International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 12
ISSN: 2454-4248
281 – 285

Web Page Enrichment using a Rough Set Based Method

Bacharaju Vishnu Swathi
Professor, Dept. of CSE
Geetanjali College of Engg. & Tech
Keesara, India
e-mail: swathiveldanda@yahoo.com

***Abstract*—**When documents are matched to a given query, often the terms in the query are matched to the words in the documents for calculating similarity. But it is a good idea if the given document is represented in an enriched manner with not only the actual words occurring in the document but also with the synonyms of the important words. This would definitely improve the recall of the system. With its ability to deal with vagueness and fuzziness, tolerance rough set seems to be promising tool to model relations between terms and documents. In many information retrieval problems, especially in text classification, determining the relation between term-term and term-document is essential. In this work, the application of TRSM to web page classification was evaluated to determine its effectiveness as a way to enrich a web page.

Keywords*-web page, enrichment,classification, rough sets,text minig*

*****

I. INTRODUCTION

In the recent past, the world wide web has been witnessing an explosive growth. Information is kept on the web in various formats and the content is dynamic in nature. All the leading web search engines, namely, google, yahoo, askjeeves, etc. are vying with each other to provide the web user with the appropriate content in response to his/her query. In most cases, the user is flooded with millions of web pages in response to his query and it is common knowledge that not many users go past the first few web pages. In spite of the multitude of the pages returned, most of the time, the average user does not find what he/she is looking for in the first few pages he/she manages to examine. It is really debatable as to how useful or meaningful it is for any search engine to return lakhs of web pages in response to a user query. In spite of the sophisticated page ranking algorithms employed by the search engines, the pages the user actually needs may actually get lost in the huge amount of information returned. Since most users of the web are not experts, grouping of the web pages into categories helps them to navigate quickly to the category they are actually interested and subsequently to the specific web page. This will reduce the search space for the user to a great extent.

It is strongly believed and felt that the experience of a person using a web search engine is enhanced multifold if the results are nicely categorized as against the case where the results are displayed in a structure less, flat manner. All manuscripts must be in English. A third approach to text classification is based on machine learning. In machine learning, the set of rules or, more generally, the decision criterion of the text classifier is learned automatically from training data. This approach is also called statistical text classification if the learning method is statistical. In statistical text classification, a number of good example documents (or training documents) from each class are required for training the classifier. The need for manual classification is not eliminated since the training documents come from a person who has labeled them where labeling refers to the process of annotating each document with its class. But labeling is arguably an easier task than writing rules. Almost anybody can look at a document and decide whether it is about cricket or not but it takes an expert to form the rules to identify a document's class.

It is customary to represent a document in a reduced form, by removing stop words and by further reducing by feature selection. When documents are matched to a given query, often the terms in the query are matched to the words in the documents for calculating similarity. But it is a good idea if the given document is represented in an enriched manner with not only the actual words occurring in the document but also with the synonyms of the important words. This would definitely improve the recall of the system.

II. BACKGROUND WORK

Tolerance Rough Set Model (TRSM) was developed [1, 2] as basis to model documents and terms in information retrieval, text mining, etc. With its ability to deal with vagueness and fuzziness, tolerance rough set seems to be promising tool to model relations between terms and documents. In many information retrieval problems, especially in text classification, determining the relation between term-term and term-document is essential. The application of TRSM in web page classification was proposed as a way to enrich web page.

The starting point of rough set theory is that each set X in a universe U can be "viewed" approximately by its upper and lower approximations in an approximation space $\Re$ = (U,R), where $R \subseteq U \times U$ is an equivalence relation. Two objects $x, y \in U$ are said to be indiscernible regarding R if $x\,R\,y$. The lower and upper approximations in R of any $X \subseteq U$, denoted respectively by L ($\Re$, X) and U ($\Re$, X), are defined by

$$L(\Re,X) = \{x \in U : [x]_R \subseteq X\}$$

$$U(\Re,X) = x \in U : [x]_R \cap X \neq \phi\}$$

where $[x]_R$ denotes the equivalence class of objects indiscernible with x regarding the equivalence relation R. All early work on information retrieval using rough set was based on traditional RST with a basic assumption that the set T of index terms can be divided into equivalence classes determined by equivalence relations [3]. The three properties of an equivalence relation R (reflexi, xRx ; symmetric, $xRy \rightarrow yRx$;

