# Rough Set Theory for Supporting Decision Making on Relevance in Browsing Multilingual Digital Resources

**Jolanta Mizera-Pietraszko and Jolanta Tancula**

**Abstract** Browsing digital library (DL) collections seems to pose a challenge for a user owning to the number of factors like for instance, operability of the system, interface readability or clarity, and retrieval efficiency directly related to it, or the number of digital items within the user's domain. However, when it comes to searching for an item in a foreign language to the user, the number of the factors arises even more which translates proportionally to the growing number of clicks aimed to retrieve the target item. Such a procedure usually leads to disheartening the user from browsing the digital collections. Our study into the user's behavior interacting with multilingual DL system is set out to propose a rough set theory based model which automatically generates a decision rule based on the minimum number of the decision factors. Analyzed is a set of the predefined factors specifically influencing the user's decision on clicking an item of his special interest. We aim to limit the number of the factors however, without losing the precision of the final user's decision. To our best knowledge, rough set theory has not been implemented for multilingual decision making purposes.

**Keywords** Rough set theory · Decision support systems · Multilingual digital libraries · Information retrieval

J. Mizera-Pietraszko (✉) · J. Tancula
Institute of Mathematics and Computer Science, Opole University,
Oleska 48, 45-052 Opole, Poland
e-mail: jmizera@math.uni.opole.pl

J. Tancula
e-mail: jtancula@math.uni.opole.pl



D. Król et al. (eds.), *Advanced Topics in Intelligent Information and Database Systems*, Studies in Computational Intelligence 710,
DOI 10.1007/978-3-319-56660-3_12

# 1 Introduction

Decision making support is becoming the more challenging task in case of uncertainty and incompleteness of information needed to take an action whether in industry, office or academia. We apply rough set theory to support the user's decision making.

Rough set theory created by Pawlak [1] captures a concept of a finite set with a lower and upper approximation defined by a human. The lower approximation of the concept is a set of the attributes sufficient to make the reliable decision about the object while the upper approximation is a finite set of other attributes which can be classified with some probability to the set (fuzzy set) allowing to support the same decision making on the comparison basis [2]. Both the upper and lower approximation, which are non-empty sets, can be called rough sets with imprecise boundary region, on the contrary to crisp sets, when the boundary region is precise.

We define the set approximations to determine the minimum number of factors allowing the user to make the decision on the selection of the full-text items from multilingual digital resources. As browsing digital collections profiles the information need, making decision on which of the numerous digital items is relevant or not, without taking into consideration the selection criteria, is a challenge. Such a motivation justifies our approach to applying rough set theory to support the user in making the crucial decision on clicking the full-text items available for download.

The remainder of this paper is structured as follows: Sect. 2 introduces the literature overview in the field of the rough set theory, section three presents our decision table of supporting decision on relevance while browsing multilingual digital resources, other sections discuss methodology of creating fuzzy sets and reduct sets from the multilingual project real data to support decision making on relevance. As the last part of this paper, the conclusion remarks are made. We plan to extend this project.

## *1.1 Fuzzy Idea About the User's Need*

Relevance does not refer to the information retrieval system, which makes a kind of guesses while building the ranking list of the responses to the particular query, but it refers just to the user who expresses his or her information need by clicking on the item which hopefully will be relevant. Boolean model serves as an example. Thus, fuzzy idea about the user's real information need refers not only to the system, but also to the user, however from a different perspective, since while browsing a digital collection, the user models his knowledge depending on so called random walk. Fulfilling the information need by making the shortest path of the random walk is a key problem widely studied in the area of information retrieval. Hoenkamp [3]

considers intuitive nature of information need by proposing lattice theory to formalize the notion with the aim at enabling the user to express the need as independent of a query and of a language model, even when the relevant results to such a query do not exist. Analysis of information need based on geo-localization positioning systems integrated with the user's profile while browsing cultural heritage digital resources is discussed from the context-aware perspective of the benefits [4]. Chronological recommendation of highly-cited papers assumes that the initial information need evolves with the time progressing while browsing research papers. The process can be modeled twofold: dynamic ranking feature construction and dynamic evolving feature weight. Some experimental studies with the ACM corpus reveal the recommendation of the highly-cited papers can be enhanced by time-series ranking [5]. Another approach relies on bound with query versus without it to measure the user's need domain. Making decision on relevance can improve precision of the information [6].

