Sheaf Theoretic Model for Information Retrieval

[1]P.VamsiSagar,[2]M. Phani Krishna Kishore
[1]Research Scholar, [2]Professor
[1]Department of Mathematics, [2]Gayatri Vidya Parishad College of Engineering(Autonomous)
[1]Rayalaseema University, Kurnool,Andhra Pradesh, India,[2]Madhurawada, Visakhapatnam,Andhra Pradesh,India

ABSTRACT: ***Representation and storage of meta data plays a crucial role in information retrieval. Several mathematical models are being used for building effective systems. As the data is growing exponential level in the quest for new techniques sheaf theory offers a promising approach which deals with connections between local and global properties of sets of objects. Ever since Jean Leray's introduction of the sheaves it attracted many of the scientists and researchers from various fields of Science and Engineering. In the present paper, a data representation model is proposed based on sheaf representation via tolerances.***

Key Words: **Tolerances, Sheaves, Information System.**

1. INTRODUCTION:

In the recent past digital explosion has been creating huge volumes of data in both structured and unstructured forms. Ever since Edger Codd has created the relational model for structured data bases, there is a paradigm shift in the information management. The concepts of Internet of Things has revolutionised the data generation and transmission in unstructured domain as well. To extract meaningful and relevant information from huge volumes of data, several mathematical models are applied in practice. Mere keyword based search is not giving good results as the same word may represent different entities in domains. To address this problem NLP techniques are investigated by several researchers. Semantic web concepts are proposed. However not much progress is being made in this direction as it lacks a sound mathematical base.

Jean Leray introduced Sheaf representations. To deduce global properties from the local data, Sheaves constructed over topological spaces are suitable. There are several approaches and notations to study the theory of sheaves. It has been successfully applied to Differential Geometry, complex analysis, algebraic topology. Recently these concepts are being applied to the field of parallel programming, understanding Networks [17]. Comer [1], Keimel [2], Hofmann [3], Davey [5], Swamy [4] studied about the representations of algebraic structures through sheaves defined over topological spaces. Swamy [4] and Wolf [7] independently developed the sheaf construction mechanism based on Chinese Remainder Theorem for universal algebra. Using equivalence relations (congruences) on the set (algebra). Recently M.P.K.Kishore et al., [6] gave a construction mechanism of sheaves of sets suing tolerances. Sheaves over tolerances offer flexibility to apply the concepts to real time entities as most of the relations are reflexive and symmetric only.

In the present work the concept of tolerance relation based sheaves is applied to an information system. A general frame work is developed to store the meta data which is stored in the forms of stalks of the sheaf. Once a query is posed, first the query is operated on the meta data and identifies the relevant stalk and the corresponding tolerance classes.

2. PRELIMINARIES:

Basic definitions related to sets and sheaves were presented in this section.

Definition 2.1. A relation defined on a set is said to be an equivalence relation if it is a reflexive, transitive and symmetric.

Definition 2.2. A relation defined on a set is said to be a tolerance relation if it is a reflexive, and symmetric.

Definition 2.3: Let X be a non-empty set and W be a collection of subsets of Xsatisfying the following three conditions:

1. W shall contain empty set $\emptyset$ and the total set X.
2. Under finite intersections W is closed.
3. Under arbitrary unions W is closed.

Then W is called a topology on X, the members of W are called open sets and the pair $(X, W)$is called a topological space.

A set X for which a topology W has been specified is called a topological space.

Example 2.4: Let $X = \{1, 2, 3, 4, 5\}$ and $W = \{\phi, X, \{1,2\}, \{1,2,3\}, \{1,3,4\}, \{1,2,4\}, \{2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}\}$ are some collection of subsets of X.

Clearly ϕ and X are in W.

$\{1, 2, 3\} \cup \{1, 2\} = \{1, 2, 3\} \in W$ and $\{1, 2, 3\} \cup \{1, 3, 4\} = \{1, 2, 3, 4\} \in W$

Also $\{1, 2, 3\} \cap \{1, 2\} = \{1, 2\} \in W$

So we can say that X is a topological space.