 ISSN: 2454-4248
281 – 285

and transitive, $xRy \wedge yRz \rightarrow xRz$ for $\forall x, y, z \in U$), the transitive property does not always hold in certain application domains, particularly in natural language processing and information retrieval. This remark can be illustrated by considering words from Roget's thesaurus, where each word is associated with a class of other words that have similar meanings. Figure 1 shows associated classes of three words, *root*, *cause*, and *basis*. It is clear that these classes are not disjoint (equivalence classes), but overlapping, and the meaning of the words is not transitive.

Overlapping classes can be generated by *tolerance relations* that require only reflexive and symmetric properties. A general approximation model using tolerance relations was introduced in [4] in which generalized spaces are called *tolerance spaces* that contain overlapping classes of objects in the universe (tolerance classes). In [4], a tolerance space is formally defined as a quadruple $R = (U, I, v, P)$ where U is a universe of objects, $I : U \rightarrow 2^U$ is an uncertainty function, $v : 2^U \times 2^U \rightarrow [0,1]$ is a vague inclusion, and $P : I(U) \rightarrow \{0,1\}$ is a structurality function.

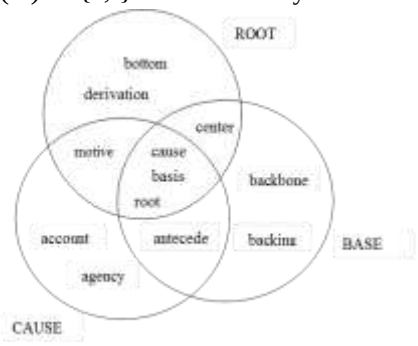

Figure 1. Overlapping classes of words

Assume that an object x is perceived by information $Inf(x)$ about it. The uncertainty function $I : U \rightarrow 2^U$ determines $I(x)$ as a tolerance class of all objects that are considered to have *similar* information to *x*. This uncertainty function can be any function satisfying the condition $x \in I(x)$ and $y \in I(x)$ iff $x \in I(y)$ for any $x, y \in U$. Such a function corresponds to a relation $\tau \subseteq U \times U$ understood as *xTy* iff $y \in I(x)$. τ is a tolerance relation because it satisfies the properties of reflexivity and symmetry. The vague inclusion $v : 2^U \times 2^U \rightarrow [0,1]$ measures the degree of inclusion of sets; in particular it relates to the question of whether the tolerance class $I(x)$ of an object $x \in U$ is included in a set *X*. There is only one requirement of *monotonicity* with respect to the second argument of *v*, that is, $v(X, Y) \leq v(X, Z)$ for any $X, Y, Z \subseteq U$ and $Y \subseteq Z$.

Finally, the structurality function is introduced by analogy with mathematical morphology [5]. In the construction of the lower and upper approximations, only tolerance sets being structural elements are considered. The structurality function, $P : I(U) \rightarrow \{0,1\}$, classifies $I(x)$ for each $x \in U$ into two classes – structural subsets $(P(I(x)) = 1))$ and non-structural subsets $(P(I(x)) = 0))$. The lower approximation *L(R, X)* and the upper approximation *U(R, X)* in *R* of any $X \subseteq U$ are defined as

$$L_R(X) = \{x \in U \mid P(I(x)) = 1 \,\&\, v(I(x), X) = 1\}$$

$$U_R(X) = \{x \in U \mid P(I(x)) = 1 \,\&\, v(I(x), X) > 0\}$$

The basic problem of using tolerance spaces in any application is in the suitable determination of I, v and P.