The information need evolves from generic, meaning the user has only heard about something without any knowledge about the subject matter, to more and more specific, by expanding the knowledge during the searching process which then transfers to the higher precision of the information retrieval system.

## *1.2 Conditional Attributes as Criteria in Decision Making Under Uncertainty*

We define the following attributes contributing to support the user's decision making:

- Digital Library Project
- Languages other than English that is usually treated as lingua franca
- Target language of the digital item
- Number of the collection items for each language
- Average number of the query matches
- Number of the collection items within the user's domain
- Average target language competence is accessed by the user within the scale 0–5 according to the Common European Framework of References for Languages proposed by the Council of Europe

All the attributes are computed in the target language collection as we assume that the user who does not feel that the foreign language competence is sufficient to browse the collection, withdraws from the task. The attribute *Average number of the query matches* indicates the extent to which it is worth to undertake the task including the system bandwidth. Also, the attribute *Number of the collection items within the user's domain* has been found essential from the perspective of the language semantics, in particular the user's familiarity of technical terminology.

## 2 Related Work

Application of rough set theory to analysis of relationship between the language competence and accessibility to multilingual digital collections with the support on relevance is the first approach to the problem presented.

In addition to the substantial collection of the original works on rough set theory authored by its creator Pawlak [2, 7, 8], around three hundred thousand works follow his research. According to him, a system is defined as a quadruple $S = (X, A, V, \delta)$, where $X$ is a set of objects with their upper and lower bounds, $A$—their attributes, $V$ is a set of the values of these attributes and $\delta$ is a decisive function, whose values are Yes, or No. Pawlak proposed to express the values of the decision function $\delta$: $X \times A \rightarrow V$ by the means of Boolean model in which discernible binary relations form reduct sets and decision rules [9]. Going further, he introduced topological operations like approximation space $(U, R^{(U)})$ of the universe $U$ of objects and $R^{(U)}$: $U \times U$ being an indiscernibility relation between the attributes [10].

As the objects can be attributed to some data which share the same information, they are indiscernible—the reasoning allows to apply the theory to pattern recognition, or natural language processing [11]. In our work, we approximate the query vagueness language.

Other works authored by the Pawlak's followers discuss granular computing applications like information processing, or decision analysis. The application areas include association rules, concept representation, approximate knowledge, scoring and ranking information retrieval results [12]. Multigranulation of rough sets is widely studied under incomplete environment in decision making, incomplete information or neighborhood systems [13]. Thangavel et al. [14] discuss meta-heuristic algorithms of the rough set theory.

Customer satisfaction analysis, probabilistic decision making or quality of rough approximation in classifying problems are some examples of their application areas [15]. A work by Polkowski [16] dedicated to the rough set theory creator, highlights the partition concept $\{A, X \backslash A\}$, where A is a set of objects and $X \backslash A$ is a problem decidable based on the knowledge. Such a paradigm transfers the incremental knowledge to the partitions belonging to the universe $U$ and consequently making the crisp notion $A$ a non-crip one. Both lower and upper approximations solve this paradigm analyzed by our model in which the user's knowledge is becoming incremental while browsing digital collections.

