SCORING operator performances are quite similar to the other baselines.

Whereas for the queries $\mathcal{R}^+$, for which IAGGREGATOR outperforms the baseline operators, it may be seen from Figure (4b) that the performance difference is very significant. This difference is sharper especially for the first *top* 30 retrieved *tweets* with an average value of about $34,41\%$, in contrast to an average value of about $14,10$ for $\mathcal{R}^-$ and for the same retrieved *tweets*. If we take for instance, the topic number 73: *"Iran nuclear program"*, we notice that IAGGREGATOR performs very well for this one, compared to all the other baselines. Likewise, the IAGGREGATOR performance for topic number 56: *"Hugo Chavez"*, is worthwhile compared to the other aggregation operators. These two queries are time sensitive, but unlike topics 65 and 63, they are not only relevant at a given position of time. More relevant *tweets* related to these two hot topics are published every day. This may explain the importance given to both relevance criteria (with high capacities values) *topicality* and *recency* (*Cf.,* Section 6.3), after the application of the Least square based approach. An indepth analysis of the nature of topics as well as the returned relevant *tweets* may reveal other interesting issues in order to improve the accuracy of our aggregation approach.

6.4.2 Comparative evaluation with learning to rank methods

Operator	**Precision**				**% change**
	P@10	**P@20**	**P@30**	**MAP**	
RankSVM	0.2500 ‡	0.2250 †	0.2218 †	0.0871	**+5.17%**
RankNet	0.2448 †	0.2198 †	0.2201 †	0.0858	**+5.89%**
ListNet	0.0931 ⋆	0.1009 ⋆	0.1115 ⋆	0.0485	**+52.33%**
iAggregator	**0.2345**	**0.2293**	**0.2339**	**0.1252**	–
	-6.60%	**+1.87%**	**+5.17%**	**+30.43%**	

Table 10: Comparative evaluation of retrieval effectiveness with conventional learning to rank methods. % change indicates the iAggregator improvements in terms of $P@30$. The symbols †, ‡ and $\star$ denote the student test significance: "†": $0.05 < t \leqslant 0.1$; "‡": $0.01 < t \leqslant 0.05$; " $\star$ ": $t \leqslant 0.01$. The last row shows the iAggregator improvement in terms of $P@X$ and MAP with the best baseline (*i.e.,* RankSvm).

In the following, we present a comparative evaluation of iAggregator versus conventional state-of-the-art learning to rank approaches. More specifically, we test our approach with two pairwise algorithms namely, RankNet and RankSVM and with a listwise learning to rank algorithm namely, ListNet. We used the open source code for RankSVM from (Joachims, 2006) and the RankLib library for the algorithms RankNet and ListNet[7]. For all the settings, all these algorithms were ran for 200 iterations with the measure $P@30$ as a loss function. The training models of these methods are learned with the same ground truth used for tuning the best capacity combination (*Cf.,* Section 5.3).

[7] *http://people.cs.umass.edu/∼vdang/ranklib.html*

From Table 10, we see that iAggregator significantly outperforms both pairwise and listwise algorithms. The improvement is up to 5% for RankNet and RankSVM and more than 52% for the ListNet algorithm. The result for RankSVM is quite lower than the other methods with an improvement varying between 1.87% and 5.17%. We also notice that iAggregator enhances the MAP obtained by all the tested approaches with an improvement of about 30.43% for the best baseline RankSVM.

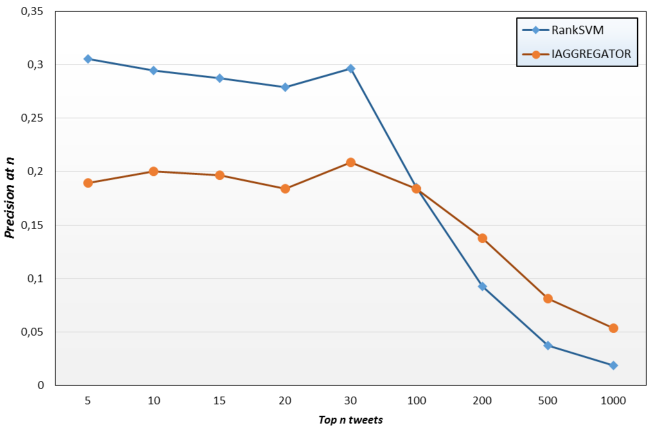
To provide an in-depth understanding of the iAggregator improvement in comparison to its counterparts, we present in the following a gain and failure analysis of the iAggregator approach. Table 11 presents the percentage of queries $\mathcal{R}^+$ and $\mathcal{R}^-$ for which iAggregator performs better (*resp.* lower) than the different learning to rank methods, in terms of $P@30$. Clearly, we can see that the percentage of queries for which