A. *Determination of Tolerance spaces*

Let $P = \{p_1, p_2, \ldots p_N\}$ be a set of document and $T = \{t_1, t2, \ldots\ldots t_M\}$ set of *index terms* for P. With the adoption of VSM each page p_i is represented by a weight vector $[w_{i1}, w_{i2}, \ldots\ldots w_{iM}]$ where w_{ij} denoted the weight of term *j* in page i . In TRSM, the tolerance space is denoted over a universe of all index terms

$$U = T = \{t_1, t_{21} \ldots\ldots t_M\}.$$

The idea is to capture conceptually related index terms into classes. There are several ways to identify conceptually related index terms, for example, human experts, thesaurus, term co-occurrence, and so on. In this work, term co-occurrence is employed to determine tolerance relation and tolerance class. The co-occurrence of the index terms is chosen for the following reasons: (i) It gives a meaningful interpretation about the dependency and semantic relation of index terms [6]; and (ii) it is relatively simple and computationally efficient.

B. *Tolerance class of term*

Let $f_P(t_i, t_j)$ denotes the number of web pages in P in which both terms t_i and t_j occurs. The uncertainty function I with regards to threshold θ is defined as

$$I_\theta(t_i) = \{t_j \mid f_P(t_i, t_j) \geq \theta\} \cup \{t_i\}$$

clearly, the above function satisfies conditions of being reflexive; $t_i \in I_\theta(t_i)$ and symmetric: $t_j \in I_\theta(t_i) \Leftrightarrow t_i \in I_\theta(t_j)$ for any $t_i, t_j \in T$. Thus, the tolerance relation $\tau \subseteq T \times T$ can be defined by means of function I:

$t_i \,\tau\, t_j \Leftrightarrow t_j \in I_\theta(t_i)$ where $I_\theta(t_i)$ is the tolerance class of index term t_i . In context of Information Retrieval, a tolerance class represents a concept that is characterized by terms it contains. By varying the threshold θ (e.g. relatively to the size of web page collection), one can control the degree of relatedness of words in tolerance classes (or in other words the preciseness of the concept represented by a tolerance class). To measure degree of inclusion of one set in another, vague inclusion function is defined as

 ISSN: 2454-4248
281 – 285

$$v(X,Y) = \frac{|X \cap Y|}{|X|}$$

It is clear that this function is monotonous with respect to the second argument. The membership function μ for $t_i \in T, X \subseteq T$ is then defined as

$$\mu(t_i, X) = v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|}$$

With the assumption that the set of index terms T doesn't change in the application, all tolerance classes of terms are considered as structural subsets: $P(I_\theta(t_i)) = 1$ for all $t_i \in T$. Finally, the lower and upper approximations of any subset $X \subseteq T$, can be determined with obtained tolerance relation respectively $R = (T, I, v, P)$ respectively as

$$L_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) = 1\}$$
$$U_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) > 0\}$$

One interpretation of the given approximations can be as follows: if X is treated as a concept described vaguely by index terms it contains, then $U_R(X)$ is the set of concepts that share some semantic meaning with X, while $L_R(X)$ is a "core" concept of X.

III. ENRICHING WEB PAGE REPRESENTATION

In standard VSM, a page is viewed as a *bag of words/terms*. This is articulated by assigning, weight values, in page's vector, to terms that occur in page. With TRSM, the aim is to enrich representation of web page by taking into consideration not only terms actually occurring in web page but also other related terms with similar meanings. A "richer" representation of web page can be acquired by representing web page as set of tolerance classes of terms it contains. This is achieved by simply representing web page with its upper approximation. Let $p_i = \{t_{i1}, t_{i2} \ldots\ldots t_{ik}\}$ be a web page in P *and* $\{t_{i1}, t_{i2} \ldots\ldots t_{ik}\} \in T$ are index terms of p_i :

$$U_R(p_i) = \{t_i \in T \mid v(I_\theta(t_i), p_i) > 0\}$$

The example below describes how to enrich the web page representation using TRSM. TRSM based experiments were conducted on a database of 449 web pages and small part of this database is used here to describe web page enrichment method. Total number of terms in this small database is 14. The keywords extracted from the pages are indexed by their order of appearance, that is t_1 = { architect}, t_2 = {build}, t_3 = {biographi}, t_4 = {inform}... t_8 ={ceram}, t_9={potteri},... t_{11}={air}...and so on. The Term co-occurrence matrix for all terms is shown in table 1. The value '3' in second row and first column of the table 1 means that the terms t_1 and t_2 have occurred together in 3 different web pages.