Definition 2.5: A subset Q of X is said to be open in X if for each $y \in Q$ there is a basis element $P \in \beta$ such that $y \in P$ and $P \subseteq Q$.

Definition 2.6: Let(X, υ) and (Y, λ) be two topological spaces. A mapping$g: X \to Y$ is said to be continuous if for each U open in Y, $g^{-1}(U)$ is open in X, we can say in other words the mapping that it brings back open sets in Y to open sets in X.

Definition 2.7: Let(X, υ) and (Y, λ) be topological spaces and let g be a mapping of Xinto Y. Then g is said to be a homeomorphism if and only if g is isomorphic, continuous and an open map between the two topological spaces X and Y.

Definition 2.8: Let(X, υ) and (Y, λ) be topological spaces and let g be a mapping of X into $Y$is a local homeomorphism if for every point $x \in X$ there exists open setsG, A containing X, $g(X)$ respectively such that the restriction $g|_G: G \to A$ is a homeomorphism.

Definition 2.9: Let A and B be sets. The disjoint union of A and B is the set represented by $A \sqcup B$ of elements of the form(a, A) or (b, B) where $a \in A$and $b \in B$, symbolically we write $A \sqcup B = \{(a, A)|a \in A\} \cup \{(b, B)|b \in B\}$

Example 2.10: Let $A = \{1, 2, 3\}$ and $B = \{2, 3\}$. The disjoint union of A and B is $A \sqcup B = \{(1, A), (2, A), (3, A), (2, B), (3, B)\}$

Sheaves are constructions that relate ideas of number theory, algebraic geometry, and topology. In topology, the sheaves encode distinctions between local and global properties of a space. So we can consider sheaf as a mathematical tool for storing locally defined information attached to the open sets of a topological space.

Definition 2.11:
A sheaf is represented by a triple(G, π, X)which admits the following conditions:

a. G *and* Xbe two topological spaces.

b. $\pi: G \to X$be the local homeomorphism from G onto X.

So is, $\pi: G \to X$ is a surjection in such a way that for any$g \in G$, there exists open sets A, B in G, X respectively such that $g \in A$, $\pi(g) \in B$ also $\pi|_A: A \to B$ is a homeomorphism. In this$G$ *and* X are named the sheaf space and the base space respectively and π is named the projection map. Often we called that (G, π, X) is a sheaf over$X$. For any $p \in X$, $\pi^{-1}(p) \neq \emptyset$ and is named the stalk at p, we represent this by G_p. Note that G is the disjoint union of all $G_p{}'s$.

Definition 2.12: Let (G, π, X) be a sheaf and $Y \subseteq X$. A section on Y is a continuous mapping$g: Y \to G$ such that $\pi \circ g = Identity$. Sections on the total space$X$ are called global sections.

Definition 2.13: Let (G, π, X) be a sheaf. If for every$g \in G$, there is a global section g and $y \in X$ such that $g \in f(X)$, then (G, π, X) is called a global sheaf.

3. STRUCTURE OF A SHEAF:

In this section, the construction of sheaf for an information system is given. Given a topological space of concepts where each concept consists of key words. Tolerance classes are constructed for each concept. The tolerance classes for each of the concepts forms a stalk. The disjoint union of stalks topologised with a topology relevant to the topology defined on concepts forms the global sheaf. The global nature of the sheaf ensures that each document is associated with at least one concept. In this model provision for explicitly stating the concept along with key word is also considered to optimize the search and retrieval process.

Let$D$ denote the set of documents and C denote the set of concepts where each $c \in C$ is described by $c = \{k_1, k_2, k_3, \dots, k_n\}$key words.

Let $X = (C, T)$be a topological space on the set of key words with some topology T.

Define for each $c \in C, R_c = \{(d_i, d_j) \in D \times D \mid k(d_i) \cap k(d_j) \neq \emptyset\}$ where $k(d_i)$ is the set of key words in the document d_i.

Let R_c be the tolerance relation corresponding to the concept$c$ and $c \mapsto R_c$ be a mapping from X into $Tol(D)$.

Let G_c be the quotient set $D|R_c$ for any $c \in C$.