## 3 Preliminaries

Following a founder of set theory, Georg Cantor, "A set is the result of collecting together certain well-determined objects of our perception or our thinking into a single whole objects called the elements of the set" [1]. As opposed to set theory, where the core concept of set is a collection of entities being either a real objects or

the conceptual entities as its elements, in rough set theory, we assume that having some data we create a set of elements. However, the elements of the set which contain the same information and therefore are similar, create elementary sets $E^{(S)}$. The sum of any elementary sets creates a definable set $S^{(D)} = \sum_{S \in N} E^{(S)}$. On the contrary to it, an undefined set is called a rough set, which can be described by two sets definable by its upper $U^{(A)}$ and lower approximations $L^{(A)}$. The difference between the upper $U^{(A)}$ and lower $L^{(A)}$ approximation is called the set boundary $S^{(B)} = U^{(A)} - L^{(A)}$. The set is called a rough set $R$ if and only if the edge of the set $E^{(S)}$ is a non-empty set $E^{(S)} \neq \{\emptyset\}$ such that $R = \{r_n : r_i \in E^{(S)}, E^{(S)} \neq \emptyset, L^{(A)} \subseteq S^{(B)} \subseteq U^{(A)}, 1 \leq i \leq n \in N\}$. Rough set theory has been applied in many fields. In this paper, we apply this theory to approximate an access to digital collections of multilingual library systems depending upon the user's competence in the target language different from English.

## 3.1 Decision System Grounded upon Rough Sets

Information system is a multilevel structure for recording and storing data that enables to construct the models which describe some processes being analyzed. In this section we are going to focus on decision support systems grounded upon the rough sets. Let us first introduce a formal definition of information system.

**Definition 1** Let $S$ be an information set such that $S = (U, A)$, where $U$ is a non-empty finite set of objects, $A$ is a non-empty finite set of attributes, $a \in A$ an $a: U \to V_a$ and $V_a$ is a domain of attribute $a$. In addition, let's denote set B: $B \subseteq A$ as a vector of information for object $x \in U$.

## 3.2 Decision Table on Supporting Relevance

Data is represented in many ways. We create decision table which is called information table built from any arbitrary data. Then, the data is divided into attributes called conditional attributes and decision attributes called decisions.

**Definition 2** Decision table is a set denoted by $S = (U, A \cup \{dec\})$ where U is a set of decision objects $U = \{u_1, \ldots, u_n\}$ called the universe and A is a set of attributes denoted by $a_i: U \to V_i$, where d is a decision $dec = \{d_1, \ldots, d_n\}$.

Table 1 serves as an example of our decision table created based on multilingual digital Project Gutenberg.

In Table 1 we define a set of decision objects $U = \{u_1, u_2, \ldots, u_{15}\}$, where $u_1$ = "Chinese", $u_2$ = "Danish",…, $u_{15}$ = "Tagalog", a set of conditional attributes $A = \{A^{(1)}, A^{(3)}, A^{(4)}\}$, where $A^{(1)}$ = "Number of the collection items", $A^{(3)}$= "Number of the collection items within the user's domain" and

**Table 1** Decision table of relevance in browsing multilingual Project Gutenberg

| Language of the collection | Number of the collection items $A^{(1)}$ | Average number of matches $A^{(2)}$ | Number of the items within the user's domain $A^{(3)}$ | Target language competence $A^{(4)}$ | Relevance $A^{(5)}$ |
|---|---|---|---|---|---|
| Chinese | 958 | 61 | 71 | 0 | Y |
| Danish | 98 | 22 | 12 | 1 | N |
| Dutch | 1260 | 45 | 214 | 2 | Y |
| Esperanto | 286 | 59 | 54 | 0 | Y |
| Finish | 1487 | 79 | 211 | 1 | N |
| French | 2079 | 62 | 268 | 4 | Y |
| German | 1785 | 89 | 311 | 5 | Y |
| Greek | 653 | 27 | 48 | 2 | Y |
| Hungarian | 206 | 11 | 17 | 1 | N |
| Italian | 1168 | 72 | 63 | 3 | Y |
| Latin | 342 | 46 | 32 | 1 | N |
| Portuguese | 786 | 89 | 18 | 1 | Y |
| Spanish | 705 | 127 | 26 | 3 | Y |
| Swedish | 265 | 35 | 15 | 3 | Y |
| Tagalog | 71 | 12 | 3 | 0 | N |

$A^{(4)}$ = "Target language competence" and one-element set of decision attribute is $A^{(5)}$ = "Relevance".

