Fuzzy Information Retrieval Based on Weighted Power-Mean Averaging Operators

Won-Sin Hong*, Shyi-Ming Chen*, and Shi-Jay Chen**

*Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R. O. C.
**Department of Information Management, Ching-Yun University, Jung-Li, Taiwan R. O. C.

Abstract

In recent years, some researchers used averaging operators (i.e., Infinite-One operators, Waller-Kraft operators, P-Norm operators and GMA operators) to deal with AND and OR operations of users' queries for fuzzy information retrieval, but they still have some drawbacks, e.g., sometimes query results do not coincide with the intuition of the human being. In this paper, we present new averaging operators, called weighted power-mean averaging (WPMA) operators, based on the weighted power-mean for dealing with fuzzy information retrieval to overcome the drawbacks of the existing methods. The proposed WPMA operators are more flexible and more intelligent than the existing averaging operators to deal with users' fuzzy queries for fuzzy information retrieval.

1 Introduction

In [1], [2], [3], [4] and [5], the T-operators (i.e., T-norms and T-cornorms) are used to deal with fuzzy information retrieval. Although T-operators support the ranking facility, they still have some drawbacks, e.g., sometimes query results do not coincide with the intuition of the human being. In [6], Lee et al. pointed out that three averaging operators (i.e. Waller-Kraft operators [7], P-Norm operators [8] and Infinite-One operators [9]) have been proposed to achieve high retrieval effectiveness for fuzzy information retrieval, where these three averaging operators can avoid the drawbacks of T-operators. However, in [10], Chen et al. pointed out that these three averaging operators still have some drawbacks, i.e., it is subjective and hard to determine appropriate values for the parameters of these averaging operators, respectively. Thus, in [10], Chen et al. presented new averaging operators based on the geometric mean, called the geometric-mean averaging (GMA) operators, to overcome the drawbacks of the Waller-Kraft operators, the P-Norm operators and the Infinite-One operators. However, the GMA operators still have some drawbacks, i.e., in some specific situations, the retrieval results do not coincide with the intuition of the human being. Thus, it is necessary to develop new averaging operators to overcome the drawbacks of the existing averaging operators for dealing with fuzzy information retrieval.

In this paper, we present new averaging operators, called weighted power-mean averaging (WPMA) operators, based on the concept of the weighted power-mean [8] to deal with fuzzy information retrieval. We also prove that the proposed WPMA operators are "positively compensatory" operators. In [3] and [6], Kim et al. pointed out that operators which have the "positively compensatory" property could provide high retrieval effectiveness. The proposed WPMA operators are more flexible and more intelligent than the average operators presented in [7], [9], [10] and [11] for dealing with fuzzy information retrieval.

2 Preliminaries

In [8], the definition of the weighted power means is defined as follows:

Definition 2.1: Assume that $\underline{a}$ and $\underline{w}$ are two positive n-tuples and $r \in R$, then the *rth* power mean $M_n^{[r]}(\underline{a},\ \underline{w})$ of $\underline{a}$ with weight $\underline{w}$ is define as follows:

$$M_n^{[r]}(\underline{a},\ \underline{w}) = \left(\frac{1}{W_n} \sum_{i=1}^{n} a_i^{\ r} w_i \right)^{\frac{1}{r}}, \qquad (1)$$

where $W_n = w_1 + w_2 + \ \cdots \ + w_n$.

Since $M_n^{[1]} = A_n$ and $M_n^{[-1]} = H_n$, in [8], Bullen et al. defined $M_n^{[0]} = G_n$ and proved that the definition of the weighted power mean is reasonable. Thus, the weighted power means form a natural extension of elementary means. Furthermore, when $r \to \infty$, $M_n^{[r]}(\underline{a},\ \underline{w}) = \max \underline{a}$; when $r \to -\infty$, $M_n^{[r]}(\underline{a},\ \underline{w}) = \min \underline{a}$.

In [3], Kim et al. pointed out that an information retrieval system based on the conventional fuzzy set model is defined by a quadruple <T, Q, D, F>, where

(1) T is a set of index terms, T=$\{t_1, t_2,\ \cdots, t_m\}$, where these index terms are used for representing queries and documents.