Define $G = \sqcup_{c \in C} G_c$be the disjoint union of $G_c{}'s$.

Define $\widehat{d_i}: X \to G$ by $\widehat{d_i}(c_j) = [d_i]_{R_{c_j}}$for each $d_i \in D$ and $c_j \in C$

Consider a topology on G such that each $\widehat{d_i}$ is open and continuous for each $d_i \in D$ and define $\pi: G \to X$ by $\pi(\{d_i, d_j\}, c_k) = c_k$. Then$(G, \pi, X)$forms a triple.

Theorem 3.1: The triple (G, π, X) defined as above is a global sheaf if and only if for every $d_1, d_2 \in D$, $X(d_1, d_2) = \{c_k \in X | \widehat{d_1}(c_j) = \widehat{d_2}(c_j)\}$is open in X.
Proof of the above theorem is given in [6].

Note3.2:
In the above construction we may choose a convenient topology on the set of all concepts and obtain the corresponding sheaf of sets over an information system. Similarly by considering different tolerance relations on the set of all documents we get suitable sheaf spaces.

4. ILLUSTRATION: In the following we have taken a set of documents and the set of concepts consisting set of key words in an information system form.
Consider an information system given by

$C(Concepts)$ $/D(Documents)$	c_1 (Programming)	c_2 (Mathematics)	c_3 (Hardware)
d_1	Python, Logic	Numerical Methods, Logic	PLC, C-language
d_2	Logic, Mat lab	Mathematica, Cobinatorics	C-language, HDLC
d_3	Java, Mat lab	Algorithms, Logic	HDLC, ARM Processor
d_4	Scala, Functonal Programming	GraphTheory, Combinatorics	ARM Processor, VLSI
d_5	R-Programming, Scala	Algoritms, Numerical Methods	PLC, VLSI

For the sake of example consider discrete topology on the set of concepts X = $\{c_1, c_2, c_3\}$.

Now we write the set of all Tolerance relations defined on the set of documents $\{d_i\}_{i=1 to\, 5}$denoted by R_c corresponding to the concepts$\{c_i\}_{i=1 to\, 3}$.

$$R_c = \left\{\begin{matrix} (d_1,d_1),(d_2,d_2),(d_3,d_3),(d_4,d_4),(d_5,d_5),(d_1,d_2),(d_1,d_3),(d_1,d_5),(d_2,d_1), \\ (d_2,d_3),(d_2,d_4),(d_3,d_1),(d_3,d_2),(d_3,d_4),(d_3,d_5),(d_4,d_2),(d_4,d_3),(d_4,d_5), \\ (d_5,d_1),(d_5,d_3),(d_5,d_4) \end{matrix}\right\}$$

Also the tolerances related to the individual documents$\{d_i\}_{i=1 to\, 5}$ are as follows.

$$Tol(d_1) = \{d_1, d_2, d_3, d_5\}$$

$$Tol(d_2) = \{d_1, d_2, d_3, d_4\}$$

$$Tol(d_3) = \{d_1, d_2, d_3, d_4, d_5\}$$

$$Tol(d_4) = \{d_2, d_3, d_4, d_5\}$$

$$Tol(d_5) = \{d_1, d_3, d_4, d_5\}$$

The tolerance relations corresponding to the concepts $\{c_i\}_{i=1 to\, 3}$are given as follows.

$$R_{c_1} = \left\{\begin{matrix} (d_1,d_1),(d_1,d_2),(d_2,d_1),(d_2,d_2),(d_2,d_3), \\ (d_3,d_2),(d_3,d_3),(d_4,d_4),(d_4,d_5), \\ (d_5,d_4),(d_5,d_5) \end{matrix}\right\}$$

$$R_{c_2} = \left\{\begin{matrix} (d_1,d_1),(d_1,d_3),(d_1,d_5),(d_2,d_2),(d_2,d_4), \\ (d_3,d_1),(d_3,d_3),(d_3,d_5),(d_4,d_2), \\ (d_4,d_4),(d_5,d_1),(d_5,d_3),(d_5,d_5) \end{matrix}\right\}$$

$$R_{c_3} = \left\{ \begin{array}{c} (d_{1,}d_1), (d_{1,}d_2), (d_{1,}d_5), (d_{2,}d_1), (d_{2,}d_2), \\ (d_{2,}d_3), (d_{3,}d_2), (d_{3,}d_3), (d_{3,}d_4), \\ (d_{4,}d_3), (d_{4,}d_4), (d_{4,}d_5), (d_{5,}d_1), (d_{5,}d_4), (d_{5,}d_5) \end{array} \right\}$$