The objects in Table 1 are divided into classes related to the user's final decision on whether or not to click on the digital item based upon the attributes considered $Rel^{(Y)} = \{1, 3, 4, 6, 7, 8, 10, 12, 13, 14\}$ (positive decision) and $Rel^{(N)} = \{2, 5, 9, 11, 15\}$ (negative decision).

Figure 1 shows relationships between some attributes for each of the Project Gutenberg collection languages. The collections with the greatest number of the full-text digital items, specifically those within the user's domain which determine the users' positive decision, are in German perhaps because the project is German,. Still, the most dominant attribute is the user's competence in the target language, since otherwise browsing the collection seems counterproductive. The attribute marked in red indicates the system's precision.

## 4 Indiscernible Sets

In this section defined are indiscernible sets of objects in relation [17] to our example in Fig. 1.

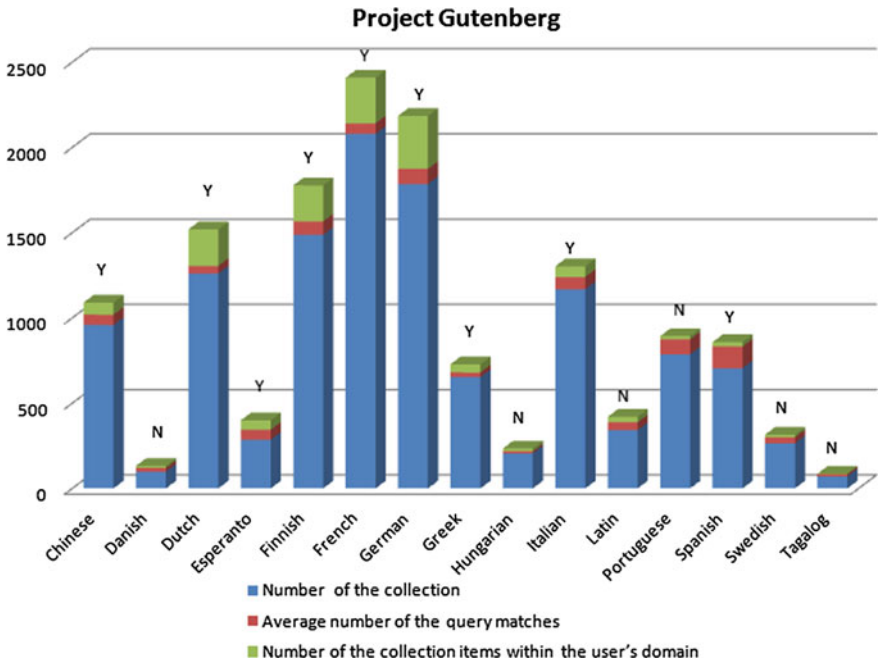


**Fig. 1** Object-related comparative stacked column chart of some attributes

**Definition 3** Objects $x$ and $y$ belonging to $U$ and a set of attributes $B \subseteq A$ are called discernible if and only if there exists such an $a \in B$ that $a(x) \neq a(y)$, otherwise $x$ and $y$ are indiscernible such that $a(x) = a(y)$.

Thus, the decision support system can be denoted

$$IND_S(B) = \{x, y \in U \times U | \bigvee_{a \in B} a(x) = a(y)\} \quad (1)$$

Indiscernibility relation $IND(B)$ is said to be an equivalence relation or alternatively reflexive relation, when an object is in relation with itself $xIND(B)x$, symmetric (if $xIND(B)y$ then $yIND(B)x$) and transitive (if $xIND(B)y$ and $yIND(B)z$ then $xIND(B)z$).

