The Extension of Weight Determining Method for Weighted Zone Scoring in Information Retrieval

Sergey Sakulin[*1], Alexander Alfimtsev[2]

[1,2]Bauman Moscow State Technical University, 2 Baumanskaya st., 5-1, Moscow, 105005, Russian Federation

[*1]sakulin@bmstu.ru; [2]alfim@bmstu.ru

***Abstract*-Information retrieval based on weighted zone scoring means the assignment weight for each zone or each field in the document metadata. All these weights are obtained using machine learning methods. The paper presents a method of determining the weights using the fuzzy Choquet integral. This allows taking into possible account interdependence between the zone parameters when calculating the relevance and allows to obtain higher scoring accuracy.**

Keywords- Information Retrieval; Aggregation Operator; Fuzzy Measure; Choquet Fuzzy Integral

I. INTRODUCTION

Information retrieval is the search for documents that are relevant to the text query using various techniques [1]. When working with large collections of documents the search results can be so big that the user will simply not be able to see them all. So one of the important tasks of information retrieval is to rank search results according to their relevance to the query.

If we use the documents' metadata in this ranking, we need to take into account the expert's knowledge about metadata's structure and its characteristics. Documents' metadata are the fields (such as the date the document was created, type of document, the book's cost, etc.) and the zones (title, author, publisher, abstract, keywords, the text etc.). The difference between the zones and fields lies in the fact that the field may have a limited predefined set of values and the zone's set of values is not limited. Further, we consider the fields as a special case of zones. Search results ranking method was described in [1]. This method is based on allocating weight g_h to each zone h. The weights are setting using machine learning based on training examples. Denote the text query as q and the document as d. In weighted zone scoring each pair $(q,\ d)$ is assigned a value on the unit interval by calculating the linear combination of each zone scores. Consider a set of documents each of which has H zones. Let $g_h \in [0,\ 1],\ 1 \le h \le H$, such that $\sum_{h=1}^{H} g_h = 1$, $s_h \in [0,1]$, while zone score s_h considering the degree of compliance (or non) between the query and the h-th zone of the document. This value can be calculated in different ways for each of the zones [1]. Consider one of the most common ways to calculate it. For example, if all the query terms contained in the particular zone, value s_h is equal to 1; if only one term is contained in the zone, value s_h is equal to 1 / r; if any term is not contained in the zone, value s_h is equal to zero, where r is the number of terms in the query. Other ways to compute this value involve using the frequency with which the query term occurs in the particular zone as input information or may be based on quality indicators of the document, age of the document, its length and so on. In particular, there is a zone score calculating method based on the band function BM25F [2], which takes into account the query term occurrence frequency in the document zones. BM25F is based on function BM25 [3] which is a linear combination of three main attributes: the term frequency, the document frequency, and the length of the document. In this paper, the focus is made not on how to calculate the zone scores s_h but on the aggregation of these scores into a single score of document's relevance to the query $score(q,d)$. This aggregation was performed by a linear combination of zone scores [1]:

$$score(q,d) = \sum_{h=1}^{H} g_h s_h \tag{1}$$

Suppose that we have a set of training examples each of which is a tuple consisting of the query q, the document d, and rating of relevance for the pair $(q,\ d)$. Usually each query q is linked with a set of documents which is completely ordered by an expert according to their relevance. In accordance to this order the rating of the relevance can be assigned by the expert within unit's interval. Then the weights g_h are determined by machine learning using available examples so that the resulting values of the weights allow to approximate the rating of relevance of the training examples. Getting weight coefficients is reduced to an optimization problem with the objective function in the form of total error corresponding to training examples. There are also empirical rules for weights assigning to the document zones. For example, the authors of paper [4] believe that

they can achieve higher ranking accuracy by assigning the relatively high weight to the document title zone. The authors of paper [5] made the assumption that the ranking accuracy of news' documents can be increased by separating the first sentence to a separate zone and assigning increased weight to this zone. These and other similar rules can be applied in machine learning within the zone scores aggregation using a weighted arithmetic mean aggregation operator [6].

The approach described above in all its varieties assumes an implicit assumption of the mutual independence of the values s_i. However, it can be shown that the values s_i can be dependent of each other. For example, if the query term is in the title of the new document most likely to meet this term in the first sentence of the document. In this case, we are dealing with a positive correlation between values s_i and if we calculate the relevance score by a weighted arithmetic mean (1) we obviously get some redundancy of result. This phenomenon of aggregated values positive correlation and ways of compensation corresponding redundancy to the result is discussed in detail for example in [7]. A possible example of a more complex dependence will be the next one. Suppose an expert knows the following: query term occurs both in the body and in the abstract of several documents. These documents are ordered by relevance using the following rule. The document with the same term in the head zone is more relevant to the query than a document with the same term in the "document type" zone. Such dependence between the zone scores s_i is known as the preferred dependence of criteria [7]. This dependence cannot be expressed by any of the additive operators including the weighted arithmetic mean operator. Such knowledge cannot be formalized by the form of rules for the zones' weights obtaining using machine learning with weighted average aggregation operators. Thus, we coarsen the result when applying the weighted average operator to compute the relevance of documents to the query and assuming that the values s_i are always independent of each other.

