# Soft Computing: Dealing with Vagueness in Intelligent Information Retrieval

Anil Sharma [1], Suresh Kumar [2]

[1]*University School of Information, Communication and Technology, Guru Govind Singh Indraprastha University, Delhi, India*

[2]*Department of Computer Science and Engineering, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India*

***Abstract.*** Information Retrieval (IR) deals with uncertain information due to vagueness in query specified by users. Employing soft computing techniques in web based search have enabled the Information Retrieval Systems with the capability to deal with imprecise and vague knowledge. Rough Set Theory (RST) is a soft computing concept that handles uncertainty or vagueness in data. Rough Set Theory has found its applications in various fields due to its ability to deal with imprecise knowledge. Hence, in this paper capabilities of RST for information retrieval are explored along with its advantage. Moreover, RST is applied on different classes of sports to index different documents on the basis of similarity. It is also observed, Rough Set Approximations or RSA enable us to give better performance of IR systems in terms of "indexing", "recall", and "precision". It is concluded, to discard the less similar document(s) from indexed objects predefine threshold level of similarity should be specified.

***Keywords:*** *Rough Set, Information Retrieval, Lower Approximation, Upper Approximation, Boundary Region.*

## 1 Introduction

The primary goal of information retrieval system is to differentiate relevant documents from non-relevant documents in a pool of documents and present them in order of their relevance with regard to user's query. Information retrieval deals with uncertain information due to vague and imperfect query formulation by user. To deal with vagueness typical of human knowledge Rough Set theory or RST was given by Pawlak, first time in [1]. RST expresses by Pawlak is not related to degree of belongingness but vagueness. This concept is based on boundary region (BoR) of a set. Rough Set or RS have non vacant boundary region. It means if BoR= φ then set is classical otherwise set is rough set. Non-empty BoR signifies about uncertain or partial knowledge, it means if BoR is there then set is not a crisp set, but RS.

Let, ∪ is Universal Set that consists of all the predefine entity, and R is indiscernibility R ⊆ ∪ x ∪. Moreover R is "equivalence relation", and X ⊆ ∪. The statements 'I-III' are defining X with respect to R or 'X wrt R'.

I. The "lower approximation" or $R_{*_S}$ of set X wrt R is the set of all items, which can be for certain classified as X wrt R.

II. The "upper approximation" or $R^{*_S}$ of set X wrt R is the set of all items, which can be possibly classified as X wrt R.

III. The BoR of set X wrt R is the set of all items, which can be classified neither as X nor ~X wrt R (where ~X represents complement of X).

Now a day's 'RST' is applied in various fields other than IR like robotics, machine learning, data mining, decision support and analysis, knowledge discovery, and artificial intelligence . Since the inception of World Wide Web applications, a number of approaches have been employed to improve the quality of search. Rough set theory to information retrieval is such an attempt.

Proposed work utilized the capability of RST and produced an improve IR system in sports domain. Section II includes related work on rough set theory. Section III consists of elementary concepts of RST. Section IV describes the implementation of different query for sports domain. The outcome are discussed in Section V. Section VI provides conclusion and future directions.

## 2 Related Work

The RST is one of the soft computing tools [1] that are applied by scientist and researchers for IR and feature selection. It is reported [5-8], IR uses rough set approximations (RSA) concept to enhance IR system efficiency. RST based approach uses approximate match between query and documents to retrieve indexed documents. In [2], author uses RSA to organize indexing terms into equivalence classes. Equivalence classes are created on the basis of keywords similarity. In [3], authors introduced RST to IR in which web page similarity served as basis for creating equivalence classes.

In [4], author applied fuzzy set to the RST for IR, which allows using RSA even when fuzzy logic was utilized to illustrate query and web page. The fuzzy logic offers easy

technique to characterize the changeable quantity of relations between "documents, queries, and indexing terms" [7]-[10]. Thus, this expansion makes IR method more efficient.