With reference to our example, it is for attribute A1, then A1 and A2 and finally A1, A2 and A3

$$IND(\{A1\}) = \{1, 4, 8, 11, 12, 13\}$$
$$IND(\{A1, A2\}) = \{\{1, 4, 12, 13\}_{SD}, \{2, 14\}_{MS}, \{5, 6, 7, 10\}_{DD}, \{9, 15\}_{MM}, \{8, 11\}_{SS}\}$$
$$IND(\{A1, A2, A3\})$$
$$= \{\{1, 4\}_{SDS}, \{2, 14\}_{MSM}, \{5, 6, 7\}_{DDD}, \{8, 11\}_{SSS}, \{9, 15\}_{MMM}, \{12, 13\}_{SDM}\}$$

The types of the relations in our example refer to their attributes described in Table 1.

### 4.1 Approximation of the Set Boundaries

A set $X$ of objects described by the attributes $A$ can be defined by a lower (positive region) or upper approximation (negative region) according to the following formula

$$\underline{B}X = \{m|[m]_B \subseteq X\} \text{ or } \overline{B}X = \{m|[m]_B \cap X \neq \varnothing\} \quad (2)$$

where $[m]_B$ is a set of indiscernible objects. On the contrary to boundary region, which does not allow to associate unambiguously the attributes to set X, the inside region does. In case of non-empty edge, the set is called an approximate, otherwise it is called a crisp edge.

**Definition 4** A set $X$ is called a rough set when $BN_B(X)$ is a non-empty set meaning $BN_B(X) = \overline{B}(X) - \underline{B}(X) \neq 0$.

The upper bound of our set is then $B = \{A1, A2, A3\}$ for the class of negative decision on clicking the item by the user $Rel^{(N)} = \{2, 5, 9, 11, 15\}$ such that $\underline{B}(X) = \{2, 14, 9, 15\}$, and consequently $\overline{B}(X) = \{2, \underline{8}, 9, 11, 12, \underline{13}, 14, 15\}$.

Since $BN_B(X) = \overline{B}(X) - \underline{B}(X) = \{8, 13\}$ is not empty, we have here a rough set.

In Fig. 2, the area in dark orange denoted with $U$ is the universe, the area in light orange is the upper bound $\overline{B}(X)$, the area in yellow is the lower bound $\underline{B}(X)$, whereas the area in blue is our set $X$.

### 4.2 Precision of Approximation

Both crisp and approximate sets are comparable in terms of their power based on coefficient of approximation precision.

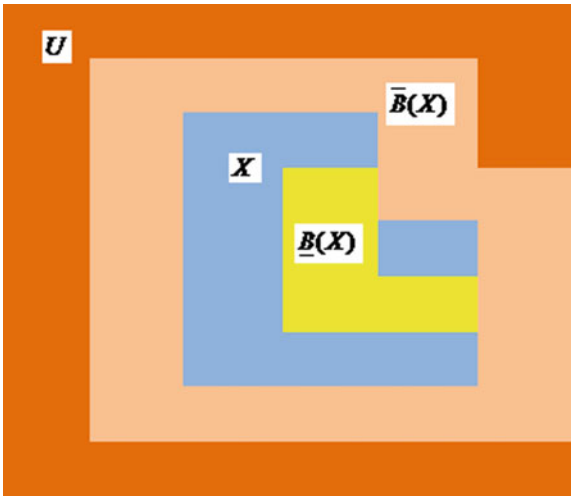


**Fig. 2** Graphical representation of approximation of our set X

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \text{ where } 0 \leq \alpha_B \leq 1 \tag{3}$$

For a crisp set it is $\alpha_B(X) = 1$, otherwise $X$ is a rough set $0 \leq \alpha_B(X) < 1$.

