A rough-fuzzy document grading system for customized text information retrieval

Shailendra Singh, Lipika Dey *

Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India

Received 25 April 2003; accepted 24 September 2003

Abstract

Due to the large repository of documents available on the web, users are usually inundated by a large volume of information, most of which is found to be irrelevant. Since user perspectives vary, a client-side text filtering system that learns the user's perspective can reduce the problem of irrelevant retrieval. In this paper, we have provided the design of a customized text information filtering system which learns user preferences and modifies the initial query to fetch better documents. It uses a rough-fuzzy reasoning scheme. The rough-set based reasoning takes care of natural language nuances, like synonym handling, very elegantly. The fuzzy decider provides qualitative grading to the documents for the user's perusal. We have provided the detailed design of the various modules and some results related to the performance analysis of the system.

Keywords: Text information retrieval; Rough-set based reasoning; Fuzzy membership; Document relevance computation; User preference learning

1. Introduction

The World Wide Web, with its large collection of documents, is a storehouse of information for any user. Search engines help users locate information. But these search engines usually return a huge list of url's which are ordered according to a general relevance computation function. Most of the users find a large proportion of these documents to be irrelevant. However, since no two users usually have identical perspectives it is very difficult to find a general relevance computation function that can satisfy all users simultaneously. It is also not feasible to load a server with profiles of all the clients to serve them better. A viable way to provide relevant documents to every user is to use client side information filtering systems, which can learn a client's perspective and grade documents according to a relevance function specific to the client.

In this paper, we have presented the design of a client-side text information filtering system based on a rough-fuzzy reasoning paradigm. This can pro-actively filter out irrelevant documents for a user, in his or her domain of long-term interest, after learning the user's preferences. To begin with, the user rates a set of training documents retrieved as a result of posing a query to a standard search-engine. The user response is then analyzed to formulate a modified query which represents the user's interests in a more focused way. This modified query is again fed to the search engine and has been found to retrieve better documents. However, since these documents are also ordered by the grading scheme of the search engine, their ordering do not still reflect the client's preferences. A rough-fuzzy grading scheme is thereafter employed to re-evaluate these documents and order them according to the user preferences.

The most unique aspects of this work are:

- The use of discernibility to represent the user's relevance feedback.
- We have presented how a client-side user relevance feedback based text retrieval system can be designed using rough-fuzzy reasoning. Most of the existing systems of this category use probabilistic reasoning (Intarka, Inc., 1999; Pazzani, Muramatsu, & Billsus, 1996). Rough-Fuzzy reasoning paradigm helps in modeling natural language based information more elegantly through the use of equivalence relations.

The remaining paper is organized as follows. Section 2 presents a brief review of related work on text-filtering systems in general and also on application of rough-set theory to text-information retrieval. Sections 3-7 present the details of the various modules of our system. Section 8 provides some results and their analysis.

2. Review of related work

Significant work has been done towards building client side text retrieval systems based on user ratings. In this section, we first provide a brief overview of these. Later in this section we present some of the recent developments in applying rough sets for text information retrieval.

2.1. User preference based text information retrieval

User preference plays a major role in text information retrieval. Different schemes are used to store and analyze user preference. Currently, many interactive systems are being designed to provide better interfaces, simple interaction metaphors and learn user's preferences. WebPlanner (Jochem, Ralph, & Frank, 1999) is a system that guides users towards their needs by using structured domain-specific queries. However, it is not feasible to have domain-specific structured queries to be stored for all possible domains.

A better way to provide individual satisfaction is to use a dynamic representation for user's preferences. One of the important advances in this area is the consideration of user profiles. User profiling aims at *determining a representation of the user preferences so that the stored values may serve as input parameters for a filtering action operating on the available offer* (Roldano, 1999). User profiling is often coupled with learning from user feedback. Relevance feedback by the user provides an assessment of the actual relevance of a document to a query. This can be utilized effectively to better the performances of retrieval mechanisms (Bodoff, Enache, Kambil, Simon, & Yukhimets, 2001; Korfhage, 1997). A relevance feedback method is either based on a document-oriented view or a query-oriented view. In the document-oriented view (Salton, 1971), the users' feedback are used to change the document's internal representation to the search engine. This allows the return of similar documents to a set of similar queries that are posted by different users. However most of the recent IR systems are built using the query-oriented view. Systems built around this model use the user feedback to modify the initial query posted by the user and tries to improve the retrieval performance. This method was initially proposed by Rocchio (1971). Salton, Fox, and Voorhees (1985) proposed a query expansion technique based on the extended Boolean model. Crestani (1993) proposed a neural-network based method which learnt the user preference through adjustment of initial weights to get the desired performance. Allan, Ballesteros, Callar, and Croft (1995) presents an overview of TREC experiments related to query expansion. Fuhr and Buckley (1993) use a massive expansion of query using co-occurrence of words in good documents. "ProspectMiner" (Intarka, Inc., 1999) is another retrieval system that learns user's interests based on a user rating. The retrieval system suggests better queries that will fetch more relevant pages. The software agent also takes into account the co-occurrence and nearness of the words. Apart from the document rating, the retrieval system requires a term-feedback from the user and maintains a thesaurus with respect to the words present in the initial query.

A third approach uses the relevance feedback from the user to eliminate bad documents in future. These are usually installed at the client side and are client-specific. ‘‘Syskill & Webert’’ (Pazzani et al., 1996) is a client-side software agent that learns to rate web pages based on the user's rating of a set of training pages. The system converts HTML source codes of a web page into a boolean feature vector of words, indicating the presence or absence of words. It then analyses the user's feedback to determine the words to be used as features by finding the expected information gain that the presence or absence of a word W gives toward the classification of elements of a set of pages. A Bayesian classifier is used to determine the degree of interest of a new page to the user. Balabanovic's Fab system (Balabanovic, 2000) recommends web sites to users based on a personal profile that has been adapted to the user over time. Individual user ratings of pages are used to generate the user's profile adaptively, so that the recommendations gradually become more personalized. These systems have used probabilistic measures for judging the relevance of a document to a user.

2.2. Rough-set based text information retrieval

The systems presented above mostly worked with two-valued crisp logic to reason with user preferences. The chief problem with this approach is that it cannot handle complexities of natural language like synonymous words or polymeric words etc.

Rough-set based reasoning technique proposed by Pawlak (1982) provides a granular approach to reasoning. Rough sets are a tool to deal with inexact, uncertain or vague knowledge. Specifically, it provides a mechanism to represent the approximations of concepts in terms of overlapping concepts. Stefanowski and Tsoukias (2001) have shown how rough reasoning can be applied to classify imprecise information. Srinivasan, Ruiz, Kraft, and Chen (2001) and Das-Gupta (1988) have proposed the use of rough-approximation techniques for query expansion based on this model. Bao, Aoyama, Du, Yamada, and Ishii (2001) have developed a hybrid system for document categorization using latent semantic indexing and rough-set based methods. This system extracts a minimal set of co-ordinate keywords to distinguish between classes of documents. Chouchoulas et al. have shown the applicability of rough-set theory to the information filtering by categorizing e-mails in Chouchoulas and Shen (2001). Jensen et al. have used rough-set theory for automatic classification of WWW bookmarks in Jensen and Shen (2001). Menasalvas, Millan, and Hochsztain (2002) have provided a rough-set based analysis of affability of web pages. They have also used rough-set based approaches to compute user interest parameters for web usage mining.