Query set	**RankSVM**	**RankNet**	**ListNet**
$\mathcal{R}^+$	$67,24\%$	$67,24\%$	$72,41\%$
$\mathcal{R}^-$	$32,76\%$	$32,76\%$	$27,59\%$

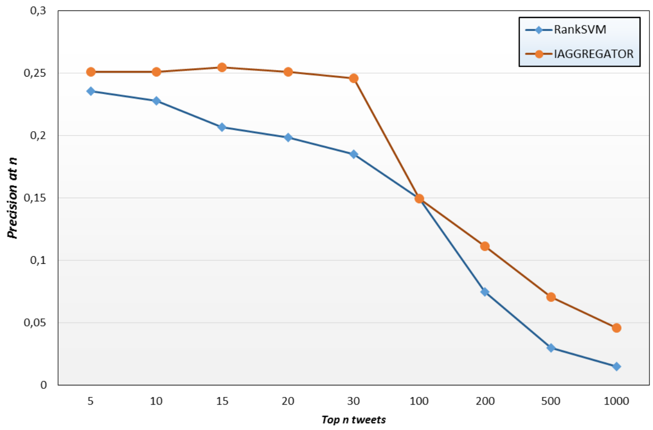
Table 11: Percentage of queries $\mathcal{R}^+$ and $\mathcal{R}^-$ for which iAggregator performs better (resp. lower) than the different learning to rank methods, in terms of $P@30$.

iAggregator performs better than the learning to rank methods is up to $67,24\%$ for both pairwise algorithms and of about $72,41\%$ for the listwise one. Despite the similar percentages obtained for $\mathcal{R}^+$ and $\mathcal{R}^-$ $w.r.t$ RankSVM and RankNet, the analysis of these queries reveals that they are not totally the same for both algorithms. The high percentage for $\mathcal{R}^+$ queries is attempted for the ListNet algorithm with a percentage of about $72,41\%$.

In Figure (5), we plot the difference of performances in terms of $P@5 \ldots P@1000$ between iAggregator and RankSVM (the best baseline) for both $\mathcal{R}^+$, $\mathcal{R}^-$. Obviously, we can notice from figure (5a), that the difference of performance between iAggregator and the baseline is quite significant for queries $\mathcal{R}^-$. This is not surprising given the fact that the percentage of queries for which iAggregator performs better than RankSVM is



(a) Comparison for queries $\mathcal{R}^-$



(b) Comparison for queries $\mathcal{R}^+$

Fig. 5: Average precision at n comparison between iAggregator and the RankSVM learning to rank algorithm for both queries $\mathcal{R}^-$ and $\mathcal{R}^+$.

relatively high (up to $67,24\%$) and given that the improvement in terms of $P@30$ despite being significant, is quite low (of about +5.17%).

For the queries $\mathcal{R}^+$, for which iAggregator outperforms the baseline learning to rank methods, we can see from Figure (5b) that the performance difference is less significant. In contrast to $\mathcal{R}^-$, for which RankSVM outperforms iAggregator only for the first top 100 tweets, we notice that for $\mathcal{R}^+$, RankSVM is outperformed for all the top K tweets. This may explain the high improvement marked by iAggregator in terms of MAP (30.43%) against the baselines. Likewise, we may further enhance these results by improving the ranking of the relevant tweets returned in the bottom (*i.e.*, beyond the top 30 tweets).

6.4.3 Comparative evaluation with official TREC Microblog results

In the following, we compare our results with the high performing official results from the TREC Microblog 2012 track (Ounis et al., 2012), in terms of the official measures ($P@30$ and MAP).

Model	$P@30$	MAP
Best 2012 TREC *run*	0.2701	0.2642
Second best *run*	0.2559	0.2277
TREC *median*	0.1808	0.1480
iAGGREGATOR	0.2339	0.1252

Table 12: Comparison with the official TREC Microblog 2012 Track results.