For the attributes in Table 1, coefficient of approximation precision is $\alpha_B(X) = \frac{4}{8} = \frac{1}{2}$. This coefficient $\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|}$ denotes a power for the set class of $Rel^{(N)} = \{2, 5, 9, 11, 15\}$.

## 5 Reduct Sets in Decision Making

In a decision support process not all the attributes are necessary to make a decision. Reduct sets or simply reducts RED(S) are the granular subsets that enable to define the characteristics of the whole set of its attributes. One of such reducts in decision table is called a decision reduct.

**Definition 5** A set of attributes is called a reduct when for any of two objects $x$ and $y$ satisfied is a condition $a(x) \neq a(y)$ and both $x$ and $y$ are discernible by $A$ and $B$ simultaneously on condition that $B$ is a granular discernible set.

**Definition 6** Core is a set of attributes belonging to each of the reducts and denoted by

$$COR(B) = \bigcap_{B \in RED(S)} B \tag{4}$$

For computing the reducts, we are going to apply Boolean algebra and Boolean functions. Following the data in Table 1, we create a discernible matrix.

Using the data from Table 2, we create a Boolean function *f(B)*

$$f(B) = (A1)(A2)(A3)(A4)(A1 \cup A2)(A1 \cup A3)(A1 \cup A2 \cup A3)(A1 \cup A3 \cup A4)$$
$$(A1 \cup A2 \cup A4)(A2 \cup A3 \cup A4)(A1 \cup A2 \cup A3 \cup A4)$$

then we apply an absorption rule such that $p \wedge (p \vee q) \equiv p$, so we get

$$f(B) = (A1)(A2)(A3)(A4)$$

following, we use conjunction rule with respect to alternative $p \wedge (q \vee r) = p \wedge q \vee p \wedge r$ like here $f(B) = (A1A2A3A4)$, so we get one reduct only $R_1 = \{A1, A2, A3, A4\}$ On applying the reduct C = X − R1 = {4, 8, 13} we get our reduct set presented in Table 3.

Reduction decision table allows to narrow down the set of searching to a few sets of objects only and associated to them attributes, as it is shown in Table 3. Our approach proves to produce quite promising results in the support of decision

**Table 2** Table of discernible set X

| F | 2 | 9 | 11 | 12 | 14 | 15 |
|---|---|---|---|---|---|---|
| 1 | A1, A2, A3 | A1, A2, A3 | A2 | A3 | A1, A2, A3, A4 | A1, A2, A3 |
| 3 | A1, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A3 | A1, A2, A3, A4 |
| 4 | A1, A2, A3 | A1, A2, A3 | A3 | A1 | A1 | A1, A2, A3 |
| 5 | A1, A2, A3 | A1, A2, A3 | A1, A2, A3 | A1, A3 | A1, A2, A3, A4 | A1, A2, A3 |
| 6 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 |
| 7 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A3, A4 | A1, A2, A3, A4 | A1, A2, A3, A4 |
| 8 | A1, A3, A4 | A1, A2, A3, A4 | – | A2, A3, A4 | A1, A3 | A1, A2, A3, A4 |
| 10 | A1, A2, A3, A4 | A1, A2, A3, A4 | A1, A2, A4 | A1, A3, A4 | A1, A2, A3 | A1, A2, A3, A4 |
| 13 | A1, A2, A4 | A1, A2, A4 | A2, A3, A4 | A4 | A1, A2 | A1, A2, A4 |

**Table 3** Reduction decision table for Project Gutenberg digital library

| Language of the collection | Number of the collection items $A^{(1)}$ | Average number of matches $A^{(2)}$ | Number of the items within the user's domain $A^{(3)}$ | Target language competence $A^{(4)}$ | Relevance $A^{(5)}$ |
|---|---|---|---|---|---|
| Esperanto = 4 | 286 | 59 | 54 | 0 | y |
| Greek = 8 | 653 | 27 | 48 | 2 | y |
| Spanish = 13 | 705 | 127 | 26 | 3 | y |

making on how far the user can limit the searching by reducing the number of sets with no impact on the user's decision accuracy of the relevance.

