

覆盖粗糙模糊集的不确定性研究

郑婷婷 朱凌云

(安徽大学数学科学学院 合肥 230601)

摘要 不确定性度量是粗糙集理论中的基础问题之一。粗糙模糊集的不确定性一方面来自上、下近似集间差异产生的粗糙性,另一方面来自概念外延不清晰产生的模糊性。目前对于粗糙模糊集的不确定性研究仍不够透彻。针对覆盖近似空间下的粗糙模糊集不确定性,提出更加严格的度量修正准则,并借助上、下近似集隶属度与原模糊集隶属度之间的差异,给出修正粗糙度的概念。算例分析表明该方法能够更加准确地刻画实际问题。

关键词 覆盖近似空间,粗糙模糊集,粗糙度,粗糙熵,不确定性

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.11.048

Uncertainty Measure Research on Rough Fuzzy Sets in Covering Approximation Space

ZHENG Ting-ting ZHU Ling-yun

(School of Mathematical Sciences, Anhui University, Hefei 230601, China)

Abstract Uncertainty measure is one of the basic problems of rough sets theory. The uncertainty of rough fuzzy sets is caused by both the roughness of the difference of the upper and lower approximate sets and the fuzziness of the uncertain concept extension. Now the uncertainty research on rough fuzzy sets is not yet well understood. This paper discussed the uncertainty of rough fuzzy sets in covering approximation space, proposed the more strict revised measurement criterion, and brought out the revised roughness by the membership difference among the original fuzzy sets, the upper and lower approximation sets. Example analysis shows the measurement approach can describe the actual problems more accurately.

Keywords Covering approximation space, Rough fuzzy sets, Roughness, Rough entropy, Uncertainty

1 引言

不确定性是人们对客观世界或实际问题认识不足所导致的,只能用现有的知识对问题做出评价和判断。它广泛存在于现实生活中,表现形式也是多种多样。它最初起源于量子运动的不确定性^[1],如今人们发现无论在数学、统计学、化学、物理学等自然科学领域,还是在哲学、经济学、金融学、社会学、心理学等社会科学领域,到处都有它的身影。进入21世纪以来,信息不确定性度量越来越得到人工智能研究者的热情追捧。

粗糙集^[2]和模糊集^[3]是人工智能领域中处理不确定性问题的重要理论,它们都是对经典集合理论的扩展。粗糙集是用上、下近似集来描述不确定性问题的边界;模糊集是用隶属函数来刻画对象与集合之间的不确定关系。虽然不确定性在模糊集和粗糙集的表现形式上不一致,但它们对不确定性的刻画本质上是一致的^[4]。为了更加贴近实际问题,人们在研究粗糙集时,提出了将模糊集理论应用到粗糙集中,从而提出了模糊粗糙集和粗糙模糊集模型^[5]。其中粗糙模糊集是用模糊等价关系代替等价关系来确定上、下近似集。对于该模型

的不确定性研究,既要考虑到近似集间的差异,也要考虑到集合本身的模糊性。

本文通过分析现有覆盖近似空间下模糊粗糙集的不确定性度量方法,指出它们的不合理性,提出修正的度量准则,并从上、下近似模糊集与原模糊集的差异出发,给出了修正模糊度,讨论了其与一般粗糙集的粗糙度之间的关系。

2 覆盖近似空间下的粗糙模糊集模型

设 U 是一个非空有限论域, $C=\{K_1, K_2, \dots, K_m\}$ 是 U 上的一个经典覆盖,即 $\bigcup_{i=1}^m K_i = U$ (这里任意 $K_i \neq \emptyset$),称 (U, C) 是一个覆盖近似空间。

定义 1^[11] 设 (U, C) 是一个覆盖近似空间,对于任意 $x \in U$,称 x 在该近似空间中的全描述为: $Ad_C(x) = \{K \in C | x \in K\}$,简记为 $Ad(x)$;称 x 在该近似空间中的最小描述为: $Md_C(x) = \{K \in C | x \in K \wedge \forall S \in C (x \in S \wedge S \subseteq K \Rightarrow K = S)\}$,简记为 $Md(x)$ 。

定义 2^[12] 设 C 是论域 U 上的一个覆盖,对于任意 $K \in C$,若存在 $K_1, K_2, \dots, K_i \in C - \{K\}$,使得 $K = \bigcup_{j=1}^i K_j$,则称 K