Results shown in Table 12 are rather promising, since we outperform the scores of the TREC $P@30$ and MAP medians. This fact holds despite the quite small exploited number of criteria, which wasn't the case for most of the participating groups. Moreover, apart from the capacities learning performed over the Microblog 2012 Track topics, we didn't made use of any external evidence. It can be seen from Tables 8, 10 and 12 that the MAP values obtained in our IR setting are relatively low compared to those of the official $P@30$ measure. As this fact holds for our Choquet based method as well as all the tested baselines, we may assert that these low values are not related to the aggregation phase. The major reason for that lie in the rankings returned by the query-likelihood $BM25$ model (topical criterion), on which were based the computation of the recency and authority document's scores. Still, we could say that our results are promising regarding the IR task setting and the track official evaluation measure used to judge the TREC participant results.

7 Conclusion

Aggregation of multiple relevance criteria is grasping an increasing attention in the IR community. Research directions that have been recently addressed in this concern, shows performance improvement on the quality of IR systems, when many relevance dimensions are combined together. The scrutiny of the experience arisen from previous state-of-the-art works reveals that there is a compelling need to design general effective multi-criteria aggregation frameworks to accurately combining the whole relevance criteria by taking into account their dependency. In this paper, after a critical review of the literature concerning the multi-criteria relevance aggregation, we proposed a new fuzzy integral-based approach, called iAggregator, based on the well studied and theoretically justified Choquet mathematical operator, for multidimensional relevance aggregation. This operator supports the observation that relevance criteria may interact with each other and may have a significant value on how well a ranking is assessed in a real-world IR setting. The effectiveness of the aggregation approach has been evaluated within a social microblogging IR setting. More particularly, a *tweet* search task where we made use of three relevance criteria. The iAggregator performance evaluation conducted within the TREC Microblog 2011 and 2012 tracks showed that the proposed operator allows to improve the ranking of the documents, in comparison to state-of-the-art aggregation operators, when relevance criteria interactions are taken into account by means of the fuzzy measure. An analysis of the success and failure of the search at the query level, revealed that our approach performs well for time sensitive hot topics for which *tweets* are not only relevant on a given position of time and that there is a need to further improve the performed capacity tuning. The study also showed that iAggregator performs well than the other baselines, for most of the TREC Microblog 2012 track topics. Afterwards, we have compared our approach with some representative learning to rank methods and showed that it performs better in terms of precision at different ranks and MAP.

This study has some limitations that can be explored in future work. First, it may be instructive to determine whether the results are generalizable by exploring the evaluation of other retrieval IR settings with a high number of incomparable relevance criteria and then gauge the consistency of the results obtained with those presented in this paper. Second, further research is needed to dynamically learn the capacity values through the study of large-scale query profiles; while several works studied the query sensitivity to orthogonal facets (such as navigational, transactional and informational (Jansen, Booth, & Spink, 2008), it would be interesting to shift the study towards multi-facet query sensitivity to dependent criteria and then attempt to tune the user preference criteria, leading to capacity values, along within the user's search sessions. The main outcome of this future research study would be the design of hypotheses supporting optimal tuning of the capacity values considering IR applications where multidimensional relevance is involved.

Acknowledgements This work is partially supported by the franco-tunisien PHC Utique project $11G1417$, entitled EXQUI.

References

Aczel, J. (1948). On mean values. *Bulletin of the American Mathematical Society*, *54*(4), 392-400.

Ah-Pine, J. (2008). Data fusion in information retrieval using consensus aggregation operators. In *Proceedings of the 2008 ieee/wic/acm international conference on web intelligence and intelligent agent technology - volume 01* (pp. 662–668). Washington, DC, USA: IEEE Computer Society.

Akritidis, L., Katsaros, D., & Bozanis, P. (2011). Effective rank aggregation for metasearching. *Journal of System Software*, *84*(1), 130–143.

Arrow, K. J. (1974). *Choix collectif et préférences individuelles* (Vol. 1; I. Aubin, Ed.). Calmann-Lévy.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th annual international* ACM SIGIR *conference on research and development in information retrieval* (pp. 276–284). New York, NY, USA: ACM.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Icwsm.* The AAAI Press.