## 6 Conclusion

Our model shows that Boolean reasoning integrated with decision tables built on the basis of rough set theory allows to determine mutual relationships between the attributes such as language competence, the number of the items in a target language and the system efficiency, expressed as the average number of matches.

In our further work, we plan to extend the model to some more attributes related to searching for multilingual items, but especially we want to add different weights to the attributes with the aim to study their influence on the multilingual ranking list.

## References


1. Inuiguchi, M., Hirano, S., Tsumoto, S.: Studies in Fuzziness and Soft Computing: Rough Set Theory and Granular Computing. Springer, Berlin (2003)
2. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**, 341—356 (1982)
3. Hoenkamp, E.: On the notion of an information need, In: Hoenkamp, E. (ed.) Advances in Information Retrieval Theory, In: 2nd International Conference on the Theory of Information Retrieval, ICTIR 2009, LNCS, vol. 5766, pp. 354–356, Springer (2009)
4. Jailani, A.K., Kusakabe, S., Araki, K.: Adaptive contex-awareness model for cultural haritage information based on user needs. In: 4th International Congress on Advances in Applied Informatics, Okayma, Japan, pp. 339–342. IEEE Computer Society (2015)
5. Jiang, Z., Xiaozhong, L., Liangcai, G.: Chronological citation recommendation with information-need shifting, In: 24th ACM International Conference on Information and Knowledge Management, pp. 1291–1300. ACM (2015)
6. Wang, B., Gao, G.: Bound on information need in information retrieval, In: Proceedings—2010 International Conference on Intelligent Computing and Cognitive Informatics, ICICCI 2010, Kuala Lumpur, Malaysia, pp. 75–78. IEEE Computer Society (2010)
7. Pawlak, Z.: Rough Sets, Rough Sets & Data Mining, pp. 1–7. Kluwer Academic Publishers (1997)
8. Pawlak, Z.: Rough Sets; Theoretical Aspects of Reasoning About Data. Springer Science & Business, Media BV (1991)
9. Pawlak, Z., Skowron A.: Rough sets and boolean reasoning. Int. J. Inf. Sci. (Elsevier) **177**, 41–73 (2007)
10. Pawlak, Z., Skowron A.: Rough sets; some extensions. Int. J. Inf. Sci. (Elsevier) **177**, 41—73 (2007). Elsevier
11. Pawlak, Z., Skowron A.: Rudiments of rough sets. Int. J. Inf. Sci. (Elsevier) **177**, 28—40 (2007). Elsevier
12. Lin, T.Y., Yao Y.Y., Zadeh L.A.: Data mining, rough sets & Granular computing. In: Kacprzyk, J. (Ed.) Studies in Fuzziness & Soft Computing. Springer (2002)
13. Yang, X., Yang, J.: Incomplete Information Systems & Rough Set Theory Models and Attribute Reductions. Science Press, Springer, Beijing (2012)
14. Thangavel, K., Pethalaksmi, A.: Dimensionality reduction based on rough set theory: a review. Appl. Soft Comput. (Elsevier) **9**, 1–12 (2009)
15. Greco, S., Hata, Y., Hirano, S., Inuiguhi, M., Miyamato, S., Nguyen, H.S., Slowinski, R.: Rough sets & current trends in computing. In: 5th Conference on Rough Sets and Current Trends in Computing (RSTCT 2006), LNAI, vol. 4259. Springer, Heidelberg (2006)
16. Polskowski, L.: Rugh sets; mathematical foundations. In: Kacprzyk, J. (dd.) Advances in Soft Computing, p. 303, Springer (2002)
17. Nguyen, H.S.: Applied Mathematics: Decision Systems. Warsaw University (2011)