到稿日期:2014-01-11 返修日期:2014-03-11 本文受国家自然科学基金青年科学基金(61203290),安徽省高校省级优秀青年人才基金重点项目(2013SQRL005ZD)资助。

郑婷婷(1978—),女,博士生,副教授,主要研究方向为粒计算理论, E-mail: tt-zheng@163.com;朱凌云(1989—),女,硕士生,主要研究方向为粗糙集、模糊集理论。

是 C 可约的,否则称 K 是 C 不可约的。若 C 中每个块都是不可约的,则称 C 是不可约的,否则是可约的。若 K 是 C 可约的,则将 K 从 C 中去掉,直到 C 不可约,则得到 C 的一个约简,称为 C 的最简覆盖,记为 $red(C)$ 。

文献[12]中已经证明 A 关于近似空间 (U, C) 和 $(U, red(C))$ 的上下近似是相等的,因此后面若不加说明, (U, C) 均表示最简覆盖近似空间。

关于覆盖粗糙模糊集的定义有很多,目前公认的主要有3类。

定义 3^[13] 设 (U, C) 是一个覆盖近似空间, A 是 U 上的一个模糊集,即 $A \in F(U)$,则 A 关于 (U, C) 的下近似隶属函数和上近似隶属函数分别为:

$$\mu_{\underline{C}1(A)}(x) = \inf\{\mu_A(y) | y \in \bigcup Md(x)\}$$

$$\mu_{\overline{C}1(A)}(x) = \sup\{\mu_A(y) | y \in \bigcup Md(x)\}$$

称 $(\underline{C}1(A), \overline{C}1(A))$ 为 A 关于覆盖 C 的第一种覆盖粗糙模糊集。

定义 4^[14] 设 (U, C) 是一个覆盖近似空间, $A \in F(U)$,则 A 关于 (U, C) 的下近似隶属函数和上近似隶属函数分别为:

$$\mu_{\underline{C}2(A)}(x) = \inf\{\mu_A(y) | y \in \bigcap Md(x)\}$$

$$\mu_{\overline{C}2(A)}(x) = \sup\{\mu_A(y) | y \in \bigcap Md(x)\}$$

称 $(\underline{C}2(A), \overline{C}2(A))$ 为 A 关于覆盖 C 的第二种覆盖粗糙模糊集。

定义 5^[15] 设 (U, C) 是一个覆盖近似空间, $A \in F(U)$,则 A 关于 (U, C) 的下近似隶属函数和上近似隶属函数分别为:

$$\mu_{\underline{C}3(A)}(x) = \sup_{K \in Md(x)} \{\inf_{y \in K} \{\mu_A(y)\}\}$$

$$\mu_{\overline{C}3(A)}(x) = \inf_{K \in Md(x)} \{\sup_{y \in K} \{\mu_A(y)\}\}$$

称 $(\underline{C}3(A), \overline{C}3(A))$ 为 A 关于覆盖 C 的第三种覆盖粗糙模糊集。

文献[15]曾介绍过这3种覆盖粗糙模糊集之间满足如下关系:对于任意 $A \in F(U)$, $\underline{C}1(A) \subseteq \underline{C}3(A) \subseteq \underline{C}2(A) \subseteq A \subseteq \overline{C}2(A) \subseteq \overline{C}3(A) \subseteq \overline{C}1(A)$ 。文献[16]证明了这3种覆盖粗糙模糊集等价当且仅当对于任意 $x \in U$, $\inf\{\mu_A(y) | y \in \bigcup Md(x)\} = \inf\{\mu_A(y) | y \in \bigcap Md(x)\}$ 且 $\sup\{\mu_A(y) | y \in \bigcup Md(x)\} = \sup\{\mu_A(y) | y \in \bigcap Md(x)\}$ 。