II. FUZZY MEASURES AND THE CHOQUET INTEGRAL

If the aggregated criteria are interdependent, then Choquet integral with respect to fuzzy measure can be used instead of the weighted arithmetic mean operator for formalizing that dependence. Choquet integral is a generalization of the weighted arithmetic mean operator in case of dependence between values s_h (which we call the criteria of aggregation for following established terminology [7, 8]).

Fuzzy measure expressed the subjective weight or importance of each subset of criteria and defined as follows [7].

Fuzzy (discrete) measure is a function $\psi:\ 2^J \to [0,1]$, where 2^J is the set of all subsets of the criteria index set $J = \{1,...,\ H\}$, which satisfies the following conditions:

1) $\psi(\varnothing) = 0,\ \psi(J) = 1$;

2) $\forall D, B \subseteq J:\ D \subseteq B \Rightarrow \psi(D) \le \psi(B)$

Further, we will omit the curly brackets writing $i,\ ij$ instead of $\{i\},\ \{i, j\}$ respectively. Instead of the "criterion of the index $i \in J$", we will also use the "criterion i" instead of the "criteria index set J", we will use the "set of criteria J", both done for brevity reason.

Firstly, we consider the basic concepts used in the fuzzy measures theory. Shapley [9] proposed a definition of the criterion importance coefficient based on several natural axioms. In the context of the fuzzy measures theory Shapley index for the criterion $i \in J$ with respect to fuzzy measure ψ is determined by the following expression:

$$\Phi_{Sh}(i) := \sum_{D \subseteq (J-i)} \frac{\left(|J| - |D| - 1\right)!|D|!}{|J|!}\left[\psi\left(D \cup i\right) - \psi\left(D\right)\right]$$

Murofushi and Soneda proposed an interaction index between criteria [10]. This index is used to express the sign and degree of interaction between criteria and is determined by the following expression:

$$I(i, j) := \sum_{D \subseteq (J-\{i,j\})} \frac{\left(|J| - |D| - 2\right)!|D|!}{\left(|J| - 1\right)!}\left[\psi(D \cup ij) - \psi(D \cup i) - \psi(D \cup j) + \psi(D)\right]$$

Choquet integral using for dependent criteria aggregation was considered in [7, 8]. Particularly criteria preferred dependence modeled by Choquet integral is discussed in [8]. In [11] discussed in detail the application of a new method of machine learning based on the Choquet integral in different application areas, and concluded the feasibility of its use. In the

field of information retrieval, Choquet integral can be used for modeling expert preferences formalized by rules similar to the rules described in the previous section.

The Choquet integral of the criteria $s_1,..., s_H$ with respect to ψ is defined by $CH_\psi(s_1,..., s_H) := \sum_{h=1}^{H} s_{(h)}[\psi(A_{(h)}) - \psi(A_{(h+1)})]$, where (*) indicates a permutation of J, such that $s_{(1)} \le \cdots \le s_{(H)}$. Also $A_{(h)} = \{(h),..., (H)\}$ and $A_{(H+1)} = \varnothing$ [7].