(2) Q is a set of queries, where query $q \in$ Q is a Boolean expression composed of index terms t_j, $1 \le j \le$ m, and the logical operators "AND", "OR"

and “NOT”.

(3) D is a set of documents, $D=\{d_1, d_2, \cdots, d_n\}$, where each document $d_i \in D$ is represented by $((t_1, e_{i1}), (t_2, e_{i2}), \cdots, (t_m, e_{im}))$, e_{ij} denotes the degree of strength of term t_j in document d_i, $e_{ij} \in [0,1]$, $1 \le i \le$ n, and $1 \le j \le$ m.

(4) F is an evaluation function,

$$F\text{: D} \times \text{Q} \rightarrow [0, 1], \tag{2}$$

which assigns a real value in the closed interval $[0,1]$ to each pair (d, q). It is a similarity measure between document d and query q.

From [12], we can see that the weight e_{ij} of term t_j in document d_i is determined either subjectively by domain experts or objectively by some algorithmic procedures, where $1 \le i \le$ n and $1 \le j \le$ m. One way for determining the degree of strength e_{ij} of term t_j in document d_i objectively is to consider the frequency of occurrence of index term t_j in document d_i.

3 Fuzzy Information Retrieval Based on the Weighted Power-Mean Averaging Operators

In this section, we present new averaging operators, called the Weighted Power-Mean Averaging (WPMA) operators, for fuzzy information retrieval, shown as follows:

$$F(d_i,\ \mathrm{q_{AND}}) = F(d_i,\ \mathrm{t_1\ AND\ t_2\ AND\ \cdots AND\ t_m})$$

$$= \left[\frac{1}{m^2}\sum_{k=1}^{m}(2m-2k+1)\, {e^*_{ik}}^{r}\right]^{\frac{1}{r}}, \tag{3}$$

$$F(d_i,\ \mathrm{q_{OR}}) = F(d_i,\ \mathrm{t_1\ OR\ t_2\ OR\ \cdots OR\ t_m})\ \text{`}$$

$$= 1-\left[\frac{1}{m^2}\sum_{k=1}^{m}(2k-1)(1-e^*_{ik})^{r}\right]^{\frac{1}{r}}, \tag{4}$$

where $r \in \{0.0001, 0.5\}$, e_{ij} denotes the degree of strength of term t_j in document d_i, e^*_{ik} denotes the kth smallest value of e_{ij}; the weight of the term in document d_i which is associated with the kth smallest value of e_{ij} in the AND query q_{AND} is $2m-2k+1$; the weight of the term in document d_i which is associated with the kth smallest value of e_{ij} in the OR query q_{OR} is $2k-1$, $1 \le i \le n$, $1 \le j \le m$, $1 \le k \le m$, $F(d_i, q_{\mathrm{AND}}) \in [0,1]$ and $F(d_i, q_{\mathrm{OR}}) \in [0,1]$.

In the following, we use two cases to discuss how the proposed WPMA operators are controlled by a parameter r. Assume that there are four documents d_1, d_2, d_3 and d_4, and assume that there are two queries q_1 and q_2, shown as follows：

$d_1 = \{(t_1, 0), (t_2, 0)\}$,
$d_2 = \{(t_1, 0), (t_2, 1)\}$,
$d_3 = \{(t_1, 1), (t_2, 0)\}$,
$d_4 = \{(t_1, 1), (t_2, 1)\}$,
$q_1 = t_1\ \mathrm{AND\ t_2}$,
$q_2 = t_1\ \mathrm{OR\ t_2}$.

Case 1: If the parameter $r = 0.0001$, then the degree of satisfaction $F(d_3, q_1)$ of the document d_3 with respect to the query q_1 and the degree of satisfaction $F(d_2, q_2)$ of the document d_2 with respect to the query q_2 can be evaluated, shown as follows:

$$F(d_3,\ q_1) = \left[\frac{1}{4}\left(3\times 0^{0.0001} + 1\times 1^{0.0001}\right)\right]^{10000} = 0,$$

$$F(d_2,\ q_2) = 1-\left[\frac{1}{4}\left(1\times(1-0)^{0.0001} + 3\times(1-1)^{0.0001}\right)\right]^{10000} = 1.$$

In the same way, we can calculate the values of $F(d_1, q_1)$, $F(d_2, q_1)$, $F(d_4, q_1)$, $F(d_1, q_2)$, $F(d_3, q_2)$ and $F(d_4, q_2)$, respectively, as shown in Table 1. From Table 1, we can see that when $r = 0.0001$, the proposed WPMA operators become the traditional Boolean operators. The operator graph [10] of the proposed WPMA operators is shown in Fig. 3.