Since documents cannot be categorized uniquely on the basis of presence or absence of words, we argue that a rough-reasoning scheme is very appropriate to design text-retrieval systems. We have used the rough-theoretic concept of discernibility to analyze a set of user feedback. This analysis provides us with an enhanced query which better represents the user interest. Using rough sets allows us to handle synonymous words very efficiently. However, since the relationships among various concepts in the real world are vague, so a mechanism is needed to model the various degrees of equivalence (Srinivasan et al., 2001; Szczepaniak & Gil, 2003). We have thus opted for a rough-fuzzy approach to design a text retrieval system which uses the user's interests to rate a set of new documents effectively for the user.

3. A brief overview of rough-set based reasoning for text information retrieval

Rough sets were introduced by Pawlak (1982). An information system can be defined as a pair A = *ðU;AÞ* where *U* is a non-empty finite set of objects called the universe and *A* is a non-empty finite set of attributes. For every *a 2 A*, V_a*ðxÞ* represents the value of attribute *a* for object *x*. An information system is called a decision system if it has an additional decision attribute. The core of all rough-set based reasoning contains an equivalence relation called the indiscernibility relation. For any *B* ⊆ *A*, the equivalence relation IND_AðBÞ is defined as:

$$R = \mathrm{IND}_A(£) = \{(x,x>)\ \mathrm{G}\ U^2\backslash Vb\ \mathrm{G}\ B, V_b\{x) = V_b\{x')\} \qquad (1)$$

This relation is called a B-indiscernibility relation. We denote the equivalence classes of this relation as $[x]_B$. In the context of text-information retrieval the equivalence relation that is generally used is the synonymy relation which establishes equivalence of two synonymous words. Thus two texts can be said to be roughly similar if they contain synonymous words but not necessarily the same words.

The equivalence classes obtained from the indiscernibility relation are used to define set approximations.

Let *B* ⊆ *A* and *X* ⊆ *U*. Let *U* be represented by the collection of disjoint equivalence classes with respect to the relation *R*, i.e., *U* = fC1; C2;...; C_n} = $[x]_B$. The pair *ðU;RÞ* is called an approximation space. The lower approximation of *X*, denoted by $\underline{\mathrm{apr}}_R$ *ðXÞ* is defined by the set

$$\underline{\mathrm{apr}}_R\ (X)\ =\ x2Cij\ \subseteq \mathrm{I}\ \} \qquad (2)$$

and the upper approximation of X, denoted by $\overline{\text{apr}}$RðXÞ is defined by the set

$$\overline{\mathrm{apr}}_R(X) = \{x \in C_i | C_i \cap X \neq \text{- 4>\}} \qquad (3)$$

The objects in $\underline{\text{apr}}_R$ ðXÞ can be definitely classified as members of X on the basis of the information in *B*, while the objects in $\overline{\text{apr}}$R ðXÞ can only be classified as possible members of *X*. In crisp set theory, the similarity of two subsets can be defined as their degree of overlap. In rough-set theory, two subsets of the universe can be compared with respect to an indiscernibility relation using their approximations.

3.1. Rough similarity measures for text documents

We will now state some rough similarity measures introduced in (Srinivasan et al., 2001) to compute document overlaps. Two approximation spaces are first introduced to measure the degree of overlap between two subsets. Let *S1* and *S2* represent two subsets, which are collections of weighted words. Let *S1* represent the words in a retrieved document and *S2* represent the words in a query. One would be interested to find the similarity between the query and the document in order to judge the relevance of the document with respect to the query. Let *R* denote the synonymy relation and *S* be a set of weighted words. Let l_S*ðy* G [0,1] denote the degree of synonymy of the word *y* with *x*. The lower and upper fuzzy approximations of the word *x* are defined as follows:

$$\begin{aligned} \mu_{\underline{\mathrm{apr}}_R(S)}(x) &= \inf\{\mu_S(y) | y \in U, (x,y) \in R\} \\ \mu_{\overline{\mathrm{apr}}_R(S)}(x) &= \sup\{\mu_S(y) | y \in U, \text{(x,y) G R\},} \end{aligned} \qquad (4)$$

Eq. (4) helps us in finding the word *y* which has the minimum/maximum degree of equivalence withx. Since the presence of a word can be related to the presence of its synonyms also, where the degree of equivalence can be judged by the degree of their synonymy, one can compute rough approximations for the entire set of words in a query *S* using the degree of synonymy of two words.

Let l_R*ðx,yÞ* denote the degree of synonymy between words *x* and *y*. Then fuzzy lower and upper approximations for the word *x* is computed as

$$\begin{aligned} \mu_{\underline{\mathrm{apr}}_R(S)}(x) &= \inf\{1 - \text{l*}_R\text{(x,y)\textbackslash y £ Sg} \\ \mu_{\overline{\mathrm{aprs}}}(\text{.})\,(x) &= \sup\{\mu_R\ \text{(x,yÞjy 2 Sg} \end{aligned} \qquad (5)$$

The two functions are now combined to give:

$$\begin{aligned} /\underline{V}_s(s)M &= \text{inf\{max[/i}_5\text{(j), 1 - n}_R\text{\{x,y)\textbackslash\textbackslash y G U\}} \\ /\text{«5pf}_{s(s)}M &= \text{sup\{min[n}_s\text{(y),n}_R\text{(x,y)]\textbackslash y G U\}} \end{aligned} \qquad (6)$$

The approximations for a set *S* are computed as union of the fuzzy approximations of all the words occurring in it. These approximations are then used to find the similarity between two-weighted set of words *S1* and S_2, where as defined earlier *S1* represents the words in the document and S2 represents those in the query. We will explain in the next few sections how we have used these concepts of rough theoretic analysis and enhanced them to perform customized text information filtering by our proposed system.

4. Architecture of the customized text filtering system

In our system text documents are represented as weighted vector of words like that used by google (Google Search Engine Optimization, 1999). The information system for classifying documents is constructed by taking words to represent attributes, and their weights in documents to represent the values of these attributes. The information system is converted to a decision system by including the user relevance feedback for each document.

This decision table is analyzed using rough-set based reasoning techniques to generate a user profile and provide a basis for more focused relevance computation, which can eliminate irrelevant documents for the user effectively in future. Fig. 1 gives a schematic view of the complete