这里仅采用第三种覆盖粗糙模糊集模型,以下若不加说明, $(\underline{C}(A), \overline{C}(A))$ 均指第三种覆盖粗糙模糊集。

3 覆盖粗糙模糊集的不确定性度量

文献[16]中曾经介绍过覆盖近似空间的层次模型,给出了覆盖粗细关系的定义。

定义 6^[16] 设 C_1, C_2 为论域 U 上的两个覆盖, $x \in U$,若对于任意 $K' \in Md_{C_1}(x)$,始终存在 $K'' \in Md_{C_2}(x)$,使得 $K'' \subseteq K'$,则称 C_2 较 C_1 细,记为 $C_2 \leq C_1$ 。若 $C_2 \leq C_1$ 且 $C_1 \leq C_2$,则称 C_1 与 C_2 相等,记为 $C_1 = C_2$;否则称 C_1 与 C_2 不等,记为 $C_1 \neq C_2$ 。若 $C_2 \leq C_1$ 且 $C_1 \neq C_2$,则称 C_2 较 C_1 严格细,记为 $C_2 < C_1$ 。

归纳现有覆盖近似空间下的粗糙集的不确定性度量方法,可以发现它们均满足统一的度量基本准则,总结如下:

定义 7 设 (U, C) 是一个覆盖近似空间,对于任意 $A \in F(U)$, $(\underline{C}(A), \overline{C}(A))$ 是 A 在 U 上的粗糙模糊集,若映射 $\epsilon: (\underline{C}(A), \overline{C}(A)) \rightarrow R$ 满足:

(1)非负性: $\epsilon(\underline{C}(A), \overline{C}(A)) \geq 0$;

(2)不变性:设 C_1, C_2 为论域 U 上的两个覆盖,若 $C_1 = C_2$,则 $\epsilon(\underline{C}_1(A), \overline{C}_1(A)) = \epsilon(\underline{C}_2(A), \overline{C}_2(A))$;

(3)单调性:设 C_1, C_2 为论域 U 上的两个覆盖,若 $C_2 \leq C_1$,则 $\epsilon(\underline{C}_2(A), \overline{C}_2(A)) \leq \epsilon(\underline{C}_1(A), \overline{C}_1(A))$,称 ϵ 满足不确定性度量的基本准则。

这里若 $\epsilon(\underline{C}(A), \overline{C}(A))$ 越小,则 A 的不确定性越小。

覆盖近似空间下粗糙模糊集的不确定性度量实际上包含两方面:一方面是覆盖近似空间下粗糙集本身概念的不确定性,即粗糙性;另一方面是来自所给的模糊集的不确定性,即模糊性。模糊集的截集是联系模糊集和普通集合的桥梁,借助于截集的概念可以将这一类粗糙模糊集转化成一般粗糙集,从而定义其不确定性。文献[6, 7]曾提出两种度量方法——粗糙度和粗糙熵。

定义 8^[7] 设 (U, C) 是一个最简覆盖近似空间, $A \in F(U)$, $0 \leq \beta \leq \alpha < 1$,则 A 在 (U, C) 中的粗糙度为:

$$\rho_C^{\alpha, \beta}(A) = 1 - \frac{|\underline{C}(A)_\alpha|}{|\overline{C}(A)_\beta|}$$

这里 $\underline{C}(A)_\alpha = \{x | \mu_{\underline{C}(A)}(x) \geq \alpha, x \in U\}$, $\overline{C}(A)_\beta = \{x | \mu_{\overline{C}(A)}(x) \geq \beta, x \in U\}$ 。

定理 1 A 在 (U, C) 中的粗糙度 $\rho_C^{\alpha, \beta}(A)$ 满足不确定性度量基本准则。

证明:非负性和不变性显然成立。下面证明单调性。

若 $C_2 \leq C_1$,则对于任意 $K' \in Md_{C_1}(x)$,都存在 $K'' \in Md_{C_2}(x)$,使得 $K'' \subseteq K'$,故对任意 $x \in U$, $\mu_{C_2(A)}(x) \geq \mu_{C_1(A)}(x)$, $\mu_{\overline{C}_2(A)}(x) \leq \mu_{\overline{C}_1(A)}(x)$,因此 $\rho_{C_2}^{\alpha, \beta}(A) \leq \rho_{C_1}^{\alpha, \beta}(A)$ 。定理证毕。