III. FUZZY MEASURE IDENTIFICATION FOR WEIGHTED ZONE SCORING

If we use the weighted arithmetic mean operator for weighted zone scoring then weights g_h can be directly set by the expert. But due to the great complexity of this task in most cases these weights are determined based on machine learning [1]. If we use the Choquet integral for weighted zone scoring it is required to obtain a fuzzy measure ψ instead of weights g_h. Direct assignment of fuzzy measure by an expert is even more difficult task than weights setting due to exponentially increasing complexity. For example, for the four criteria an expert will have to set $2^4 = 16$ fuzzy measure's coefficients. Such setting is impossible in practice. Therefore, the coefficients of fuzzy measure ψ are obtained using machine learning as it is done for the weighted arithmetic mean operator. For realization of such machine learning procedure it is necessary to form a set of training examples and a set of formal empirical rules like those described above. Each of the training examples is a triple $\Delta_k = (d_k,\ q_k,\ r(q_k,\ d_k))$ in which the assessment of relevance $r(q_k,\ d_k)$ of the document d_k to the query q_k is assigned by an expert on the unit interval or these assessments are ranked by an expert. The rules are the limitations both on the fuzzy measure and the Choquet integral as weak partial orders on the set of zone scores realizations, results of aggregation (final relevance of the document), the Shapley indices, and interaction indices of criteria. Methods used to formalize these rules were considered in detail in [7]. In particular, if the rule states that the zones cores are correlated then it will be formalized by assigning a positive sign to the interaction index of these scores. In practice, to enable the expert to create such rules it is common to use 2nd-order fuzzy measures and, accordingly, 2nd-order Choquet integrals. Remaining relatively simple it allows to model the interaction between the criteria which are described by the rules similar to those mentioned above. The paper [12] is entirely devoted to the question under what conditions such a simplification (using of the 2-order Choquet integral) is correct. This paper presents necessary conditions that should satisfy the expert preferences in order that they can be formalized using the 2nd-order Choquet integral.

For each training example, we have the s_h values that are appropriated for any area of the document. Relevance of the document d_k to the query q_k will be determined as $score(q_k,\ d_k) = CH_\psi(s_1,..., s_H)$. Because of the nature of available information in the form of rules described above we need to choose a method of identification of fuzzy measure.

Method based on minimization of fuzzy measure variance or maximization of fuzzy measure entropy is the most suited for solving many practical problems [13]. One of the advantages of this method is the lack of any strict requirements to input information, in contrast to other methods of identification of fuzzy measure. This method is based on the principle of maximum entropy proposed in 1957 by Jaynes [14]. In relation to the construction of aggregation operators that principle involves the use of all available information about the aggregation criteria but the most unbiased attitude to the inaccessible information. We will follow this principle in weighted zone scoring of the documents, that is, taking into account the expert knowledge in the form of training examples and rules we will consider the missing information without bias. Kojadinovic [13] extended the principle of maximum entropy on the utility theory and developed fuzzy measures identification method based on this. The objective function of this method is defined as the variance of fuzzy measure:

$$F_{MV}(\psi) := \frac{1}{|J|}\sum_{i\in J}\sum_{G\subseteq J-i}\frac{(|J|-|G|-1)!|G|!}{|J|!}\left(\sum_{D\subseteq G} a(D\cup i) - \frac{1}{|J|}\right)^2.$$

Corresponding optimization problem takes the following form. Minimize $F_{MV}(\psi)$ under the following constraints:

$$\begin{cases} \sum\limits_{\substack{D\subseteq G \\ |D|\le\kappa-1}} a(D\cup i)\ge 0,\ \ \forall i\in J,\ \ \forall G\subseteq J-i \\ \sum\limits_{\substack{D\subseteq J \\ 0<|D|\le\kappa}} a(D)=1 \\ CH_{\psi}(\mathbf{g})-CH_{\psi}(\mathbf{g}')\ge\delta_{CH} \\ .... \end{cases}$$

Here $G\subseteq J;\ \ \kappa$ is the order of fuzzy measure ψ; δ_{CH} - indifference threshold that is set by an expert to compare the two results of aggregation; $a(D)$ is set function of a set J, this is called the Möbius function and is defined by the following expression and is given by $a(D)=\sum\limits_{G\subseteq D}(-1)^{|D|-|G|}\psi(D)$.

IV. THE PROCEDURE FOR DETERMINING THE WEIGHTS FOR WEIGHTED ZONE SCORING

If the aggregation operator is the Choquet integral with respect to the fuzzy measure, this procedure consists of the following steps.

Step 1. Form a set of zones $J=\{1,...,\ H\}$ for the document and a method of zone scores s_h calculating.

Step 2. Generate training examples using a collection of documents, these examples being relevance estimation and(or) non-strict partial order on the set of the estimates, i.e. implement expert ranking of documents relative to the query. Create rules in the form of partial weak orders on sets of Choquet integral parameters.

Step 3. Formalize obtained on the previous step information in the form of restrictions on the Choquet integral parameters in the form of inequalities with indifference thresholds. Set the indifference thresholds from training examples and scales that have been applied.

Step 4. Identify fuzzy measure on the basis of obtained in the previous step information by the minimizing dispersion method.

When new available information is added to the set of training examples and the set of rules the procedure is repeated from step 3. The Choquet integral with respect to the fuzzy measure ψ is aggregating operator for zone scores s_h through which the documents are ranked according to their relevance to the query.

V. EXPERIMENT

During the experiment we do not attempt to create a complete search engine. The purpose of experimental study was to obtain an answer to the question about the practical applicability of fuzzy measures and the Choquet integral in the field of information retrieval.