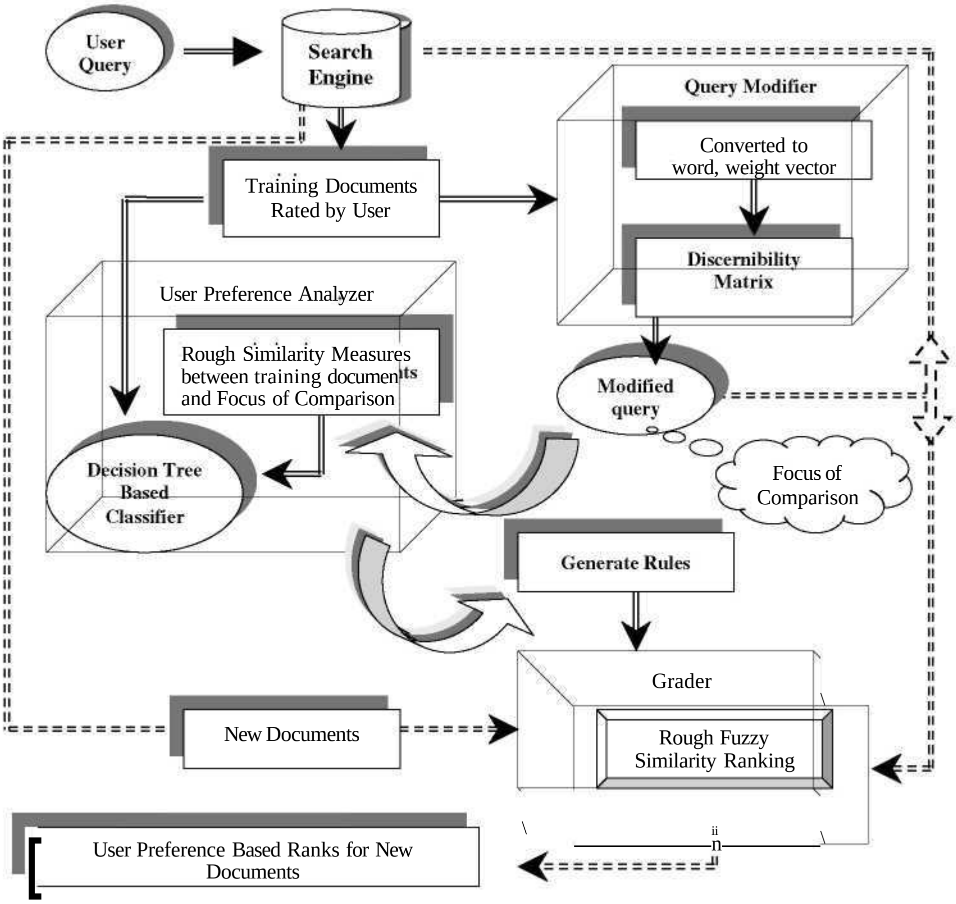

User Query
Search Engine
Query Modifier
Converted to word, weight vector
Discernibility Matrix
Training Documents Rated by User
User Preference Analyzer
Rough Similarity Measures between training documents and Focus of Comparison
Modified query
Focus of Comparison
Decision Tree Based Classifier
Generate Rules
Grader
Rough Fuzzy Similarity Ranking
New Documents
User Preference Based Ranks for New Documents

Fig. 1. Schematic view of the customized text information retrieval system.

system. The system works in co-operation with a backend search engine. Once the user specifies a query to the search engine, a training set is formed with a subset of the top graded documents to get the user feedback. The user is asked to rate each training document on a three-point scale in which 1 stands for bad, 2 for average, and 3 for good. The system then starts analysing the training set for generating the user profile and the grading scheme for future documents. Grossly, the functions of the various modules are as follows:

Query modifier—This module generates the modified query that can retrieve better documents for the user. Every training document which is rated by the user is converted to a weighted vector of words. The most unique aspect of our system is the introduction of the rough-set theoretic concept of *discernibility* to identify words which help in distinguishing between relevant and irrelevant documents. A *discernibility matrix* is constructed based on the user's rating and the weighted word vector for each training document. The matrix is scanned to extract the set of most discerning words which yields the *modified query*. We propose to identify the most discerning words from a document as follows:

(i) Words which are present with a high relative importance in good documents and are not present or have low importance in the bad documents are *positively discerning* words, and are desirable in a document.
(ii) Words which are present with a high relative importance in bad documents and have low importance in the good documents are *negatively discerning* words, and are to be avoided while searching for relevant documents.

We use the positive and negative discerning words to formulate an improved search query for the user. The details of this are presented in Section 5. We will show that this can also imbibe the user's preference. When this modified query is fed to the search engine, it is observed that the documents returned are in general better. However, the relative ordering of the documents is influenced by the relevance computing function of the underlying search engine. The documents are therefore graded according to user preferences.

User preference analyzer—This module uses *rough similarity measures* to learn the user preferences for rating documents. Srinivasan et al. (2001) had introduced some similarity measures for queries and documents, though it was not indicated how these measures can be actually used to judge the relevance of a document. We observed that the values of similarities do not provide any indication about the actual relevance of a document, rather it is the range of the similarity values and their co-relations, that provide indication about the quality of documents in a domain specific way. Based on these observations, we have developed a complete grading scheme using these measures. The modified query is used as the focus of comparison. A new decision system is constructed using the ratings of the training documents and their lower and upper similarity measures with respect to the modified query. This table is then used to extract fuzzy grading rules, which relate the similarity measures to the user rating decisions. Section 6 presents the details of the scheme.

Grader—This module grades the documents retrieved with the help of the modified query, using the fuzzy rules extracted by the user preference analyzer. Now rough similarity measures between the modified query and each new document is computed, keeping the query as the focus of comparison. The classification rules are then employed to assign a rating to the document.

The rating is expressed both as a crisp decision value and a fuzzy membership to a decision class.

We explain the detailed design of each of these modules in the next section.

5. Query modifier—forming modified query with most discerning words

As stated in the earlier section, every training document is converted to a weighted vector of words appearing in the document. To calculate the weights of the words, we use the HTML source code of the pages. Since each tag like ÆTITLEæ, ÆBæ etc. in an HTML document has a special significance, we assign separate weights to each one of them. We have given tag weights in the range of 1-10, with 10 for ÆTITLEæ, then 8 for ÆMETAæ, 6 for ÆBæ etc. The plain words have a weighing factor 1. The weight of a particular word *s* is then computed as follows:

$$W\{s) = \sum_{i=\backslash}^{m} w_t \text{ x } n_i$$

where w_i represents the weight of tag *i* and n_i represents the number of times the word *s* appears within tag *i*, and *m* is the total number of tags (including no tag) considered. The weights are normalized by dividing the weight of each word in a document by the maximum weight of a word in that document. We use a word vector of size *n*, where *n* P 30 to represent the document. All our results have been obtained with *n* = 50, as no significant improvement has been observed with *n* > 50. We now take a look at the decision table constructed with the weighted word vectors. Suppose the user has rated the documents *D1*, D_2, *D3*, and D_4 as 1, 2, 3, and 3 respectively. Further, suppose that *D1*, D_2, *D3*, and D_4 have words *W1*, W_2, *W3*, and W_4 with weights as shown in Table 1.

Table 1 shows that the word *W3* does not have the capacity to distinguish between ‘‘good’’, "bad" or an 'average' document since it has a high weight in all of them. On the other hand, the word *W1* has the potential to distinguish a 'bad' document from a 'good' one. Thus we can say that *W1* may be a negatively discerning word. Similarly it may be argued that W_2 and W_4 are positively discerning word.

If we can extract the most discerning words using the decision table, these can be used to formulate a modified query as follows. The query is constructed as a Boolean function of all positive and negative discerning words. Positive discerning words are indicated as desirable words in the documents while negatively discerning words are indicated as undesirable. For example, a query ‘‘W_2 þ W_4 — *W\"* would indicate that we want documents which contain W_2 and W_4 but do

Table 1
A decision table D_T for documents

Documents	W_1	W_2	W_3	*W4*	Decision
A	1.0	0.1	0.9	0.2	1 (bad)
D_2	0.5	0.5	0.9	0.75	2 (average)
D_3	0.0	1.0	1.0	0.9	3 (good)
$Z>_4$	0.2	0.9	0.9	0.9	3 (good)

not contain *W1.* Now we will explain how the most discerning words can be extracted from the decision table.