文献[8]在研究知识的粗糙性时,提出了知识的粗糙熵概念。

定义 9^[7] 设 (U, C) 是一个最简覆盖近似空间, $C = \{K_1, K_2, \dots, K_m\}$,定义知识的粗糙熵为:

$$E(C) = \sum_{i=1}^m \frac{|K_i|}{m} \log_2 |K_i|$$

知识粗糙熵在具有偏序关系的覆盖近似空间下同样具有偏序关系,因此它能够定量反映知识粒度的粗糙性。文献[7]将粗糙度与知识的粗糙熵结合,从粗糙性和模糊性两方面定义了覆盖粗糙模糊集的粗糙熵。

定义 10^[7] 设 (U, C) 是一个最简覆盖近似空间, $A \in F(U)$, $0 \leq \beta \leq \alpha < 1$,则 A 在 (U, C) 中的粗糙熵定义为:

$$E_C^{\alpha, \beta}(A) = \rho_C^{\alpha, \beta}(A) \cdot E(C)$$

由定理 1 和定义 9 容易证明定理 2。

定理 2 A 在 (U, C) 中的粗糙熵 $E_C^{\alpha, \beta}(A)$ 满足不确定性度量基本准则。

粗糙熵比粗糙度能更为准确地度量覆盖近似空间中粗糙模糊集的不确定性,但它仍然有一定的不合理性。

例 1 设 $U = \{x_1, x_2, x_3, x_4, x_5\}$, $C_1 = \{\{x_1, x_2\}, \{x_2, x_3\}, \{x_4, x_5\}\}$, $C_2 = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4, x_5\}\}$,则 (U, C_1) , (U, C_2) 均是最简覆盖近似空间,且 $C_2 < C_1$ 。设 $A = \frac{0.8}{x_1} + \frac{0.7}{x_2} + \frac{0.5}{x_3} + \frac{0.3}{x_4} + \frac{0.2}{x_5}$ 是 U 上的模糊集,则由定义 5 可求得 A 的上、下近似集:

$$\underline{C}_1(A) = \frac{0.7}{x_1} + \frac{0.7}{x_2} + \frac{0.5}{x_3} + \frac{0.2}{x_4} + \frac{0.2}{x_5}$$

$$\overline{C_1}(A) = \frac{0.8}{x_1} + \frac{0.7}{x_2} + \frac{0.7}{x_3} + \frac{0.3}{x_4} + \frac{0.3}{x_5}$$

$$\underline{C_2}(A) = \frac{0.8}{x_1} + \frac{0.7}{x_2} + \frac{0.5}{x_3} + \frac{0.2}{x_4} + \frac{0.2}{x_5}$$

$$\overline{C_2}(A) = \frac{0.8}{x_1} + \frac{0.7}{x_2} + \frac{0.5}{x_3} + \frac{0.3}{x_4} + \frac{0.3}{x_5}$$

显然 $\underline{C_1}(A) \leq \underline{C_2}(A)$, $\overline{C_1}(A) \geq \overline{C_2}(A)$, 从直观的认知规律来看, A 在两个覆盖近似空间中的不确定性应该不等。但按照定义 8, $\rho_{C_1}^{0.8, 0.3}(A) = \rho_{C_2}^{0.8, 0.3}(A) = 0.8$, 说明此时粗糙度无法准确体现由于边界的存在而引起的不确定性。而 $\rho_{C_1}^{0.4, 0.1}(A) = \rho_{C_2}^{0.4, 0.1}(A) = 0$, 故 $E_{C_1}^{0.4, 0.1}(A) = E_{C_2}^{0.4, 0.1}(A) = 0$, 说明此时也无法衡量知识的粗糙性。这是因为当截集的阈值小于某一值时, 下近似集与上近似集元素个数比例固定, 从而导致粗糙度或者粗糙熵无法随着空间的细化而减小。因此, 虽然粗糙度和粗糙熵都能符合覆盖近似空间下粗糙模糊集的不确定性度量的基本准则, 但某些具体情形下可能刻画得仍然不够, 不能符合一般的认知规律。

事实上, 覆盖近似空间下的粗糙模糊集的不确定性既与覆盖的粗糙程度有关, 也与所取的模糊集的模糊程度有关系。为了更加准确地体现这两种不确定性, 本文提出更加严格的不确定性度量准则:

定义 11 设 (U, C) 是一个覆盖近似空间, 对于任意 $A \in F(U)$, $(\underline{C}(A), \overline{C}(A))$ 是 A 在 U 上的粗糙模糊集, 若映射 $\epsilon': (\underline{C}(A), \overline{C}(A)) \rightarrow \mathbb{R}$ 满足:

(1) 非负性与定义 7(1) 一致;

(2) 不变性与定义 7(2) 一致;

(3) 严格单调性: 设 C_1, C_2 为论域 U 上的两个覆盖, 若 $C_2 < C_1$, 则

$$\epsilon'(\underline{C_2}(A), \overline{C_2}(A)) < \epsilon'(\underline{C_1}(A), \overline{C_1}(A))$$

称 ϵ' 满足不确定性度量修正准则。

由于覆盖近似空间下粗糙模糊集本身在构造上、下近似集时已经考虑了覆盖特性, 所得的上、下近似集也均是模糊集, 因此可以通过比较上、下近似集的隶属度与原模糊集的隶属度之间的差异, 得出粗糙模糊集的修正粗糙度。

定义 12 设 (U, C) 是一个覆盖近似空间, $A \in F(U)$, A 的上、下近似集分别记为 $\overline{C}(A)$, $\underline{C}(A)$, 则 A 在 (U, C) 中的修正粗糙度为:

$$\rho_C(A) = \frac{1}{2} \sum_{j=1}^n \left\{ \frac{\mu_A(x_j) - \mu_{\underline{C}(A)}(x_j)}{\sum_{j=1}^n \mu_A(x_j)} + \frac{\mu_{\overline{C}(A)}(x_j) - \mu_A(x_j)}{\sum_{j=1}^n [1 - \mu_A(x_j)]} \right\}$$

从以上定义可知, 粗糙模糊集的修正粗糙度既涵盖了正域模糊集的相对粗糙度, 也涵盖了负域模糊集的相对粗糙度, 因而更能体现粗糙模糊集在覆盖近似空间下的不确定度量。

定理 3 A 在 (U, C) 中的修正粗糙度 $\rho_C(A)$ 满足不确定性度量修正准则。

证明: 非负性、不变性显然成立, 以下只证明严格单调性。

若 $C_2 < C_1$, 则对于任意的 $K' \in Md_{C_1}(x)$, 都存在 $K'' \in Md_{C_2}(x)$, 使得 $K'' \subset K'$, 故对任意 $x \in U$, $\mu_{C_2(A)}(x) \geq \mu_{C_1(A)}(x)$, $\mu_{\overline{C_2}(A)}(x) \leq \mu_{\overline{C_1}(A)}(x)$, 且等号并不总是成立。于是

$$\mu_A(x) - \mu_{C_2(A)}(x) \leq \mu_A(x) - \mu_{C_1(A)}(x)$$

$$\mu_{\overline{C_2}(A)}(x) - \mu_A(x) \leq \mu_{\overline{C_1}(A)}(x) - \mu_A(x)$$

成立, 因此 $\rho_{C_2}(A) < \rho_{C_1}(A)$ 。定理证毕。

续例 1 容易计算 $\rho_{C_1}(A) = 0.1$, $\rho_{C_2}(A) = 0.04$, $\rho_{C_2}(A)$

$< \rho_{C_1}(A)$ 符合认知规律。

定理 4 当 C 退化成论域 U 上的一个划分且 A 是一个普通集合时, 上述修正粗糙度退化为文献[17]提出的修正粗糙度。

证明: 因为当 $C = \{K_1, K_2, \dots, K_m\}$ 为 U 上的一个划分且 A 是一个普通集合时, 对任意 $x \in U$,

$$\mu_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

此时 A 的下近似集 $\underline{A}$ 满足: 对任意 $x \in U$,

$$\mu_{\underline{A}}(x) = \begin{cases} 1, & \exists K_i, \text{ s. t. } x \in K_i \subset A \\ 0, & \text{其它} \end{cases}$$

A 的上近似集 $\overline{A}$ 满足对任意 $x \in U$,

$$\mu_{\overline{A}}(x) = \begin{cases} 1, & \exists K_i, \text{ s. t. } x \in K_i \cap A \neq \emptyset \\ 0, & \text{其它} \end{cases}$$