Table 1. Term Co-occurrence matrix

	t1	t2	t3	t4	t5	t6
t1	-1	3	0	1	0	0
t2	3	-1	0	2	0	1
t3	0	0	-1	2	0	0
t4	1	2	2	-1	2	1

Tolerance class of a term is defined as $I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\}$ where $f_D(t_i, t_j)$ denotes the number of web pages in which both terms t_i and t_j occurs. In the example, for θ=3, tolerance class for term t_1 is:

$$I_\theta(t_1) = \{t \mid f_D(t_1, t) \geq 3\} \cup \{t_1\}$$
$$= \{t_1, t_2\}$$
$$Similarly, I_\theta(t_2) = \{t_1, t_2\} \quad .$$
$$I_\theta(t3) = \{t_3\}$$
$$I_\theta(t_4) = \{t_4, t_8, t_9, t_{11}\} \; and \; so \; on.$$

Table 4.2 shows ten web pages with the terms actually appearing in the web page and upper approximation of the page. Upper approximation of a page is defined as $U_R(p_i) = \{t_i \in T \mid v(I_\theta(t_i), p_i) > 0\}$ where $v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|}$. The concept X in the definition is the "web page", that means X contains all the terms appearing in the page. Upper approximation for web page p_1 is calculated as follows: Concept X, for web page p_1 contains all the terms appearing in that page.

$$v(I_\theta(t_1), P_1) = \frac{|I_\theta(t_1) \cap P_1|}{|I_\theta(t_1)|}$$
$$= \frac{|\{t_1, t_2\} \cap \{t_1\}|}{|\{t_1, t_2\}|}$$
$$= 1/2 > 0$$
$$v(I_\theta(t_2), P_1) = \frac{|\{t_1, t_2\} \cap \{t_1\}|}{|\{t_1, t_2\}|}$$
$$= 1/2 > 0$$
$$v(I_\theta(t3), P_1) = \frac{|\{t_3\} \cap \{t_1\}|}{|\{t_3\}|}$$
$$= 0/1$$
$$= 0$$

 ISSN: 2454-4248
281 – 285

$$v(I_\theta(t_4), P_1) = \frac{\left|\{t_4, t_8, t_9, t_{11}\} \cap \{t_1\}\right|}{\left|\{t_4, t_8, t_9, t_{11}\}\right|}$$
$$= 0/4$$
$$= 0$$

Similarly, vague inclusion $v(I_\theta(t_i), p_i)$ of all terms in page p_1 is calculated and all the terms whose vague inclusion is greater than zero will be added to the upper approximation of the page p_1 . Upper approximation for page p_1 is: $U_R(p_i) = \{t_1, t_2\}$.

In table 2, web page p_1 contains the term t_1 which is "architect" and upper approximation contains terms t_1 and t_2 that is "architect" and "build", which are conceptually related. That means, the upper approximation of the web page contains the terms appearing in the page and conceptually related terms of these terms. The upper approximation of the page p_3 contains the terms t_3, t_4, t_8, t_9 and t_{11} which are biographi, inform, Ceram, potteri and air respectively. These terms are not semantically related but they are appearing in upper approximation. This is due to the inappropriate value chosen for threshold θ. So care must be taken while choosing the threshold value. In the example if θ=13, then upper approximation of page 3 contains the terms t_8, t_9 which are semantically related.

Table 2. Upper approximation for 10 web pages

	Terms	Upper approximation
1	t_1	t_1, t_2
2	t_1	t_1, t_2
3	t_3, t_4	$t_3, t_4, t_8, t_9, t_{11}$
4	t_1, t_2, t_4	$t_1, t_2, t_4, t_8, t_9, t_{11}$
5	t_{11}	t_4, t_8, t_9, t_{11}

IV. IMPLEMENTATION AND RESULTS

In vector space model (VSM), a web page is represented by the terms that appear in the page. Since the feature space of this representation is high, feature selection is applied on this representation to reduce the feature space. The reduced dataset is generated after applying FS. This reduced dataset will also be in VSM which do not capture conceptually related terms. TRSM is applied on this reduced dataset to enrich the web page representation.