Ben Jabeur, L., Tamine, L., & Boughanem, M. (2010). A social model for literature access: towards a weighted social network of authors. In *Adaptivity, personalization and fusion of heterogeneous information* (pp. 32–39). Paris, France.

Berardi, G., Esuli, A., Marcheggiani, D., & Sebastiani, F. (2011). ISTI@TREC microblog track 2011: Exploring the use of hashtag segmentation and text quality ranking. In *Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011).* National Institute of Standards and Technology (NIST).

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

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 *Proceedings of the 7th asia conference on information retrieval technology* (Vol. 7097, p. 169-180). Berlin, Heidelberg: Springer.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, *26*(2), 211–252.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*(391), 580–598.

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* (pp. 89–96). New York, NY, USA: ACM.

Cantera, J. M., Arias, M., Cabrero, J., Garcia, G., Zubizarreta, A., Vegas, J., et al. (2008). Mymose : Next generation search engine for mobile users. In *the 3rd edition of the future of web search workshop.*

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* (pp. 129–136). New York, NY, USA: ACM.

Carterette, B., Kumar, N., Rao, A., & Zhu, D. (2011). Simple rank-based filtering for microblog retrieval: Implications for evaluation and test collections. In *Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011).* National Institute of Standards and Technology (NIST).

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., & Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international* ACM SIGIR *conference on research and development in information retrieval* (pp. 661–670). New York, NY, USA: ACM.

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000). Developing a context-aware electronic tourist guide: some issues and experiences. In *Proceedings of the* SIGCHI *conference on human factors* *in computing systems* (pp. 17–24). New York, NY, USA: ACM.

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

Church, K., & Smyth, B. (2008). Who, what, where & when: a new approach to mobile search. In *Proceedings of the 2008 international conference on intelligent user interfaces* (p. 309-312). ACM.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris: Imprimerie Royale.

Cong, G., Jensen, C. S., & Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. *Proc. VLDB Endow.*, *2*, 337–348.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, *7*(1), 19–37.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, *24*(2), 87–100.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management*, *36*(4), 533–550.

Costa Pereira, C. da, Dragoni, M., & Pasi, G. (2009). Multidimensional relevance: A new aggregation criterion. In *Proceedings of the 31st european conference on advances in information retrieval* (p. 264-275).

Costa Pereira, C. da, Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing and Management*, *48*(2), 340–357.

Cuadra, C., & Katter, R. (1967). *Experimental study of relevance judgement* (Vol. 1; final report). OH: Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

Damak, F., Jabeur, L. B., Cabanac, G., Pinel-Sauvagnat, K., Tamine, L., & Boughanem, M. (2011). IRIT at TREC microblog 2011. In *Proceedings of the 20th* Text REtrieval Conference (TREC *2011).* National Institute of Standards and Technology (NIST).

Daoud, M., & Huang, J. X. (2013). Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science*, *64*(1), 190-212.

Daoud, M., Tamine, L., & Boughanem, M. (2010). A personalized graph-based document ranking model using a semantic user profile. In *Proceedings of the 18th international conference on user modeling, adaptation, and personalization* (p. 171-182). Berlin, Heidelberg.

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

Dubois, D., & Prade, H. (1996). Semantics of quotient operators in fuzzy relational databases. *Fuzzy Sets and Systems*, *78*, 89–93.

Dubois, D., & Prade, H. (2004). On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, *142*(1), 143–161.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on world wide web* (pp. 613–622). New York, NY, USA: ACM.

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

Farah, M., & Vanderpooten, D. (2007). An outranking approach for rank aggregation in information retrieval. In *Proceedings of the 30th annual international* ACM SIGIR *conference on research and development in information retrieval* (pp. 591–598). New York, NY, USA: ACM.

Farah, M., & Vanderpooten, D. (2008). An outranking approach for information retrieval. *Information Retrieval*, *11*(4), 315–334.

Fishburn, P. C. (1972). *Mathematics of decision theory.* Mouton, The Hague.

Fox, E. A., & Shaw, J. A. (1993). Combination of multiple searches. In D. K. Harman (Ed.), *The 2nd text retrieval conference* TREC*2* (pp. 243–252). National Institute for Standards and Technology.

Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, *1*(3-4), 219–234.

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

Göker, A., & Myrhaug, H. (2008). Evaluation of a mobile information system in context. *Inf. Process. Manage.*, *44*(1), 39–65.