这里 $\sum_{j=1}^n \mu_A(x_j) = |A|$, $\sum_{j=1}^n [1 - \mu_A(x_j)] = |U - A|$, $\sum_{i=1}^n [\mu_{\underline{A}}(x_i) - \mu_{\underline{C}(A)}(x_i)] = |A - \underline{A}|$, $\sum_{i=1}^n [\mu_{\overline{C}(A)}(x_i) - \mu_{\overline{A}}(x_i)] = |\overline{A} - A|$, 因

此 $\rho_C(A) = \frac{1}{2} \left(\frac{|A - \underline{A}|}{|A|} + \frac{|\overline{A} - A|}{|U - A|} \right)$ 。

这恰好是文献[17]提出的修正粗糙度, 符合粗糙集不确定性度量的扩展准则。定理证毕。

由此可见, 本文提出的覆盖近似空间下粗糙模糊集的修正粗糙度是一般粗糙集的粗糙度的推广。由于直接考虑了模糊性的差异, 因此它更符合实际问题中对于不确定性的度量。

结束语 通过对现有覆盖近似空间下的粗糙模糊集不确定性度量进行分析, 发现其不合理性, 即当阈值小于某一值时, 由于上、下近似集的元素个数比例固定, 导致原有的粗糙度或粗糙熵无法随着空间的细化而减小。本文提出不确定性度量修正准则, 并借助上、下近似集隶属度和原模糊集隶属度之间的差异, 给出修正粗糙度的概念, 举例证明该方法更加符合实际问题的需要, 为覆盖近似空间下粗糙模糊集的不确定性度量提供了新的手段。

参考文献

- [1] 王国胤, 张清华, 马希骞, 等. 知识不确定性问题的粒计算模型[J]. 软件学报, 2011, 22(4): 676-694
- [2] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991
- [3] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8: 338-353
- [4] 胡军. 覆盖近似空间中粗糙集的不确定性度量研究[J]. 计算机应用与软件, 2011, 8(11): 180-183
- [5] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets[J]. International Journal of General Systems, 1990, 17(2): 191-209
- [6] 王丽娟, 吴陈, 杨习贝, 等. 邻域系统粗糙集和覆盖粗糙集[J]. 计算机科学, 2013, 40(1): 221-224
- [7] 徐菲菲, 苗夺谦, 李道国, 等. 基于覆盖的粗糙模糊集的粗糙熵[J]. 计算机科学, 2006, 33(10): 179-181
- [8] 黄兵, 何新, 周献中. 基于广义粗糙覆盖约简的粗糙熵[J]. 软件学报, 2004, 15(2): 215-220
- [9] 余美真, 李进金. 覆盖粗糙模糊集的不确定性度量[J]. 模糊系统与数学, 2009, 23(6): 145-148
- [10] 郭增晓, 米据生. 粗糙模糊集的模糊性度量[J]. 模糊系统与数

- [11] Bonikowski Z, Bryniarski E, Wybraniec U. Extensions and intentions in the rough set theory[J]. Information Sciences, 1998, 107(1-4):149-167
- [12] Zhu W, Wang F Y. Reduction and axiomization of covering generalized rough sets[J]. Information Sciences, 2003, 152:217-230
- [13] 魏莱, 苗夺谦, 徐菲菲, 等. 基于覆盖的粗糙模糊集模型研究[J]. 计算机研究与发展, 2006, 43(10):1719-1723

(上接第 251 页)

结束语 Web 文档数据无处不在, 数据质量评估是获得高质量 Web 文档的关键, 从而保证数据分析产生有意义的结果。本文提出了一种基于事实提取的 Web 文档内容数据质量评估方法。该方法的优越性在于: (1) 不依赖于某类特征, 根据从内容中提取的事实来度量准确性和完整性, 实现了基于语义内涵的数据质量评估; (2) 它是一种对数据质量自动化评估的方法。