The tolerance classes of documents $\{d_i\}_{i=1to\,5}$corresponding to $R_{c_1}$are given as follows.

$$[d_1]_{R_{c_1}} = \{d_1, d_2\}, [d_2]_{R_{c_1}} = \{d_1, d_2, d_3\}$$

$$[d_3]_{R_{c_1}} = \{d_2, d_3\}, [d_4]_{R_{c_1}} = \{d_4, d_5\}, [d_5]_{R_{c_1}} = \{d_4, d_5\}$$

The stalk corresponding to the corresponding to the concept $c_1$consists all the above tolerance classes, denoted by G_{c_1}

That is $G_{c_1} = \{[d_1]_{R_{c_1}}, [d_2]_{R_{c_1}}, [d_3]_{R_{c_1}}, [d_4]_{R_{c_1}}, [d_5]_{R_{c_1}}]\}$

The tolerance classes of documents $\{d_i\}_{i=1to\,5}$corresponding to $R_{c_2}$are given as follows.

$$[d_1]_{R_{c_2}} = \{d_1, d_3, d_5\}, [d_2]_{R_{c_2}} = \{d_2, d_4\}$$

$$[d_3]_{R_{c_2}} = \{d_1, d_3, d_5\}, [d_4]_{R_{c_2}} = \{d_2, d_4\}, [d_5]_{R_{c_2}} = \{d_1, d_3, d_5\}$$

The stalk corresponding to the corresponding to the concept $c_2$consisting all the above tolerance classes, denoted by G_{c_2}

That is $G_{c_2} = \{[d_1]_{R_{c_2}}, [d_2]_{R_{c_2}}, [d_3]_{R_{c_2}}, [d_4]_{R_{c_2}}, [d_5]_{R_{c_2}}]\}$

The tolerance classes of documents $\{d_i\}_{i=1to\,5}$corresponding to $R_{c_3}$are given as follows.

$$[d_1]_{R_{c_3}} = \{d_1, d_2, d_5\}, [d_2]_{R_{c_3}} = \{d_1, d_2, d_3\}$$

$$[d_3]_{R_{c_3}} = \{d_2, d_3, d_4\}, [d_4]_{R_{c_3}} = \{d_3, d_4, d_5\}, [d_5]_{R_{c_1}} = \{d_1, d_4, d_5\}$$

The stalk corresponding to the corresponding to the concept $c_3$consisting all the above tolerance classes, denoted by G_{c_3}

That is $G_{c_3} = \{[d_1]_{R_{c_3}}, [d_2]_{R_{c_3}}, [d_3]_{R_{c_3}}, [d_4]_{R_{c_3}}, [d_5]_{R_{c_3}}]\}$

And G is the disjoint union of all $G_c's$.

That is $G = \sqcup_{c \in C} G_c = \sqcup \{G_{c_1}, G_{c_2}, G_{c_3}\}$

$$=\left\{\begin{array}{c}(\{d_1,d_2\},\ c_1),(\{d_1,d_2,d_3\},c_1),(\{d_2,d_3\},c_1),\ (\{d_4,d_5\},c_1),(\{d_4,d_5\},c_1),\\(\{d_1,d_3,d_5\},c_2),\ (\{d_2,d_4\},c_2),(\{d_1,d_3,d_5\},c_2),(\{d_2,d_4\},c_2),(\{d_1,d_3,d_5\},c_2),\\(\{d_1,d_2,d_5\},c_3)\ ,(\{d_1,d_2,d_3\},c_3),(\{d_2,d_3,d_4\},c_3),(\{d_3,d_4,d_5\},c_3),(\{d_1,d_4,d_5\},c_3)\end{array}\right\}$$

Now by defining $\widehat{d_\iota}: X \rightarrow G$ by $\widehat{d_\iota}(c_j) = [d_i]_{R_{c_j}}$for each $d_i \in D$ and $c_j \in C$

Consider any topology on G such that each $\widehat{d_\iota}$ is open and continuous for each $d_i \in D$ and define $\pi: G \rightarrow X$ by $\pi(\{d_i, d_j\}, c_k) = c_k$. Then$(G,\ \pi,\ X)$forms a triple.