Grabisch, M. (1995). Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, *69*(3), 279–298.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, *89*(3), 445–456.

Grabisch, M. (2002). Set function over finite sets: transformations and integrals. In E. Pap (Ed.), *Handbook of measure theory* (pp. 1381–1401). Elsevier.

Grabisch, M., Kojadinovic, I., & Meyer, P. (2008). A review of methods for capacity identification in choquet integral based multi-attribute utility theory: Applications of the kappalab R package. *European Journal of Operational Research*, *186*(2), 766-785.

Grabisch, M., & Labreuche, C. (2010). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals OR*, *175*(1), 247-286.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, *43*(9), 602–615.

Hattori, S., Tezuka, T., & Tanaka, K. (2007). Context-aware query refinement for mobile web search. In *Proceedings of the 2007 international symposium on applications and the internet workshops.* Washington, DC, USA: IEEE Computer Society.

Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval*, *4*(1), 33–59.

Hwang, C.-L., & Yoon, K. (1981). *Multiple attribute decision making. methods and applications: a state-of-the-art survey.* Springer-Verlag.

James, S. (2010). *Use of aggregation functions in decision making.* Unpublished doctoral dissertation, School of Information Technology, Faculty of Science and Technology, Deakin University. Retrieved from http://trove.nla.gov.au/work/168487579

Jankowski, P. (1995). Integrating geographical information systems and multiple criteria decision-making methods. *International Journal of Geographical Information Systems*, *9*(3), 251–273.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, *44*(3), 1251–1266.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th* ACM SIGKDD *international conference on knowledge discovery and data mining* (pp. 217–226). New York, NY, USA: ACM.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value tradeoffs* (Vol. 42; H. Raiffa, Ed.) (No. 3). Cambridge University Press.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1), 81–93.

Kishida, K. (2010). Vocabulary-based re-ranking for geographic and temporal searching at NTCIR geotime task. In *Proceedings of the 8th* NTCIR *workshop meeting on evaluation of information access technologies* (pp. 181–184).

Kolmogorov, A. N. (1930). On mean values. *Rend. Accad. dei Lincei*, *12*(4), 388–391.

Larkey, L. S., Connell, M. E., & Callan, J. (2000). Collection selection and results merging with topically organized u.s. patents and TREC data. In *Proceedings of the ninth international conference on information and knowledge management* (pp. 282–289). New York, NY, USA: ACM.

Le Calvé, A., & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing and Management*, *36*(3), 341–359.

Leung, C. W.-k., Chan, S. C.-f., & Chung, F.-l. (2006). A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. *Knowl. Inf. Syst.*, *10*(3), 357–381.

Liang, F., Qiang, R., Hong, Y., Fei, Y., & Yang, J. (2012). Pkuicst at trec 2012 microblog track. In *Proceedings of the 21th* T*ext* RE*trieval* C*onference (*TREC *2012).* National Institute of Standards and Technology (NIST).

Liu, F., Yu, C., & Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, *16*(1), 28–40.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225-331.

Ma, Z., Pant, G., & Sheng, O. R. L. (2007). Interest-based personalized search. *ACM Trans. Inf. Syst.*, *25*(1).

Macdonald, C., Santos, R., & Ounis, I. (2013). The whens and hows of learning to rank for web search. *Information Retrieval*.

Mata, F., & Claramunt, C. (2011). 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* (pp. 5–20). Berlin, Heidelberg: Springer-Verlag.

Menger, K. (1942). Statistical metrics. *In Proceedings of the National Academy of Sciences of the United States of America*, *28*(12), 535–537.

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).

Miyanishi, T., Seki, K., & Uehara, K. (2012). Trec 2012 microblog track experiments at kobe university. In *Proceedings of the 21th* T*ext* RE*trieval* C*onference (*TREC *2012).* National Institute of Standards and Technology (NIST).

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, *10*(3), 303–320.

Murofushi, T., & Soneda, S. (1993). Techniques for reading fuzzy measures (iii) : Interaction index. In *Proceedings of the 9th fuzzy systems symposium, sapporo, japan* (pp. 693–696).

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

Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior* (3rd ed.). Princeton University Press.

Ounis, T., Macdonald, J., & Soboroff, I. (2011). Overview of the TREC-2011 microblog track. In *Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011).* National Institute of Standards and Technology (NIST).