A set of training examples included 30 queries, 100 terms, and 300documents (publications in the field of artificial intelligence).

The procedure discussed above was put in practice to determine the fuzzy measures for the weighted zone scoring.

Step 1. We considered five zones of document: title $(h=1)$, abstract $(h=2)$, keywords $(h=3)$, main text $(h=4)$, and references $(h=5)$. These zones correspond to the zone indicators s_h which are calculated based on the function BM25F [2].

Step 2. Initial data for machine learning comprised both set of training examples and the following empirical rules similar to those discussed above.

Set of training examples Δ_k, where $k=1,...,\ 1000$ was received with experts' support. As noted above, each of these examples is a triple: $\Delta_k=(q_k,\ d_k,\ r(q_k,\ d_k))$. Relevance of the document d_k to the query q_k was evaluated on a scale which is the set S={0, 1, 2, 3, 4} in the same manner as it was done in [6]. In this set, "0" means that the document does not fully matches the query (no relevance), "4" means full compliance (document is relevant to the query), other values correspond to intermediate gradations of relevance.

Also, we obtained the following empirical rules in this step with experts' support:

Rule 1. If the query term was met in the title, it is likely to meet the same term both in the abstract and in the main text.

This rule means that the corresponding criteria are positively correlated and their interaction indices are less than zero. Then interaction indices of these criteria are defined by the following inequalities:

$$I(1,3) < 0;\ \ I(1,4) < 0;\ \ I(3,4) < 0 \tag{2}$$

Rule 2. In order to have the document relevant to the query it is least important that the query term is contained in the list of references; more importantly, that the query term is contained in the main text; more importantly to meet the term in the keywords; and finally, most importantly to meet the query term in the title and (or) in the annotation.

This rule means the following. Importance of the criterion s_5 is less than the importance of the criterion s_4. Similarly, importance of the criterion s_4 is less than importance of the criterion s_3. Importance of the criterion s_1 is the same as importance of the criterion s_2 and more than importance of the criterion s_3. This reasoning can be expressed by a partial weak order $\succeq_J$ on the set J of document's zones:

$$5 \prec_J 4 \prec_J 2 \prec_J 1 \sim_J 3 \tag{3}$$

Rule 3. If the query term is found in the main text and in the abstract, in order to get the document being more relevant to the query it is preferable that the same term is contained in the title rather than it is contained in the "keywords". This rule can be expressed by the following preference relations on the set S of available realizations of criteria:

$$\mathbf{s}_1 \prec_S \mathbf{s}_2$$

$$\mathbf{s}_3 \prec_S \mathbf{s}_2$$

Here $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$ are the realizations of criteria for three documents from the training set.

Step 3. Inequalities (2) are translated into inequalities with indifference thresholds:

$$-\delta_I < I(1,3) < 0;\ \ -\delta_I < I(1,4) < 0;\ \ \delta_I < I(3,4) < 0$$

Here δ_I - indifference threshold defined by an expert. This threshold is interpreted as the minimum significantly non-zero absolute value of interaction index.

Partial weak order (3) is translated into inequalities with Shapley indexes of the criteria:

$$\Phi_{Sh}(4) - \Phi_{Sh}(5) \geq \delta_{Sh};\ \ \Phi_{Sh}(2) - \Phi_{Sh}(4) \geq \delta_{Sh};$$

$$\Phi_{Sh}(1) - \Phi_{Sh}(2) \geq \delta_{Sh};\ \ \Phi_{Sh}(3) - \Phi_{Sh}(2) \geq \delta_{Sh};$$

$$-\delta_{Sh} \leq \Phi_{Sh}(3) - \Phi_{Sh}(1) \leq \delta_{Sh}$$

Here δ_{Sh} is the indifference threshold defined by an expert. Shapley indices are significantly distinguished if their absolute difference exceeds indifference threshold δ_{Sh}.

Step 4. Training examples and rules formed the restrictions imposed on the Choquet integral and its parameters during the identification process of fuzzy measures. Fuzzy measure was identified by the minimum variance method using specialized package Kappalab [7] by the above described optimization problem. An important question that arose in the identification process related to the need for expert's assignment of indifference thresholds. These values were chosen on the basis of the document relevance scale: for the aggregation result indifference threshold was taken to be δ_c=0,25. In addition, restrictions imposed on the indifference thresholds have been met (these thresholds can be set so that the fuzzy measure identification problem obviously does not have a solution), thus inequality proposed in [15] constraints the implementation of which allows to exclude such a situation.

Experiments for evaluating the accuracy of proposed method were performed on a statistically significant sample of 500 search queries containing the terms of training examples in various combinations.