Consider a set of documents denoted by D. Let Cdenote the set of key words. Let $F: D \rightarrow C$ be a map that assigns for each document the key words present in the concept it C.
Then$(D,\ F,\ C)$denotes an information system.We can Construct a sheaf over a prescribed topology for this given information system. In the sheaf every stalk contains the tolerance classes for a prescribed set of key words.

The storage of the metadata in the above format for the given information system facilitates quick query processing on the corresponding stack. Given a query with a set of key words the corresponding smallest element in the topology is considered and the corresponding information from the sheaf is retrieved from the relevant stack.

And $\pi^{-1}(c_i)_{i=1\ to\ 3} = \{d_i\}_{i=1\ to\ 5}$

$\pi^{-1}(c_1)_{python} = \{d_1\}$and $\pi^{-1}(c_2)_{logic} = \{d_1, d_3\}$

Now $\pi^{-1}(c_1)_{python} \cup \pi^{-1}(c_2)_{logic} = \{d_1, d_3\}$

And $\pi^{-1}(c_1)_{python} \cap \pi^{-1}(c_2)_{logic} = \{d_1\}$

For the above example suppose for a given query to know the details about the key words Python and logic. The search engine immediately shows about a python a snake details. But actually we want the information of Python programming. So when we want to identify the correct details about our query, under the sheaf structure we have to give the concept (c_i)also clearly so that the information retrieval can easily retrieve the data from the documents which consists of the concepts. From the above mathematical sheaf representation consisting of the stalks which stores the information about all the key words. Hence when we mention the query clearly, the search engine it searches for the concerned topology applied on the sheaf structure it identifies the corresponding stalk where the information is related to the python is stored in the document. Python key word exists in the document d_1 and the key word logic presented in the documents d_1, d_3 but we want the information related to both Python and logic, if we observe the intersection of these two documents we can easily identify that $d_1$is the only document which consists both key words Python and logic. From this the information system retrieves the information related to Python and logic.

Also considering another example to retrieve the data related to the key words Scala and ARM Processor, the system performs the following operations.

$\pi^{-1}(c_i)_{i=1\ to\ 3} = \{d_i\}_{i=1\ to\ 5}$

$\pi^{-1}(c_1)_{Scala} = \{d_4, d_5\}$ and $\pi^{-1}(c_3)_{ARM\ Processor} = \{d_3, d_4\}$

Now $\pi^{-1}(c_1)_{Scala} \cup \pi^{-1}(c_3)_{ARM\ Processor} = \{d_3,\ d_4,\ d_5\}$

And $\pi^{-1}(c_1)_{Scala} \cap \pi^{-1}(c_3)_{ARM\ Processor} = \{d_4\}$

By the above procedure we can observe that the data relevant to the Scala and ARM Processor is stored in the document d_4. From there the information system retrieves the data related to the key words Scala and Arm Processor using the sheaf model consisting the stalks which stores the data in the form of tolerances related to the corresponding key words.

Similarly to retrieve the information suitable to the key words GraphTheory and Combinatorics, the system perform the following procedure.

$$\pi^{-1}(c_i)_{i=1\ to\ 3} = \{d_i\}_{i=1\ to\ 5}$$

$\pi^{-1}(c_2)_{Graph\ theory} = \{d_4\}$ and $\pi^{-1}(c_2)_{Combinatorics} = \{d_2, d_4\}$

Now $\pi^{-1}(c_2)_{Graph\ Theory} \cup \pi^{-1}(c_2)_{Combinatorics} = \{d_2,\ d_4\}$

And $\pi^{-1}(c_2)_{Graph\ Theory} \cap \pi^{-1}(c_2)_{Combinatorics} = \{d_4\}$

From the above strategy we can conclude that the data relevant to the Graph Theory and Combinatorics is stored in the document d_4. From there the information system retrieves the data related to the key words Graph Theory and Combinatorics using the sheaf representation mathematical model consisting the stalks which stores the data in the form of tolerances related to the corresponding key words Graph Theory and Combinatorics.