Ounis, T., Macdonald, J., & Soboroff, I. (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).

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.*, *11* (2), 91–109.

Rees, A., & Schultz, D. (1967). *A field experiment approach to the study of relevance assessments in relation to document searching* (Vol. 2). OH: Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

Renda, M. E., & Straccia, U. (2003). Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 acm symposium on applied computing* (pp. 841–846). New York, NY, USA: ACM.

Saracevic, T. (1996). Relevance reconsidered. In (pp. 201–218).

Saracevic, T. (2000). Digital library evaluation: Toward evolution of concepts. *Library Trends*, *49* (2).

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*, *58* (13), 2126–2144.

Saracevic, T., Rothenberg, D., & Stephan, P. (1974). Study of information utility. In (Vol. 11, p. 234-238).

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. *In Proceedings of the 54th* ASIS *Annual Meeting*, *28*, 126–133.

Schilit, B. N., LaMarca, A., Borriello, G., Griswold, W. G., McDonald, D., Lazowska, E., et al. (2003). Challenge: ubiquitous location-aware computing and the "place lab" initiative. In *Proceedings of the 1st* ACM *international workshop on wireless mobile applications and services on wlan hotspots* (pp. 29–35). New York, NY, USA: ACM.

Schweizer, B., & Sklar, A. (1960). Statistical metrics. *Pacific Journal of Mathematics*, *10*(1), 313-334.

Schweizer, B., & Sklar, A. (1983). *Probabilistic metric spaces.* New York: North-Holland Publishing Co.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. 28, pp. 307–317). Princeton: Princeton University Press.

Si, L., & Callan, J. (2002). Using sampled data and regression to merge search engine results. In *Proceedings of the 25th annual international* ACM SIGIR *conference on research and development in information retrieval* (pp. 19–26). New York, NY, USA: ACM.

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

Smith, M., Barash, V., Getoor, L., & Lauw, H. W. (2008). Leveraging social context for searching social media. In *Proceedings of the 2008* ACM *workshop on search in social media* (pp. 91–94). New York, NY, USA: ACM.

Steuer, R. E. (1986). *Multiple criteria optimization: Theory, computation and application.* New York: John Wiley & Sons.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Inf. Process. Manage.*, *28*(4), 503–516.

Su, L. T. (1994). The relevance of recall and precision in user evaluation. *J. Am. Soc. Inf. Sci.*, *45*(3), 207–217.

Taylor, A. R. (2012). User relevance criteria choices and the information search process. *Information Processing and Management*, *48*(1), 136–153.

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*, *43*(4), 1071–1084.

Taylor, R. S. (1986). *Value-added processes in information systems.* Ablex Pub. Corp.

Torra, V. (2005). Aggregation operators and models. *Fuzzy Sets Syst.*, *156*(3), 407–410.

Triantaphyllou, E. (2000). Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study* (Vol. 44, p. 5-21). Springer US.

Vickery, B. C. (1959). Subject analysis for information retrieval. In *Proceedings of the international conference on scientific information* (Vol. 2, pp. 855–865).

Vogt, C. C., & Cottrell, G. W. (1999). Fusion via a linear combination of scores. *Information Retrieval*, *1*(3), 151-173.

Wei, F., Li, W., & Liu, S. (2010). iRANK: A rank-learn-combine framework for unsupervised ensemble ranking. *Journal of the American Society for Information Science and Technology*, *61*(6), 1232–1243.

Wei, Z., Zhou, L., Li, B., Wong, K.-F., Gao, W., & Wong, K.-F. (2011). Exploring tweets normalization and query time sensitivity for twitter search. In *Proceedings of the twentieth text retrieval conference.*

Wolfe, S. R., & Zhang, Y. (2010). Interaction and personalization of criteria in recommender systems. In *Proceedings of the 18th international conference on user modeling, adaptation, and personalization* (pp. 183–194). Berlin, Heidelberg: Springer-Verlag.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions On Systems Man And Cybernetics*, *18*(1), 183–190.

Yau, S. S., Liu, H., Huang, D., & Yao, Y. (2003). Situation-aware personalized information retrieval for mobile internet. In *Proceedings of the 27th annual international conference on computer software and applications* (pp. 638–). Washington, DC, USA: IEEE Computer Society.