Table 1. Query Result of the Proposed WPMA Operators (when r = 0.0001)

Documents	Terms		Queries	
	t_1	t_2	$q_1 = t_1$ AND $\mathrm{t_2}$	$q_2 = t_1$ OR $\mathrm{t_2}$
d_1	0	0	0	0
d_2	0	1	0	1
d_3	1	0	0	1
d_4	1	1	1	1

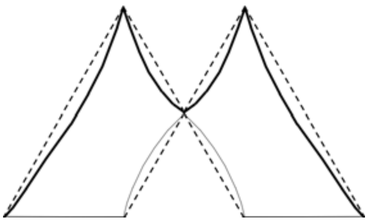

Fig. 3. The operator graphs of the proposed WPMA operators (when $r = 0.0001$).

Case 2: If the parameter $r = 0.5$, then the degree of satisfaction $F(d_3, q_1)$ of the document d_3 with respect to the query q_1 and the degree of satisfaction $F(d_2, q_2)$ of the document d_2 with respect to the query q_2 can be evaluated, shown as follows:

$$F(d_3,\ q_1) = \left[\frac{1}{4}\left(3\times 0^{0.5} + 1\times 1^{0.5}\right)\right]^2 = 0.063,$$

$$F(d_2,\ q_2) = 1-\left[\frac{1}{4}\left(1\times(1-0)^{0.5} + 3\times(1-1)^{0.5}\right)\right]^2 = 0.938.$$

In the same way, we can calculate the values of $F(d_1,\ q_1)$, $F(d_2,\ q_1)$, $F(d_4,\ q_1)$, $F(d_1,\ q_2)$, $F(d_3,\ q_2)$ and $F(d_4,\ q_2)$, respectively, as shown in Table 2. From Table 2, we can see that when $r = 0.5$, the proposed WPMA operators are compatible with the extended Boolean operators [11]. The operator graph of the proposed WPMA operators is shown in Fig. 4.

Table 2. Query result of the Proposed WPMA Operators (when $r = 0.5$)

Documents	Terms		Queries	
	t_1	t_2	$q_1 = t_1$ AND t_2	$q_2 = t_1$ OR t_2
d_1	0	0	0	0
d_2	0	1	0.063	0.938
d_3	1	0	0.063	0.938
d_4	1	1	1	1

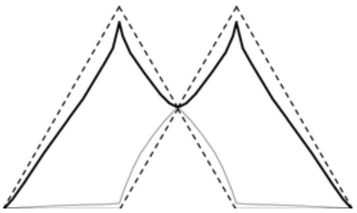

Fig. 4. The operator graph of the proposed WPMA operators ($r = 0.5$).

In [6], Lee pointed out that an operator which has the "*positively compensatory*" property could provide higher retrieval effectiveness. The "*positively compensatory*" operators are functions of the form $p:[0,1]\times[0,1]\rightarrow[0,1]$. They must satisfy the following two properties:

(1) $p(x,x) = x$; i.e., p is an idempotent function.

(2) $Min(x,y) < p(x,y) < Max(x,y)$, where $x \neq y$.

In the following, we assume that $0 \le x \le 1$, $0 \le y \le 1$, and prove that the proposed WPMA operators satisfy the following two properties.

Property 3.1: $F(d,\ t_1 \text{ AND } t_2)$ and $F(d,\ t_1 \text{ OR } t_2)$ are idempotent, where document $d = \{(t_1,\ x),\ (t_2,\ x)\}$, t_1 and t_2 are terms, and $0 \le x \le 1$.