Initially we calculated the documents' relevance $score(q_k, d_k)$ to these search queries using the set of training examples Δ_k, the empirical rules 1-3, and the method described above.

Then we calculated the relevance $score(q_k, d_k)$ on the basis of the machine learning method described in [1] and the set of training examples Δ_k.

Finally, it was found that the accuracy of search results ranking when using 2nd order Choquet integral aggregation has improved by an average of 4.5% when compared to the weighted average aggregation operator. The ranking accuracy considered is the difference between the relevance assigned by an expert and document relevance prepared on the basis of weighted zone scoring aggregation with one of two aggregation operators considered in this paper.

VI. CONCLUSIONS

The paper considers the practical application of the fuzzy measure and the Choquet integral in the field of information retrieval.

Experimental results have shown that increasing the accuracy of documents' relevance ranking can be achieved by using the Choquet integral as an aggregation operator for zone scores. The increase of accuracy of documents' relevance ranking is about 4.5% compared to using the weighted average operator. Further, it is assumed to investigate the application of proposed method for determining the weights on the various collections of documents as well as to investigate the practical applicability of the Choquet integral and fuzzy measure in other tasks of information retrieval such as automatic error correction, automatic abstracting, and annotating of texts.

REFERENCES

[1] Manning C., Raghavan P., and Schutze H., *Introduction to information retrieval*, Cambridge University Press, 2008, 544 p.

[2] Robertson S., Zaragoza H., and Taylor M., "Simple BM25 Extension to Multiple Weighted Field," *Proc. of ACM conference on Information Knowledge Management (CIKM)*, pp. 42-49, Nov. 2004.

[3] Robertson S. and Walker S., "Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval," *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232-241, 1994.

[4] Cohen W. and Singer Y., "Context-sensitive learning methods for text categorization," *ACM transactions on Information Systems*, vol. 17(2), pp. 141-173, 1999.

[5] Murata M., Ma Q., Uchimoto K., Ozaki H., Utiama M., and Isahara H., "Japanese probabilistic information retrieval using location and category information," *Proc. of the Fifth International Workshop on Information Retrieval with Asian Languages*, pp. 81-88, 2000.

[6] Krysta S. and Burges C., "A Machine Learning Approach for Improved BM25 Retrieval," *Proc of the 18th ACM Conference on Information and Knowledge Management*, pp. 1811-1814, 2009.

[7] Grabisch M., Kojadinovic I., and Meyer P., "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*, vol. 186(2), pp. 766-785, 2008.

[8] Marichal J.-L., "An axiomatic approach to the discrete Choquet integral as a tool to aggregate interacting criteria," *IEEE Transactions on Fuzzy Systems*, vol. 8(6), pp. 800-807, 2000.

[9] Shapley L., "A value for n-person games," Kuhn H, Tucker A., Eds., Contributions to the Theory of Games, Princeton: Princeton University Press, pp. 307-317, 1953.

[10] Murofushi T and Soneda S., "Techniques for reading fuzzy measures (III): interaction index," *9th Fuzzy System Symposium*, pp. 693-696, 1993.

[11] Fallah T. A., Cheng W., and Hüllermeier E., "Preference Learning using the Choquet Integral: The Case of Multipartite Ranking," *IEEE Transactions on Fuzzy Systems*, pp. 5-28, 2012.

[12] Mayag B., Grabisch M., and Labreuche Ch., "A representation of preferences by the Choquet integral with respect to a 2-additive capacity," *Theory and Decision*, № 71, pp. 297-324, 2011.

[13] Kojadinovic I., "Minimum variance capacity identification," *European Journal of Operational Research*, vol. 177(1), pp. 498-514, 2007.

[14] Jaynes E., "Information theory and statistical mechanics," *Physical Review*, № 106, pp. 620-630, 1957.

[15] Alfimtsev A., Sakulin S., and Devyatkov V., "Web personalization based on fuzzy aggregation and recognition of user activity," *International Journal of Web Portals*, vol. 4(1), pp. 33-41, 2012.

Sergey Sakulin graduated the Bauman Moscow State Technical University in 2001. He is a Ph.D. MSTU n.a. N.E. Bauman in 2009. Today he is assistant professor of Information systems and telecommunications department. He has ten scientific papers. Scientific interests lie in the fields of artificial intelligence methods and expert knowledge formalization and visualization.

Alexander Alfimtsev graduated the Bauman Moscow State Technical University in 2005. He is an associated professor at BMSTU, Information systems and telecommunications department. He has fifty scientific papers, including three patents for inventions. Scientific interests lie in the fields of intelligent multimodal interfaces, patterns recognition and computer vision.