Figure 2 shows the reduced dataset. The extended representation of this reduced data set is shown in figure 3. In both the figures, rows represents the web page, columns represents terms ( ), "1" represents the presents of the term and "0" represents the absence of the term in web page. In figure 2, web page 42 contains the term 8, where as in extended representation (see figure 3), the same web page contains the terms 8 and 9.

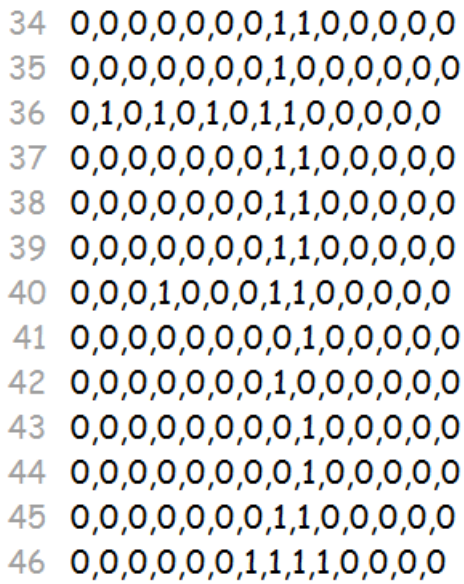

Figure 2. Web –page representation before applying TRSM

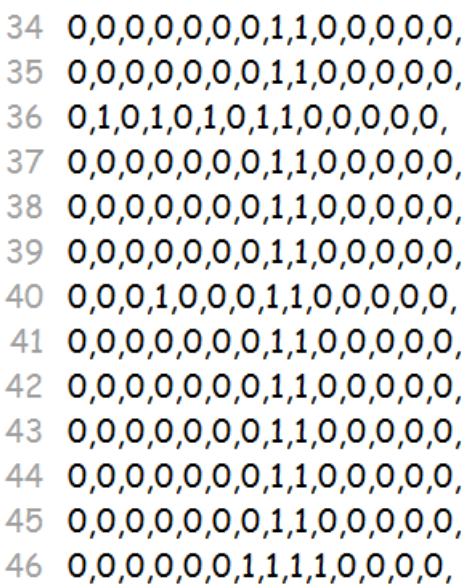

Figure 3. Enriched web page representation

The classification performance is evaluated for TRSM based web page representation and the accuracy of this representation is compared with the non-TRSM based web page representation (or VSM web page representation). Results of the classification performance are shown in table 3. From the table, it can be observed that TRSM based web page representation yields a higher accuracy compared to the normal vector space model.

 ISSN: 2454-4248
281 – 285

Table 3. Classification accuracy

Web page representation	Accuracy (%)	
	Non-TRSM	TRSM
Title of the web page	75.0557	76.5011
Meta data of the page	91.0594	93.2574

V. CONCLUSION

In general, Vector Space Model (VSM) will be used for web page representation. This model does not capture the conceptually related words. In this work, web page representation is enriched by adding the conceptually related terms to the web page. Tolerance Rough Set Model, which is an extension of traditional RST, was used to enrich the web page representation. Since enriching the original web page further increases the dimensionality, TRSM is applied on after applying FS on web pages. TRSM based web page representation has increased the classification accuracy obtained after applying Feature Selection by 2-3%.

REFERENCES

[1] Saori Kawasaki, Ngoc Binh Nguyen, T. B. H. Hierarchical document clustering based on tolerance rough set model. In Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000.

[2] Tu Bao Ho, N. B. N. Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems 17, 2 (2002), 199-212. http://dir.yahoo.com.

[3] T. Tsukada, M. Washio and H. Matoda. Automatic web-page classification by using machine learning methods. Proceedings of the First Asia-Pacific Conference on Web Intelligence(WI2001), LNAI 2198:303–313, 2001.

[4] Toshiko Wakaki, H. Itakura and Masaki Tamura. Rough Set-Aided Feature Selection for Automatic Web-page Classification. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2004.

[5] G. Salton, A. Wong and C.S. Yang. A vector space model for automatic indexing. Communications of the ACM, Vol. 18, No. 11, pp. 613–620. 1975