Proof: Based on formula (3), we can see that

$$F(d,\ t_1 \text{ AND } t_2) = \left[\frac{1}{4}(3\times x^r + 1\times x^r)\right]^{\frac{1}{r}}$$
$$= \left[x^r\right]^{\frac{1}{r}}$$
$$= x.$$

Based on formula (4), we can see that

$$F(d,\ t_1 \text{ OR } t_2) = 1-\left[\frac{1}{4}(1\times(1-x)^r + 3\times(1-x)^r)\right]^{\frac{1}{r}}$$
$$= 1-\left[(1-x)^r\right]^{\frac{1}{r}}$$
$$= 1-(1-x)$$
$$= x.$$

Thus, the proposed WPMA operators are idempotent.

Q.E.D.

Property 3.2: Assume that there is a document $d = \{(t_1,\ x),\ (t_2,\ y)\}$, where t_1 and t_2 are terms, and $0 \le x \le 1$ and $0 \le y \le 1$. Then, $Min(x,y) < F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ OR } t_2) < Max(x,y)$, where $x \neq y$.

Proof:

(i) If $x > y$, then we can see that $Min(x,y) = y$ and $Max(x,y) = x$. Furthermore, based on formula (3), we can see that

$$F(d,\ t_1 \text{ AND } t_2) = \left[\frac{1}{4}(3\times y^r + 1\times x^r)\right]^{\frac{1}{r}}$$

$$\Rightarrow \left[\frac{1}{4}(3\times y^r + 1\times y^r)\right]^{\frac{1}{r}} < \left[\frac{1}{4}(3\times y^r + 1\times x^r)\right]^{\frac{1}{r}} < \left[\frac{1}{4}(3\times x^r + 1\times x^r)\right]^{\frac{1}{r}}$$

$$\Rightarrow F(d,\ t_2 \text{ AND } t_2) < F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ AND } t_1)$$

(by formula (3))

$$\Rightarrow y < F(d,\ t_1 \text{ AND } t_2) < x$$ (by Property 3.1)

$$\Rightarrow Min(x,y) < F(d,\ t_1 \text{ AND } t_2) < Max(x,y).$$

Based on formula on (4), we can see that

$$F(d,\ t_1 \text{ OR } t_2) = 1-\left[\frac{1}{4}(1\times(1-y)^r + 3\times(1-x)^r)\right]^{\frac{1}{r}}$$

$$\Rightarrow 1-\left[\frac{1}{4}(1\times(1-y)^r + 3\times(1-y)^r)\right]^{\frac{1}{r}} <$$
$$1-\left[\frac{1}{4}(1\times(1-y)^r + 3\times(1-x)^r)\right]^{\frac{1}{r}} <$$
$$1-\left[\frac{1}{4}(1\times(1-x)^r + 3\times(1-x)^r)\right]^{\frac{1}{r}}$$

$$\Rightarrow F(d,\ t_2 \text{ OR } t_2) < F(d,\ t_1 \text{ OR } t_2) < F(d,\ t_1 \text{ OR } t_1),$$

(by formula (4))

$$\Rightarrow y < F(d,\ t_1 \text{ OR } t_2) < x,$$ (by Property 3.1)

$$\Rightarrow Min(x,y) < F(d,\ t_1 \text{ OR } t_2) < Max(x,y).$$

Then, we prove "$F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ OR } t_2)$" based on formulas (3) and (4), shown as follows:

$$F(d,\ t_1 \text{ AND } t_2) = \left(\frac{y^r + y^r + y^r + x^r}{4}\right)^{\frac{1}{r}} <$$

$$\left(\frac{y+y+y+x}{4}\right) < \left(\frac{y+x+x+x}{4}\right) =$$

$$1-\left(\frac{(1-y)+(1-x)+(1-x)+(1-x)}{4}\right) <$$

$$1-\left(\frac{(1-y)^r+(1-x)^r+(1-x)^r+(1-x)^r}{4}\right)^{\frac{1}{r}}$$

$$= F(d,\ t_1 \text{ OR } t_2).$$

From the above discussions, we can see that $Min(x, y) < F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ OR } t_2) < Max(x, y)$.