In [5], author applied RSA to query expansion. Further "*fuzzy rough set"* is also used in this work to handle "graded thesauri" and subjective queries. In [6], author presented how RST can be combined with Fuzzy Formal Concept Analysis (FFCA) to perform semantic web search as well as to retrieve information in the web. In [11], Tsang et.el. applied fuzzy rough set in PNN classifier for Feature and instance reduction, "whereas in [12] Wang et.el. proposed a naïve method for attribute reduction of covering decision systems. In [13], Wu et.el. uses generalized fuzzy rough approximation operators to calculate fuzzy implicators. In [14], Yao and Zhao anticipated attribute reduction in decision-theoretic rough set models, moreover in [15] Sun et.el. build robust fuzzy rough classifier. In [16], Zhao et.el. define model of fuzzy variable precision rough sets. In [17], Zhong et.el. uses rough sets with heuristics for feature selection. In [18], contribution of uncertainty in perception of human is encountered. Upcoming section is throwing light on elementary concept of RST.

## 3 Rough Set Theory

$\overline{P_A}(b)$: $R^{*s}$ in $P_A$ , which is "smallest composed set in A" containing b.

$\underline{P_A}(b)$: The best $R_{*s}$ in $P_A$, which is the "largest composed set in A" that is contained in b.

If φ is null, then wrt $P_A$, we can say that set is:

"Definable" $\Leftrightarrow \overline{P_A(b)} = \underline{P_A}(b)$ .

"Roughly Definable" $\Leftrightarrow \underline{P_A}(b) \neq \varphi$ and $\overline{P_A(b)} \neq \cup$ .

"Externally Definable" $\Leftrightarrow \underline{P_A}(b) \neq \varphi$ and $\overline{P_A(b)} = \cup$ .

"Internally Definable" $\Leftrightarrow \underline{P_A}(b) = \varphi$ and $\overline{P_A(b)} \neq \cup$ .

"Totally Undefinable" $\Leftrightarrow \underline{P_A}(b) = \varphi$ and $\overline{P_A(b)} = \cup$ .

Given $P_A = (\cup, R)$, we may approximate any Roughly Definable subset b of ∪ using $\overline{P_A}(b)$ and $\underline{P_A}(b)$. Also *if* α and β are two subsets of ∪, then the following hold true:

α and β are roughly bottom equal i.e. $\alpha \approx \beta, \Leftrightarrow \underline{P_A}(\alpha) = \underline{P_A}(\beta)$ . (6)

α and β are roughly top equal i.e. $\alpha \approx \beta, \Leftrightarrow \overline{P_A}(\alpha) = \overline{P_A}(\beta)$ . (7)

α and β are roughly equal i.e. $\alpha \approx \beta, \Leftrightarrow \overline{P_A}(\alpha) = \overline{P_A}(\beta)$

and

$\underline{P_A}(\alpha) = \underline{P_A}(\beta)$ . (8)

α is roughly bottom included in β i.e. $\alpha \subseteq \beta \Leftrightarrow \underline{P_A}(\alpha) \subset \underline{P_A}(\beta)$ . (9)

α is roughly top included in β i.e. $\alpha \subset \beta \Leftrightarrow \overline{P_A}(\alpha) \subset \overline{P_A}(\beta)$ . (10)

α is roughly included in β i.e. $x \subseteq y$ i.e. $\Leftrightarrow \overline{P_A}(\alpha) \subset \overline{P_A}(\beta)$

and

$\underline{P_A}(\alpha) \subset \underline{P_A}(\beta)$ . (11)

## 4 Implementation of Rough Set

In recent years, various generalization models of RST has been proposed, which provided the basis for the application of RST in web-based search systems. In [2], author adopted one such approach of partitioning objects (vocabulary terms) into equivalence classes using rough set approximations. Proposed work focuses on investigation of effectiveness of above mentioned method.