Let us suppose the number of distinct documents in the training set is N and the number of distinct words in the entire training set is k. We will now show how the *discernibility table* (Komorowski, Polkowski, & Andrzej, 1999) for this set is constructed. For each distinct word in the domain, its weights (in different documents) are arranged in ascending order. An interval set P_S is then constructed for the word *s*, which is defined as

P_s = {[/o,/i), [/i,/2), • • •, Mr+i)}, where $I_s = I_o < h < h < \cdots < h <$ A+i = *U* (7)

For each interval in the interval set, the mid point of the interval is called a *cut.* Each distinct word *s* is thus associated with a set of cuts.

{(s,c\), (s,c$_2$), *{s,* c$_3$),..., *(s,* c$_r$)} C *A* x 91, where c_t is the mid point of [/,-!, I_i] (8)

Since each word may not be present in all the documents, the number of intervals and therefore the number of cuts may be different for different words. Let us suppose word s_i has p$_i$ cuts. Then the total number of cuts for the entire set of words is P$_i$ p$_i$ where 1 6 *i* 6 k.

Let D^*_T denote the discernibility table. D^*_T is constructed with help of decision Table 1 and the cuts. *Dj* has one column for each cut induced over D_T, and one row for each pair of documents *(Di,Dj)* where D_i and *Dj* have different user categorizations i.e. different decisions. An entry *v\-* in *Dj* is decided as follows:

jk = 0 *in Dj, if the document pair* D_i *andDj have different decisions but the weight of the wordk in both the documents are on the same side of the cut,*

kij = di — dj, if the weight of the wordk in document i is more than the cut and the weight of the word in document j is less than the cut, and the documents have different decisions di and dj respectively.

Otherwise, t^. = *dj* — d_i.

Thus, a non-zero entry corresponding to a word *s* in D^*_T, indicates that the word has two different significance levels in two documents of different decisions. The absolute value of the entry determines the power of the word to distinguish between two different categories. A negative value indicates that the word has a higher weight in a bad document than in a good document, which means that the word may be a negatively discerning word. Table 2 denotes the discernibility table *Dj* constructed from the decision table presented in Table 1.

Finally, we will now analyze D^*_T, to get the most discerning words and the corresponding values of the cuts. Since, theoretically, there can be an infinite number of cuts possible, one can apply the

Table 2
Discernibility table

Document pairs	*W1,0.1*	W_u 0.35	W_u 0.75	W_2, 0.3	W_2, 0.7	W_2, 0.75	W_3, 0.95	W_4, 0.47	W_4, 0.82
(A, A)	0	0	11	1	0	0	0	1	0
(A, A)	)2	)2	22	2	2	2	2	2	2
(A, A)	0	)2	22	2	2	2	0	2	2
(A, A)	)1	01	0	0	1	1	0	0	1
(A, A)	0	01	0	0	1	1	0	0	0

MD-Heuristic algorithm presented in Komorowski et al. (1999) to obtain the minimal set of maximal discerning cuts.

However, since the original MD-Heuristic algorithm works with a discernibility table in which all decision differences were considered as identical, we have modified this algorithm to find the most discerning words. For this, we first consider those columns which induce the highest degree of difference in decision, followed by the next highest and so on, till there are no more discerning words in the set.

The steps in the modified MD-Heuristic algorithm followed by us are:

Step 1: Let W denote the set of most discerning words. Initialize W to NULL. Initialize $T = r$, where r is the maximum difference in decision possible (in our case 2).

Step 2: For each entry in D^*_T consider the absolute value of the decision-difference stored there. If none of the absolute values are equal to T, then set $T = T — 1$, if $T = 0$ then stop else go to step 3.

Step 3: Considering the absolute values of decision difference, choose a column with the maximal number of occurrences of Ts—*this column contains the word and the cut that is discerning the maximum number of documents corresponding to the current level of discernibility.*

Step 4: Select the word w^* and cut c^* corresponding to this column. In case of a tie, the leftmost column is chosen. Delete the column from D^*_T. Delete all the rows marked in this column by T since this decision-discernibility is already considered. Delete all columns for w^* from D^*_T.

Step 5: If majority of the decision differences for this column are negative, then the word is tagged with a ()) sign to indicate that it is a negatively discerning word. Otherwise it is tagged with a positive sign (+) to indicate that it is a positively discerning word.

Step 6: Add the tagged word w^* and cut c^* to W.

Step 7: If there are more rows left, then go to step 2. Else stop.

This algorithm outputs a list of words along with their cut-values, which collectively discern all pairs of documents rated by the user. The presence of the positively discerning words and the absence of negatively discerning words are desirable in good documents. A *modified query* is constructed using these words and Boolean operators. The modified query is fed to the search engine again. It is observed that performance improves significantly. However, some irrelevant documents are still retrieved and the list is not ordered according to the user preference. In the next section, we will elaborate on how the irrelevant documents can be filtered out from this set.

6. User preference analyzer—learning the users basis for rating

To help the system rate the newly fetched documents and eliminate irrelevant ones, it is essential to learn the user's rating paradigm. For this we make use of rough similarity measures between the modified query and the original documents that were rated by the user. Let $S1$ denote the set of words along with their weights, extracted from a document as explained in Section 5. Let S2 denote the set of most discerning words along with their discerning cut values. Using Eqs. (5) and (6) of Section 3.1, one can obtain the lower and upper approximations for each word. The equivalence relation *R* used is the "synonym" relation. In general, the synonym dictionary is constructed using WordNet and each word occurring with a different sense is assigned the same weight. However, the dictionary does vary to some extent according to the domain. For example the word "can" which is usually a stop word for most of the domains, becomes a word synonymous to "dustbin" for the domain ‘‘Air pollution’’.

Let $\underline{\mathrm{apr}}\mathrm{R}$ ðSÞ and $\overline{\mathrm{apr}}\mathrm{R}$ ðSÞ denote the lower and upper approximations of a set of words S respectively. These approximations can be computed using Eq. (5) as follows:

$$\underline{\mathrm{apr}}_R(S)\ \mathrm{w\,H/\underline{V}_s M > 0g}$$

$$\overline{\mathrm{apr}}_R(S) = \{w\backslash fi\wedge_s(w) > 0\}$$

Here l_R *ðx; yÞ* is the degree of synonymy between words *x* and *y*, while l_S *ðxÞ* is the weight of the word *x* in the set S. The difference in the lower and upper approximations for the sets S1 representing the document, with respect to the lower and upper approximations of the set of most discerning words represented by S_2 are computed as follows:

$$B_1 = \underline{\mathrm{apr}}_R(S_2)| - |(\underline{\mathrm{apr}}_R(S_1) \cap \underline{\mathrm{apr}}_R(S_2)) \quad \text{and} \quad B_u = \overline{\mathrm{apr}}^*(S_2)| - |(\overline{\mathrm{apr}}_R(S_1)\ \mathrm{n}\ \overline{\mathrm{apr}}^*(S_2)) \tag{9}$$

where | — | represents the bounded difference. *Bl* is called the *lower approximation* of subset S_2 with S1 and B_u is the *upper approximation* of subset S_2 with *S1.*

With these approximations, the similarity of two subsets *S1* and S_2 is defined in Srinivasan et al. (2001) as

$$\underline{\mathrm{Similarity}}_R\ \mathrm{ðS1;\ S_2Þ} = 1 - \left[\frac{\mathrm{card}(B_1)}{\mathrm{cardð}\underline{\mathrm{apr}}_R\mathrm{ðS_2ÞÞ}}\right]$$

$$\overline{\mathrm{Similarity}}\mathrm{R}\ ðS_1;S_2Þ = 1 - \left[\frac{\mathrm{cardðB_uÞ}}{\mathrm{cardð}\overline{\mathrm{apr}}\mathrm{RðS_2ÞÞ}}\right] \qquad \text{ð10Þ}$$

where (10) denotes the lower similarity and upper similarity of S1 and S_2, considering S_2 as the focus in the comparison. In both the cases the value will be 0 for no match and 1 for maximum match between S1 and S_2. We have used the set of most discerning words which also serve as the modified query, as the focus of comparison.