(ii) In the same way, if $x < y$, we can get

$F(d,\ t_1 \text{ AND } t_1) < F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_2 \text{ AND } t_2),$ (by formula (3))

$x < F(d,\ t_1 \text{ AND } t_2) < y,$ (by Property 3.1)

$F(d,\ t_1 \text{ OR } t_1) < F(d,\ t_1 \text{ OR } t_2) < F(d,\ t_2 \text{ OR } t_2),$ (by formula (4))

$x < F(d,\ t_1 \text{ OR } t_2) < y,$ (by Property 3.1)

and $F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ OR } t_2).$ (by formulas (3) and (4))

Thus, $Min(x, y) < F(d,\ t_1 \text{ AND } t_2) < F(d,\ t_1 \text{ OR } t_2) < Max(x, y)$, where $x \neq y$. Q.E.D.

According to the above properties, we can see that the proposed WPMA operators have the “*positively compensatory*” property. Moreover, they have neither the “single operand dependent” property nor the “negatively compensatory” property. Therefore, they can overcome the drawback of the T-operators.

4 Conclusions

In this paper, we have presented the weighted power-mean averaging (WPMA) operators for fuzzy information retrieval. The proposed WPMA operators are more flexible and more intelligent than the averaging operators presented in [7], [9], [10] and [11] to deal with users’ fuzzy queries for fuzzy information retrieval due to the fact that the proposed WPMA operators have the following advantages:

(1) The proposed WPMA operators can overcome the drawbacks of the existing averaging operators for fuzzy information retrieval.
(2) The retrieval results of the proposed WPMA operators are much closer to the intuition of the human being than the existing averaging operators.
(3) We can easily determine appropriate values for the parameters of the proposed WPMA averaging operators. If we use the proposed WPMA operators to deal with traditional Boolean query processing for fuzzy information retrieval, then we can set the parameter $r = 0.0001$; if we use the proposed WPMA operators to deal with extended Boolean query processing for fuzzy information retrieval, then we can set the parameter $r = 0.5$.

Acknowledgements

This work was supported in part by the National Science Council, Republic of China, under Grant NSC 93-2213-E-011-018.

References

[1] D. A. Buell, “A problem in information retrieval with fuzzy set,” *Journal of the American Society for Information Science*, Vol. 36, pp. 398-401, 1985.

[2] D. H. Kraft and D. A. Buell, “Fuzzy set and generalized Boolean retrieval systems,” *International Journal of Man-Machine Studies*, Vol. 19, pp. 45-56, 1983.

[3] M. H. Kim, J. H. Lee, and J. Lee, “Analysis of fuzzy operators for high quality information retrieval,” *Information Processing Letters*, Vol. 46, pp. 251-256, 1993.

[4] D. Lucaralla and R. Morara, “FIRST: Fuzzy information retrieval system,” *Journal of Information Science*, Vol. 17, pp. 81-91, 1991.

[5] S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Kluwer, Dordrecht, 1990.

[6] J. H. Lee, “Properties of extended Boolean model in information retrieval,” *Proceedings of the Seventeenth Annual ACM Conference on Research and Development in Information Retrieval*, Dublin, pp. 182-190, 1994.

[7] W. G. Waller and D. H. Kraft, “A mathematical model of a weighted Boolean retrieval system,” *Information Processing and Management*, Vol. 15, pp. 235-245, 1979.

[8] P. S. Bullen, D. S. Mitrinovic, and P. M. Vasic, *Means and Their Inequalities*, D. Reidel Publishing Company, Dordrecht, 1988.

[9] M E. Smith, *Aspects of the P-Norm Model of Information Retrieval: Syntactic Query Generation, Efficiency and Theoretical Properties*, Ph.D. Dissertation, Cornell University, New York, 1990.

[10] S. J. Chen and S. M. Chen, “A new method for fuzzy information retrieval based on geometric-mean averaging operators,” *Proceedings of the 2002 International Computer Symposium: Workshop on Artificial Intelligence*, Hualien, Taiwan, Republic of China, 2002.

[11] G. Salton, E. A. Fox, and H. Wu, “Extended Boolean information retrieval,” *Communications of the ACM*, Vol. 26, pp. 1022-1036, 1983.

[12] G. J. Klir and B.Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall, New Jersey, 1995.