Suppose 'W' is the set of web pages {$W_1$, $W_2$, $W_3$,……,$W_{11}$} and 'I' is the set of given indexing terms {Bat, Bounce, Run, Smash, Wimbledon, Headlock, Knockdown, Wrestler, Check, Grandmaster, Goal}. Let the equivalence relation R produces the given below equivalence classes for sports:

Cricket = {Bat, Bounce, Run} (1)

Tennis = {Smash, Wimbledon} (2)

Boxing = {Headlock, Knockdown, Wrestler} (3)

Chess = {Check, Grandmaster} (4)

Football = {Goal} (5)

**Table 1.** Indexed Web Pages

| | Bat | Bounce | Run | Smash | Wimbledon | Head lock | Knock Down | Wrestler | Check | Grand Master | Goal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $W_1$ | ϒ | ϒ | - | ϒ | ϒ | ϒ | - | - | - | - | - |
| $W_2$ | - | - | - | ϒ | ϒ | - | - | - | - | - | - |
| $W_3$ | ϒ | - | - | ϒ | ϒ | - | - | - | - | - | - |
| $W_4$ | ϒ | - | - | ϒ | - | ϒ | - | - | ϒ | - | ϒ |
| $W_5$ | - | - | ϒ | - | - | - | ϒ | - | - | ϒ | - |
| $W_6$ | - | - | ϒ | - | - | - | ϒ | - | - | - | - |
| $W_7$ | ϒ | ϒ | ϒ | - | - | - | - | - | - | - | - |
| $W_8$ | ϒ | ϒ | ϒ | - | ϒ | ϒ | - | - | ϒ | - | ϒ |
| $W_9$ | - | ϒ | - | ϒ | ϒ | - | - | - | - | - | - |
| $W_{10}$ | - | ϒ | - | - | ϒ | - | - | - | ϒ | ϒ | ϒ |
| $W_{11}$ | - | - | - | - | ϒ | - | - | - | - | - | ϒ |

Table 1 shows indexed pages for a set of web pages. Let the user fired one query to IR system containing collection of keywords $U_i$ related to I and a query is comprised of collection of terms so various approaches for searching related pages can be proposed as mentioned in Table 4. Approaches are ranked in the order of relevance (for documents) to the searched terms.

Approach A1 is relevant when $U_i$ is definable in $P_A$. It fetches all web pages with matching description to $U_i$. Approach A2 is implemented to "roughly definable queries". It claims all web pages that are "roughly equivalent" to the $U_i$ definition. Approach A3 and A4 retrieve web pages that match to lower and upper rough set belonging to query. Approach A5 matches web pages that relate to searched terms and some extra web pages that are not included in search query. Approach A6 and A7 are less rigorous flavor of A5. Approach A8 retrieves web pages whose key terms are subsets of the subject included in the search query. Here A9 and A10 approaches represents $R_{*s}(U_i)$ and $R^{*s}(U_i)$. The approaches A11, A12 and A13 identify web pages whose contents are overlapping with the contents of the query.

The above discussed approaches mentioned here identify web pages having varying degree of similarity with the query. This can be performed by estimating similarity between query and set of available web pages. The purpose of measuring similarity between two entities i.e. web pages and query is generating the index in which highly relevant web page will be on top. From the above discussion, equation (12) to (14) is concluded. Equation (12) is serving as a base for approach A5, A8 and A11 to retrieve web pages on the basis of similarity measure, whereas equation (13) intended for approaches A6, A9, and A12. Moreover, equation (14) applies to implement approach A7, A10, and A13.

The distance between document $U_i$ and query $W_j$ can be calculated as:

$$DIS(U_i, W_j) = \underline{DIS}(U_i, W_j) + \overline{DIS}(U_i, W_j) \quad (12)$$

$$\underline{DIS}(U_i, W_j) = |R_{*S}(U_i) \wedge R_{*S}(W_j)| / |R_{*S}(U_i) \vee R_{*S}(W_j)| \quad (13)$$

$$\overline{DIS}(U_i, W_j) = |R^{*S}(U_i) \wedge R^{*S}(W_j)| / |R^{*S}(U_i) \vee R^{*S}(W_j)| \quad (14)$$

Where |P| denotes the cardinality of equivalence classes in P. The above approach calculates the degree of resemblance between web pages and user's query. Table 2 describes the user's searched items.