Using Eqs. (9) and (10), we first compute the lower and upper similarities between each of the old documents and the set of most discerning words. Since there is no apparent unique association between the similarity measures and the user's rating, we decided to use a decision tree which can summarize the relationship as a set of rules. These rules typically relate the rating assigned to a document by the user to its similarity measures. The decision tree is constructed using the ID3 algorithm (Mitchel, 1997). Here are some typical rules generated by our system for the domain ‘‘alcohol addiction’’.

Rule 1: If Lower similarity > 0.027 and Upper similarity 6 *0.1111111 then Class = 1 (bad)* (9/2), where the number of training cases covered by the rule is 9 and 2 of them do not belong to the class predicted by the rule.

Rule 2: If Lower similarity 6 *0.01388* and *Upper similarity > 0.4305556 then Class = 1 (bad) (4/2).*

Rule 3: If Lower similarity 6 *0.01388, Upper similarity > 0.2777778 and 60.4166 then Class = 2 (average)* (7/3).

Rule 4: If Lower similarity > 0.01388 and Upper similarity 6 *0.52777 then Class = 3 (good)* (7/3).

We found that around 10 or 11 rules were generated for each domain.

7. Fuzzy grading of documents

The rules generated by the preference analyzer are used to rate the new set of documents retrieved using the modified query. For this we use a fuzzy reasoning scheme which provides both a crisp document grading as well as a fuzzy visualizer, to provide a qualitative idea about the relevance of a document.

Fuzzy reasoning consists of two core activities—editing the fuzzy input and output membership functions. To design the fuzzy input membership functions we have made use of the rules obtained earlier. The rules give us an idea about the cut-off values and the membership functions to be used for the input parameters i.e. the lower similarity and upper similarity and the class decisions. We have used the triangular function to represent the bad decision class, since the rate of change of quality of a document from bad to average or vice-versa is very steep. The average class is represented by the gaussian function, since it has a lower rate of change of quality. Finally we use the sigmoidal function to represent the good decision class since it is a right open function and indicates that once the quality of a document is judged good it remains so. Fuzzy Logic Toolbox also suggests use of similar functions for modeling linguistic variables like low, medium, high. The relationships of these functions with the input parameters are extracted from the rules generated by ID3.

To plot the bad decision function we consider all the ID3 rules that yield the class decision Bad. For each of the input parameters ''lower similarity'' and ''upper similarity'', we feed the range of these parameters for the Bad class as obtained from the rules. These ranges along with the type of the membership function used to represent the decision, generate the ultimate membership function curve for the class. For example the minimum and maximum values of lower similarity for the Bad decision class for the domain ''Alcohol addiction'' were obtained as (0, 0.01388). Similarly the least and maximum values for the upper similarity for the Bad decision class for the same domain are (0.027, 0.43). Since the membership function type for this is the triangular function, Fig. 2 shows the corresponding curve that was generated for the bad decision class. Membership function curves for the other decision classes are also chosen accordingly. We have used the MATLAB Fuzzy Logic Toolbox to generate the fuzzy membership values for the documents. Fig. 2 shows the example functions for the domain ''alcohol addiction''.

To rate a new document, it is first converted into a vector of weighted words. Using the modified query as the focus of comparison, we now compute the lower and upper similarity measures between the modified query and the new document using Eqs. (9) and (10). On feeding these values to the membership editor of the Fuzzy Tool Box, we get the membership of each document to all the three decision classes—good, average and bad. The user is presented with a graphical representation of the fuzzy membership values of each document, which is also generated by the Tool Box. This gives an intuitionist feel of the document to the user.

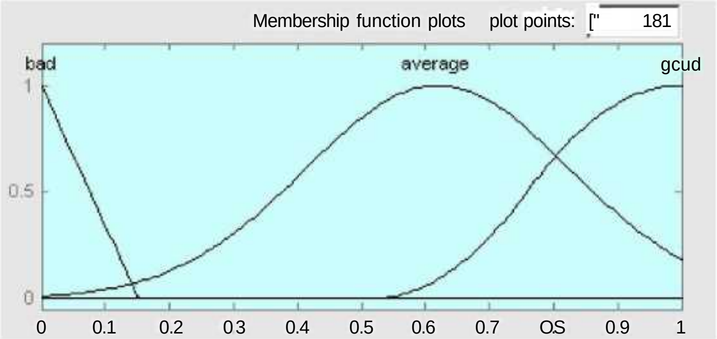


Fig. 2. Editing membership functions for different decision classes for ‘‘alcohol addiction’’.

7.1. Crisp rating of documents

For each document, a crisp membership value to each individual category may also be obtained. The document may be awarded the class with the maximum membership value. The url's are then re-arranged according to their crisp system ratings. Thus the good url's are presented first followed by the average and the bad url's. Using this approach, the system can filter out irrelevant documents by eliminating the ones which have maximum membership to the bad decision class altogether. This approach is also quite useful since the user does not have to take a look at irrelevant documents at all.

To determine the efficacy of this approach we had requested the users to rate the new documents also and then compared the user evaluations with our results. In the next section we present results obtained with different queries and present the success rate of the system grading scheme.

8. Results

In this section we will present some performance analysis of our system. We have worked with queries some of which like "HIV" and ‘‘alcohol addiction’’ were chosen because they had been mentioned in TREC topics. TREC mentioned ‘‘brain cancer’’ as a topic. But we worked with ‘‘Blood cancer’’ since we had less expertize in rating the other topic. Similarly, rather than ‘‘Thailand tourism’’ as mentioned in Chakrabarti et al. (1998) we chose ‘‘Indian Tourism’’ as a domain. We chose a new query ‘‘Alternative medicine’’ since many users have interest in this topic though from different perspectives. Table 3 shows the top 10 url's obtained with the initial query ‘‘alcohol addiction’’. However, as we can see, the user has rated 5 documents as bad and 5 documents as average out of 10. Using 50 documents retrieved from this query, we now find the most discerning words.