理论分析和实验表明了本文方法的有效性。今后将对算法的某些环节进行改进。

参考文献

- [1] Aebi D, Perrochon L. Towards improving data quality[C]// Proc. of the international conference on information systems and management of data. New York: ACM, 1993:273-281
- [2] 马茜, 古峪, 张天成, 等. 一种基于数据质量的多源多模态感知数据获取方法[J]. 计算机学报, 2013, 36(10):2010-2131
- [3] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(1):2076-2082
- [4] Pernici B, Scannapieco M. Data Quality in Web Information Systems[C]// Proc. of the 21st International Conference on Conceptual Modeling. Berlin Heidelberg: Springer, 2002:397-413
- [5] Dalip D H, Cristo M, Calado P. Automatic assessment of document quality in web collaborative digital libraries [J]. ACM Journal of Data and Information Quality, 2011, 2(3):14
- [6] Hu Mei-qun, Lim Ee-peng, Sun Ai-xin. Measuring Article Quality in Wikipedia: Models and Evaluation[C]// Proc. of the 16th CIKM. New York: ACM, 2007:243-252
- [7] Zeng H, Alhossaini M A, Li D, et al. Computing trust from revision history[C]// Proc. of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services. New York: ACM, 2006
- [8] Blumenstock J E. Size Matters: Word Count as a Measure of Quality on Wikipedia[C]// Proc. of the 17th International Conference on World Wide Web. New York: ACM, 2008:1095-1096
- [9] Knap T, Mlynkova I. Quality Assessment Social Networks: A Novel Approach for Assessing the Quality of Information on the Web[C]// Proc. of QDB of VLDB'10. 2010
- [10] Baeza-Yates R, Rello L. On Measuring the Lexical Quality of the Web[C]// Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. New York: ACM, 2012:1-6
- [11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3):993-1022
- [12] Dong Xin, Berti-Equille L, Hu Yi-fan, et al. Global Detection of Complex Copying Relationships Between Sources[C]// Proc. of

- [14] 徐忠印, 廖家奇. 基于覆盖的模糊粗糙集模型[J]. 模糊系统与数学, 2006, 20(3):141-144
- [15] 胡军, 王国胤, 张清华. 一种覆盖粗糙模糊集模型[J]. 软件学报, 2010, 21(5):968-977
- [16] 胡军, 王国胤. 覆盖粒度空间的层次模型[J]. 南京大学学报: 自然科学版, 2008, 44(5):551-558
- [17] 胡军, 王国胤. 粗糙集的不确定性度量准则[J]. 模式识别与人工智能, 2010, 23(5):606-615

- VLDB Endowment. New York: VLDB Endowment, 2010:1358-1369
- [13] Fan Wen-fei. Dependencies Revisited for Improving Data Quality [C]// Proc. of PODS 2008. New York: ACM, 2008:159-170
- [14] Rassbach L, Pincock T, Mingus B. Exploring the Feasibility of Automatically Rating Online Article Quality[EB/OL]. [2013-9-10]. <http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf>
- [15] Liu Jun, Ram S. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality[J]. ACM Transactions on Management Information Systems, 2011, 2(2):1-23
- [16] Han Jing-yu, Wang Chuan-dong, Jiang Da-wei. Probabilistic Quality Assessment Based on Article's Revision History[C]// Proc. of the 22nd International Conference on Database and Expert systems Applications (DEXA). Berlin Heidelberg: Springer, 2011:574-588
- [17] Dalvi N, Kumar R, Soliman M. Automatic Wrappers for Large Scale Web Extraction[C]// Proc. of the 37th International Conference on Very Large Databases. New York: VLDB Endowment, 2011:219-230
- [18] 肖升, 何炎祥. 基于动词论元结构的中文事件抽取方法[J]. 计算机科学, 2012, 39(5):161-164
- [19] 杨少华, 林海略, 韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报, 2008, 19(2):209-223
- [20] Etzioni O, Fader A, Christensen J, et al. Open information extraction: the second generation[C]// Proc. of 22nd International Joint Conference on Artificial Intelligence. California: AAI press, 2011:3-10
- [21] Simões G, Galhardas H, Gravano L. When Speed Has a Price: Fast Information Extraction Using Approximate Algorithms[C]// Proc. of the 39th International Conference on Very Large Databases. New York: VLDB Endowment, 2013:1462-1473
- [22] Mays E, Damerau F J, Mercer R L. Context based spelling correction [J]. Information processing and management, 1991, 27(5):517-522
- [23] Si X, Chang E Y, Gyongyi Z, et al. Confucius and its intelligent disciples: integrating social with search[C]// Proc. of the 36th International Conference on Very Large Databases. New York: VLDB Endowment, 2010:1505-1516
- [24] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]// Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2003:173-180