**Table 2.** Query Representation

| | Bat | Bounce | Run | Smash | Wimbledon | Head lock | Knock Down | Wrestler | Check | Grand Master | Goal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $U_1$ | - | - | - | ϒ | - | - | - | - | - | - | ϒ |
| $U_2$ | - | ϒ | - | ϒ | ϒ | ϒ | ϒ | - | - | - | - |
| $U_3$ | ϒ | ϒ | ϒ | - | - | - | - | - | - | - | - |

**Table 3.** Ordered Results

| Query | Approach | Order of Relevance | Web Pages |
|---|---|---|---|
| $U_1$ | A2 | I | $W_{11}$ |
| | A3 | II | $W_4$ |
| | A5 | III | $W_{10}$ |
| | A5 | IV | $W_8$ |
| | A10 | V | $W_2$ |
| | A13 | VI | $W_3, W_9$ |
| | A13 | VII | $W_1$ |

**Table 4.** Sorted Sequence of Searched Approaches

| Condititon | Approach |
|---|---|
| $U_i = W_j$ | A1 |
| $U_i \approx W_j$ | A2 |
| $U_i \approx W_j$ | A3 |
| $U_i \simeq W_j$ | A4 |
| $U_i \subseteq W_j$ | A5 |
| $U_i \subseteq W_j$ | A6 |
| $U_i \subset W_j$ | A7 |
| $W_j \subseteq U_i$ | A8 |
| $W_j \subseteq U_i$ | A9 |
| $W_j \subset U_i$ | A10 |
| $U_i$ overlap $W_j$ | A11 |
| $U_i$ overlap $W_j$ | A12 |
| $U_i$ overlap $W_j$ | A13 |

Approach A1 is applied, whenever $U_i$ is definable in the approximation space A and retrieves all documents having identical definition to $U_i$ in space A.

## 5 Results and Discussions

The results indicate that rough set approach to information retrieval has been very effective. For query $U_1$= {Smash, Goal} in Table 2, documents retrieved through various strategies (excluding duplicate documents) as shown in Table 3.

**(A2) Roughly definable**

$\overline{P}_A(U_1) = \overline{P}_A(W_{11})$ = {Tennis, Football}

$\underline{P}_A(U_1) = \underline{P}_A(W_{11})$ = {Football}

**(A3) Roughly bottom equal**

$\underline{P}_A(U_1) = \underline{P}_A(W_4)$ = {Football}

**(A5) Roughly included**

$\overline{P}_A(U_1) \subset \overline{P}_A(W_8)$ and $\underline{P}_A(U_1) \subset \underline{P}_A(W_8)$

{Tennis, Football} $\subset$ {Cricket, Tennis, Boxing, Chess, Football} and

{Football} $\subset$ {Cricket, Football}

**(A5) Roughly included**

$\overline{P}_A(U_1) \subset \overline{P}_A(W_{10}) \rightarrow$ {Tennis, Football} $\subset$ {Cricket, Tennis, Chess, Football}

$\underline{P}_A(U_1) \subset \underline{P}_A(W_{10}) \rightarrow$ {Football} $\subset$ {Chess, Football}

**(A10) Roughly top included**

$\overline{P}_A(U_1) \supset \overline{P}_A(W_2) \rightarrow$ {Tennis, Football} $\supset$ {Tennis}

**(A13) Roughly top overlap**

(i) $\overline{P}_A(U_1)$ = {Tennis, Football}

$\overline{P}_A(W_3)$ = {Cricket, Tennis}

(ii) $\overline{P}_A(U_1)$ = {Tennis, Football}

$\overline{P}_A(W_9)$ = {Cricket, Tennis}

(iii) $\overline{P}_A(U_1)$ = {Tennis, Football}

$\overline{P}_A(W_1)$ = {Cricket, Tennis, Boxing}

## 6. Conclusion and Future Directions

The present work facilitates detailed analysis of various approaches that were used on rough set for document search. It is evident from the study that IR system improved significantly in terms of document indexing and user's satisfaction with other parameters viz. precision and recall. The advantages of using rough set approach in information retrieval have been analyzed and possible improvements in the existing system are discussed as potential research areas.