Table 4 shows these initial and modified queries along with ones we obtained for different other domains. Columns 2 and 4 of Table 4 show the percentage of bad documents among the top 50 url's obtained with the initial and the modified query respectively. The list of the top 50 url's for the two queries need not be same. Usually, the total number of documents retrieved also decrease

Table 3
List of top 10 url's corresponding to initial query ‘‘alcohol addiction’’

No.	List of top 10 retrieved url's using original query	User rating
1	http://center.butler.brown.edu/	1
2	http://www.well.com/user/woa/	2
3	http://www.thirteen.org/edonline/lessons/alcohol/alcoholov.html	1
4	http://www.odadas.state.oh.us/_GD_Frame Work/template s/fwTemplate001 .asp?FW = 1&ContentId = &enumerator = &search =	1
5	http: //www. health. org/	2
6	http://www.niaaa.nih.gov/	2
7	http://www.schick-shadel.com/	2
8	http://www.family.org/topics/a0018129.cfm	2
9	http: //www. healthreco very. com/	1
10	http://www.addiction-help-line.com/-	1

with the modified query as shown in Table 4. This is because the query is more focused now. Certain documents which were not retrieved earlier may be retrieved now, and similarly certain documents which were obtained with the initial query may not be retrieved now. Thus the top 50 documents may not be identical. We see that while 50% of the top 50 documents were bad with the initial query ‘‘alcohol addiction’’, with the modified query containing the words shown in column 3, only 10% of the top 50 documents are bad. The reduction in bad documents is substantial in all the domains as the table indicates. The decrease in the percentage of bad documents with the modified query proves the effectiveness of the modified query.

Different groups of users were asked to evaluate different domains depending on their interest in the topics. For each domain, the same user(s) has been asked to rate the initial and final set of documents to maintain uniform standards of rating. In all the cases, authoritative pages containing good documentation about the topic were rated higher than hub pages containing links to other pages.

Next we present some results to show the working of the fuzzy grading scheme. Table 5 presents the top ten retrieved documents using modified query from the domain alcohol addiction. Columns 3 and 4 of Table 5 show the lower similarity and upper similarity of each document with respect to modified query. Columns 5-7 of Table 5 show the fuzzy membership values of each document to bad, average, and good decision classes respectively, using Fig. 2.

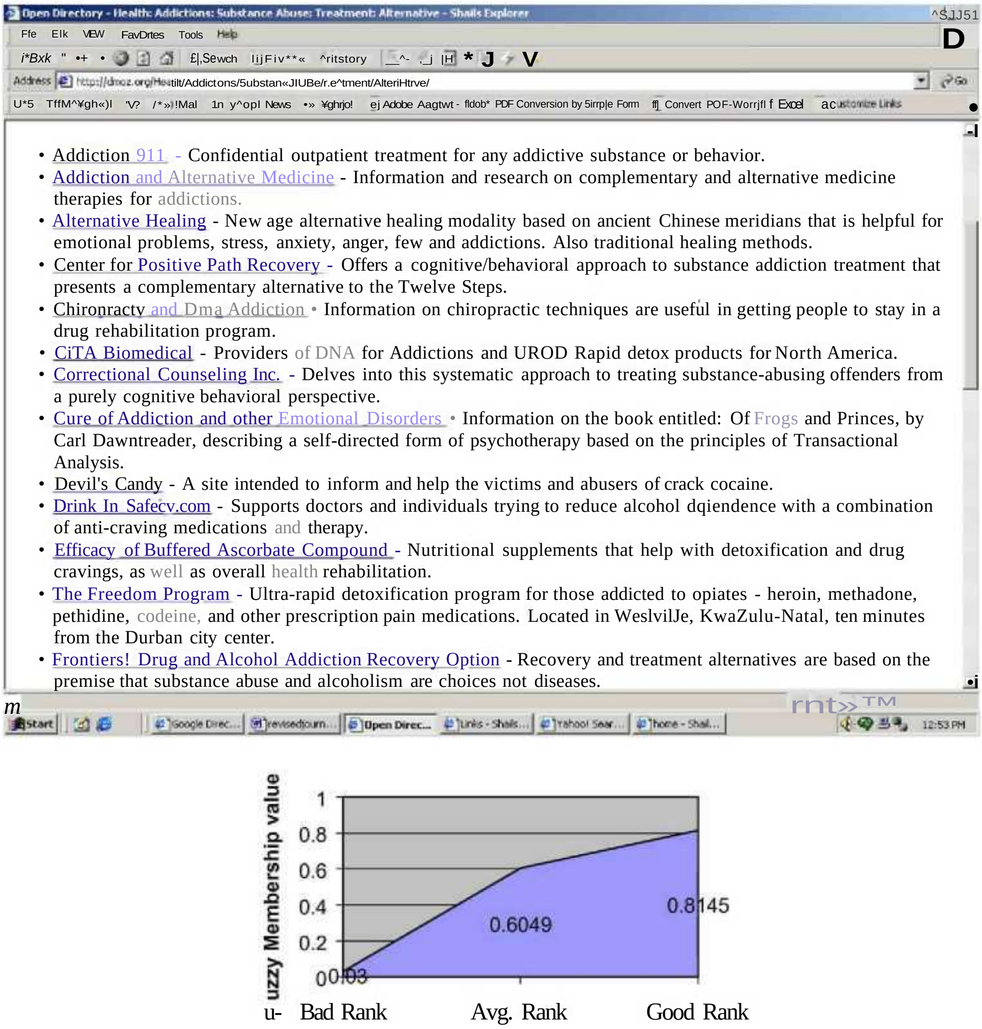
Figs. 3-5 present glimpses of some of the url's and their fuzzy memberships are presented in Table 5. These graphs give an idea about the quality of the documents to the user. Fig. 3 is corresponding to the url http://dmoz.org/Health/Addictions/Substance_Abuse/Treatment/Alternative/. This is a very informative url. We find that fuzzy membership value from the accompanying figure for Good category is also maximum. This url has a collection of links which give the information about alcohol addiction treatment, its prevention and important information about texts on alcohol addiction. Though this page is a collection of links, it provides adequate information about each link. Hence it gets a very high rating. Url numbers 2, 3, 6, 8, 9, and 10 in Table 5, all have similar lower and upper similarity measures—hence all of them have identical fuzzy membership values.

Table 4
Proportion of bad documents in top 50 url's corresponding to initial and modified query. Operator—indicates negatively discerning words which are to be avoided

Initial query and number of retrieved url's using Google	% of bad documents in top 50's from initial query	Modified query and total number of retrieved url's using Google search engine with modified query	% of bad documents in top 50's from modified query
Alcohol addiction 528,00	50	Alcohol + addictions + abuse + drugs + treatment + health + description + rehabilitation + help - revised 3270	10
Alternative medicine 1,910,000	54	Health + medicine + alternative + therapy + yoga + acupuncture + stress + diet + disease - agriculture - altvetmed - dosha 3940	12
Blood cancer	44	Cancer + health + medical + information + blood + leukaemia + help + myeloma + alive + symptom - companion - safety - poison 40	8
Air pollution 2,180,000	46	Air + pollution + health + carbon + environmental + research + smog + quality + rain + clean 8810	10
HIV 6,360,000	72	AIDS + treatment + HIV + epidemic + health + description + information + service + virus - details 6150	28
Indian tourism 695,000	76	India + tourism + information + indian + pradesh + indiaworld 190	18

Table 5
List of top ten url's retrieved with modified query for alcohol addiction along with their fuzzy memberships