Hence by the observation of the above we can say that from the storage of the metadata, the sheaf representation model retrieves the information corresponding to the query in the above process.

5. CONCLUSION:

In this paper a sheaf theoretic modal for information storage and retrieval using tolerance relations is presented.

REFERENCES:

[1] Comer. S. D., Representation by algebras of sections over Boolean spaces, *Pacific J. Math.*38,(1971) 29 -38.

[2] Keimel. K., The representation of lattice-ordered groups and rings by sections in sheaves, in *Lecture on the Applications of Sheaves to Ring Theory*, Lecture Notes in Math. 248, Springer-Verlag, Berlin and New York, pages(1971) 1- 98.

[3] Hofmann,K. H. Representation of algebras by continuous sections, *Bull. Amer.math. Soc.* 78, Number 3, (1972) 291-373.

[4] Swamy, U.M.,Representation of Universal algebras by sheaves, Proc.Amer.Math. Soc., 45, (1974) 55-58.

[5] Davey, B. A. sheaf spaces and sheaves of universal algebras, *Math. Z.*,134,(1973) 275-290.

[6] Kishore. M.P.K., Ravi Kumar. R.V.G., VamsiSagar. P.Construction of sheaves by tolerance Relations, *Asian-European Journal of Mathematics*, Vol. 9, No. 2 (2016)650035 (8 pages), World Scientific Publishing Company,DOI: 10.1142/S1793557116500352

[7 Wolf. A.,sheaf representations of arithmetical algebras, *Memoirs Amer. Math.Soc.* 148, (1974) pp. 85-97.

[8] Grant Malcolm, Sheaves, Objects and Distributed Systems, *Electronic Notes in Theoretical Computer Science, Elsevier* (225) (2009) 3-19.

[9] James R. Munkres, Topology A First Course, *Prentice-Hall International*.

[10] Skowron,A. and Stepaniuk, J.Tolerance Approximation Spaces *FundamentaInformaticae 27*, (1996) pp. 245-254.
[11] B. Zelinka Tolerance in algebraic structures, *Czechoslovak Mathematical Journal*, Vol 20, No.2 (1970), 179-183.
[12] E. C. Zeeman: The Topology of the Brain and Visual Perception. In: *The Topology of 3-Manifolds. Ed.* by M. K. Fort, (1962), pp. 240-256.
[13] Ivan Chajda; BohdanZelinka Tolerance relation on lattices *Časopis pro pěstovánímatematiky,* Vol. 99 (1974), No. 4, 394—399
[14] Christopher D. Manning, PrabhakarRaghavan and HinrichSchutzeFrontmatter-Introduction to Information Retrieval *Cambridge University Press* 978-0-521-86571-5
[15] A. McCallum, K. Nigam, J. Rennie, K. Seymore. 2000. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval Journal*, volume 3, Kluwer. (2000). pages 127-163.
[16] T.D. Wilson, Models in Information behaviour research, *Journal of Documentation,* Vol. 55, No. 3, June (1999).
[17] Michael Robinson, Understanding networks and their behaviors using Sheaf theory, *arXiv:1308.4621v1 [math AT]* 21 August 2013.
[18] David M. Butler, Sheaf Data Model, *United States Patent, Patent No. US 7,865,526 B2;* Date of Patent: Jan 4. 2011.
[19] B. Zelinka, Tolerance in algebraic structures, Czech. Math. J. 20 (2) (1970) 179-183.
[20] J. Pogonowski, *Tolerance Spaces with Applications to Linguistics* (University of Adam Mickiewicz Press, 1981).
[21] I. Chajda, *Algebraic Theory of Tolerance Relations* (Palacky University Olomouc, 1991)