In the proposed work, indexing terms are organized into equivalent classes by human experts; the above method of equivalence class partitioning is slow and biased as process is influenced by individual perception. This could be dealt with the use of automated tool for finding terms associations. Finally, generalization based rough set models may also be used in IR systems for making search strategies more practical.

## References


1. Pawlak, Z. Rough sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.
2. Das-Gupta, P. Rough sets and information retrieval, The Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 567-581, 1988.
3. Wong, S. K. M., Ziarko, W. A Machine Learning Approach to Information Retrieval, Research and Development in Information Retrieval. 1986.
4. Srinivasan, P. Ruiz, M. E., Kraft, D. H., Chen J., Kundu, S. Vocabulary mining for information retrieval: rough sets and fuzzy sets, Information Processing and Management, 1998.
5. De Cock, M. Cornelis, C. Fuzzy rough set based web query expansion. In: Proceedings of Rough Sets and Soft Computing in Intelligent Agent and Web Technology, International Workshop at WIIAT, pp. 9-16. 2005.
6. Formica, A. Semantic Web search based on rough sets and Fuzzy Formal Concept Analysis, Knowledge-Based Systems, 2012.
7. Yao, Y. Y., Li, X., Lin, T. Y., Liu, Q. Representation and Classification of rough set models. Soft Computing: Proceedings of the 3rd International Workshop on Rough Sets and Soft Computing, 1994.
8. Yao, Y. Y. Combination of rough and fuzzy sets based on α-level sets. In Rough Sets and Data Ming: Analysis for Imprecise Data, 1997.
9. Zhou, B. Yao, Y. Y. Evaluating information retrieval system performance based on user preference. Journal of Intelligent Information Systems. Volume 34. Issue 3. pp. 227-248, 2010.
10. Ziarko, W., Fei, X. VPRSM approach to WEB searching. Rough Sets and Current Trends in Computing, 2002.
11. E. C. C. Tsang, Q. Hu, and D. Chen, "Feature and instance reduction for PNN classifiers based on fuzzy rough sets," Int. J. Mach. Learning Cybern., vol. 7, pp. 1–11, 2014.
12. C. Wang, Q. He, D. Chen, and Q. Hu, "A novel method for attribute reduction of covering decision systems," Inf. Sci., vol. 254, pp. 181–196, 2014.
13. W. Z. Wu, Y. Leung, and M. W. Shao, "Generalized fuzzy rough approximation operators determined by fuzzy implicators," Int. J. Approx. Reason., vol. 54, no. 9, pp. 1388–1409, 2013.

14. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," Inf. Sci., vol. 178, no. 17, pp. 3356–3373, 2008.

15. S. Zhao, H. Chen, C. Li, X. Du, and H. Sun, "A novel approach to building a robust fuzzy rough classifier," IEEE Trans. Fuzzy Syst., vol. 23, no. 4, pp. 769–786, Aug. 2015.

16. S. Zhao, E. C. C. Tsang, and D. Chen, "The model of fuzzy variable precision rough sets," IEEE Trans. Fuzzy Syst., vol. 17, no. 2, pp. 451–467, Aug. 2009.

17. N. Zhong, J. Dong, and S. Ohsuga, "Using rough sets with heuristics for feature selection," J. Intell. Inf. Syst., vol. 16, no. 3, pp. 199–214, 2001.

18. Rahman Abdul, Beg M.M.S., "Face Sketch Recognition Using Sketching With Words", Springer Journal of Machine Learning and Cybernetics ISSN 1868-8071, April, 2014.