No.	List of top 10 retrieved url's using modified query	Lower similarity	Upper similarity	Bad	Avg.	Good
1	http://dmoz.org/Health/Addictions/Substance_Abuse/Treatment/Alternative/	0.9600	1.0	0.030	0.6049	0.8145
2	http://dmoz.org/Society/Issues/Health/Mental_Health/Substance_Abuse/	0.8570	1.0	0.030	0.6049	0.8145
3	http://uk.dir.yahoo.com/health/diseases_and_conditions/addiction_and_recovery/	0.7850	1.0	0.030	0.6049	0.8145
4	http://directory.google.com/Top/Health/Addictions/Substance_ Abuse/Resources/	0	0.130	0.045	0.500	0.0450
5	http://www.nada.org.au/links.asp	0	0.04	0.5667	0.0626	0.0434
6	http://www.drug-rehabs.com/Top_Tool_Bar/ Resources.htm	0.8210	1.0	0.030	0.6049	0.8145
7	http://www.aizan.net/families/links_alcohol_drug_abuse.htm	0	0.09	0.560	0.1299	0.0439
8	http://www.leydenfamilyservice.org/alcoholdrug.html	0.7850	1.0	0.030	0.6049	0.8145
9	http://www.nethealth.com/links/addiction.htm	0.8210	1.0	0.030	0.6049	0.8145
10	http://mainseek.pl/ca/472620/Health/Addictions/Substance_ Abuse/Resources/	0.0300	1.0	0.030	0.6049	0.8145


Open Directory - Health: Addictions: Substance Abuse: Treatment: Alternative - Shails Explorer
Ffe Elk VEW FavDrtes Tools Help
£|,Sewch
^ritstory
Address http://dmoz.org/Heatilt/Addictons/5ubstan«JIUBe/r.e^tment/AlteriHtrve/
Go
Start
Google Direc...
revisedjourn...
Open Direc...
Links - Shails...
Yahoo! Sear...
home - Shail...
12:53 PM
uzzy Membership value
1
0.8
0.6
0.4
0.2
0
0.03
0.6049
0.8145
Bad Rank
Avg. Rank
Good Rank

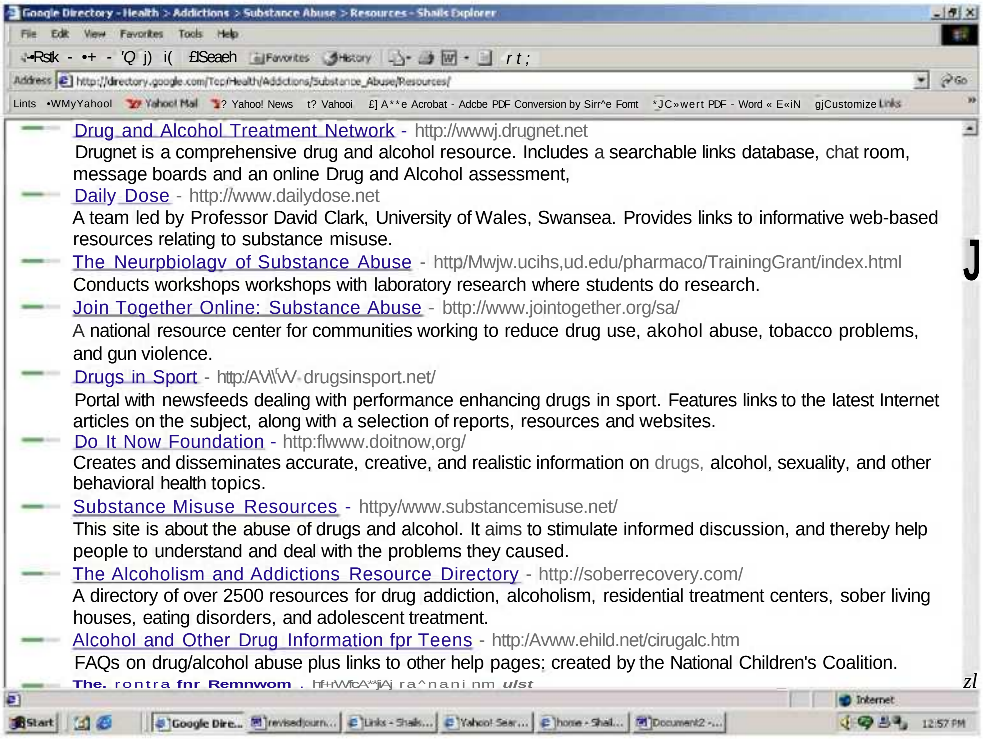
- Addiction 911 - Confidential outpatient treatment for any addictive substance or behavior.
- Addiction and Alternative Medicine - Information and research on complementary and alternative medicine therapies for addictions.
- Alternative Healing - New age alternative healing modality based on ancient Chinese meridians that is helpful for emotional problems, stress, anxiety, anger, few and addictions. Also traditional healing methods.
- Center for Positive Path Recovery - Offers a cognitive/behavioral approach to substance addiction treatment that presents a complementary alternative to the Twelve Steps.
- Chironractv and Dma Addiction • Information on chiropractic techniques are useful in getting people to stay in a drug rehabilitation program.
- CiTA Biomedical - Providers of DNA for Addictions and UROD Rapid detox products for North America.
- Correctional Counseling Inc. - Delves into this systematic approach to treating substance-abusing offenders from a purely cognitive behavioral perspective.
- Cure of Addiction and other Emotional Disorders • Information on the book entitled: Of Frogs and Princes, by Carl Dawntreader, describing a self-directed form of psychotherapy based on the principles of Transactional Analysis.
- Devil's Candy - A site intended to inform and help the victims and abusers of crack cocaine.
- Drink In Safecv.com - Supports doctors and individuals trying to reduce alcohol dqiendence with a combination of anti-craving medications and therapy.
- Efficacy of Buffered Ascorbate Compound - Nutritional supplements that help with detoxification and drug cravings, as well as overall health rehabilitation.
- The Freedom Program - Ultra-rapid detoxification program for those addicted to opiates - heroin, methadone, pethidine, codeine, and other prescription pain medications. Located in WeslvilJe, KwaZulu-Natal, ten minutes from the Durban city center.
- Frontiers! Drug and Alcohol Addiction Recovery Option - Recovery and treatment alternatives are based on the premise that substance abuse and alcoholism are choices not diseases.

Fig. 3. Glimpse of url first of Table 5 and its corresponding fuzzy membership value determined by our system.

Fig. 4 is corresponding to the url http://directory.google.com/Top/Health/Addictions/Substance_Abuse/Resources/. This url is an average document because it has collection of link which provides information mostly about drug abuse. Its fuzzy membership to different categories is also shown below it. The membership to the average category is maximum which corresponds to our judgement.

Fig. 5 is corresponding to the url http://www.nada.org.au/links.asp. This url is a bad document because it provides information on education and training, funding, and government sites links on alcohol addiction. The membership values of this document to various categories also show that it is maximum for the bad category.

To filter out irrelevant documents however, we need to get a single rating for each document. For that, we use the de-fuzzification technique. Table 6 shows the de-fuzzified value for the same


Google Directory - Health > Addictions > Substance Abuse > Resources - Shails Explorer
File Edit View Favorites Tools Help
Rstk Search Favorites History
Address http://directory.google.com/Top/Health/Addictions/Substance_Abuse/Resources/
Go
Lints •WMyYahool Yahoo! Mail ? Yahoo! News t? Vahooi £] A**e Acrobat - Adcbe PDF Conversion by Sirr^e Fomt *JC»wert PDF - Word « E«iN gjCustomize Links
Internet
Start Google Dire... revisedjourn... Links - Shails... Yahoo! Sear... home - Shail... Document2 -...
12:57 PM

Drug and Alcohol Treatment Network - http://wwwj.drugnet.net
Drugnet is a comprehensive drug and alcohol resource. Includes a searchable links database, chat room, message boards and an online Drug and Alcohol assessment,

Daily Dose - http://www.dailydose.net
A team led by Professor David Clark, University of Wales, Swansea. Provides links to informative web-based resources relating to substance misuse.

The Neurpbiolagv of Substance Abuse - http/Mwjw.ucihs,ud.edu/pharmaco/TrainingGrant/index.html
Conducts workshops workshops with laboratory research where students do research.

Join Together Online: Substance Abuse - http://www.jointogether.org/sa/
A national resource center for communities working to reduce drug use, akohol abuse, tobacco problems, and gun violence.

Drugs in Sport - http:/AWVV-drugsinsport.net/
Portal with newsfeeds dealing with performance enhancing drugs in sport. Features links to the latest Internet articles on the subject, along with a selection of reports, resources and websites.

Do It Now Foundation - http:flwww.doitnow,org/
Creates and disseminates accurate, creative, and realistic information on drugs, alcohol, sexuality, and other behavioral health topics.

Substance Misuse Resources - httpy/www.substancemisuse.net/
This site is about the abuse of drugs and alcohol. It aims to stimulate informed discussion, and thereby help people to understand and deal with the problems they caused.

The Alcoholism and Addictions Resource Directory - http://soberrecovery.com/
A directory of over 2500 resources for drug addiction, alcoholism, residential treatment centers, sober living houses, eating disorders, and adolescent treatment.

Alcohol and Other Drug Information fpr Teens - http:/Avww.ehild.net/cirugalc.htm
FAQs on drug/alcohol abuse plus links to other help pages: created by the National Children's Coalition.

The. rontra fnr Remnwom . hf+iWfcA**jiAj ra^nani nm ulst

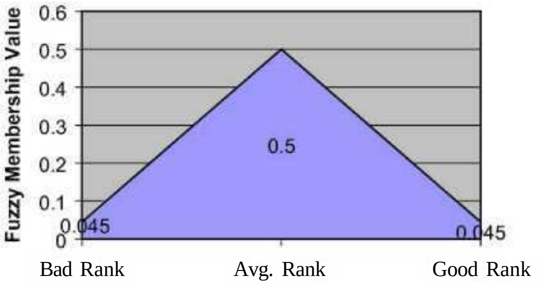
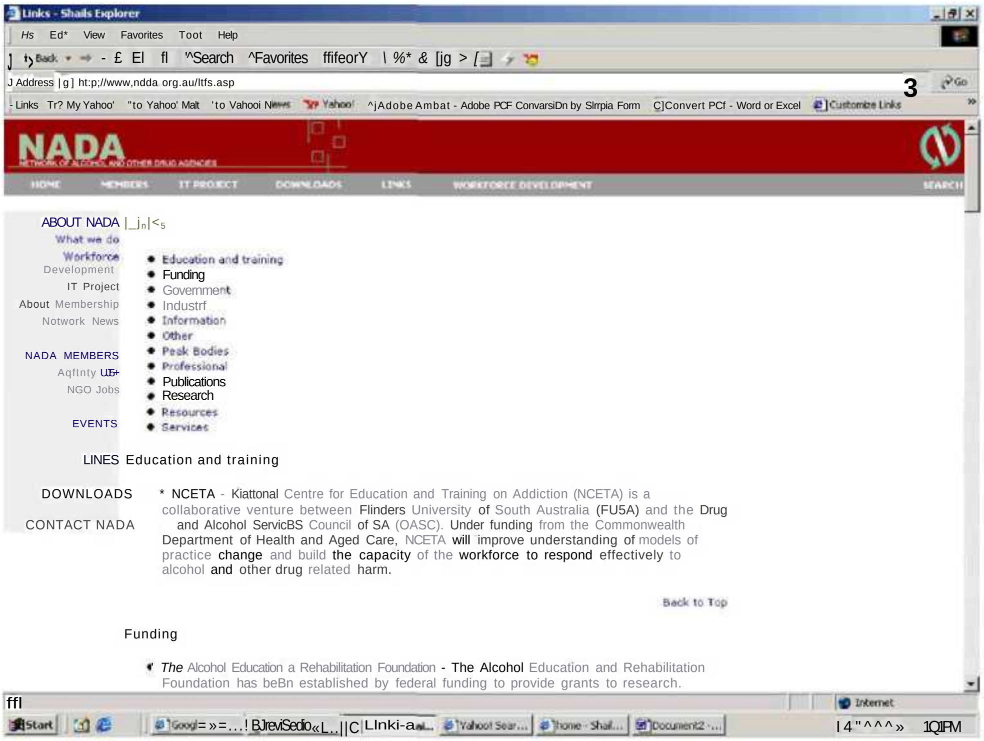


Fig. 4. Glimpse of url fourth of Table 5 and its corresponding fuzzy membership value determined by our system.

documents. These values can be used to rate the documents. Column 4 of Table 6 shows the ratings assigned by the system to the documents using this technique.

In order to judge the accuracy of the grading, we compared the system-generated grades with feedback taken from the user for top 50 documents retrieved with the modified query. For this we requested the users to rate the newly retrieved documents also. Column 5 of Table 6 shows the user assigned rating to the documents. We find that in most of the cases the system grading matches the user grading. However, the system is more harsh towards the average documents and in some cases average documents have been rated bad as Table 6 shows. It may be noted that this step is just for rating the system and is not an integral part of the system.


Links - Shails Explorer
Hs Ed* View Favorites Toot Help
Back ^Search ^Favorites ffifeorY
J Address | g] ht:p;//www,ndda org.au/ltfs.asp
Links Tr? My Yahoo' "to Yahoo' Malt 'to Vahooi News Yahoo! ^jAdobe Ambat - Adobe PCF ConversiDn by Slrrpia Form C]Convert PCf - Word or Excel Customize Links
NADA
NETWORK OF ALCOHOL AND OTHER DRUG AGENCIES
HOME MEMBERS IT PROJECT DOWNLOADS LINKS WORKFORCE DEVELOPMENT SEARCH
ABOUT NADA
What we do
Workforce Development
IT Project
About Membership
Notwork News
NADA MEMBERS
Aqftnty UJ5+
NGO Jobs
EVENTS
LINES
DOWNLOADS
CONTACT NADA
Internet
Start
10:1FM

- Education and training
- Funding
- Government
- Industrf
- Information
- Other
- Peak Bodies
- Professional
- Publications
- Research
- Resources
- Services

Education and training

* NCETA - Kiattonal Centre for Education and Training on Addiction (NCETA) is a collaborative venture between Flinders University of South Australia (FU5A) and the Drug and Alcohol ServicBS Council of SA (OASC). Under funding from the Commonwealth Department of Health and Aged Care, NCETA will improve understanding of models of practice change and build the capacity of the workforce to respond effectively to alcohol and other drug related harm.

Back to Top

Funding

* The Alcohol Education a Rehabilitation Foundation - The Alcohol Education and Rehabilitation Foundation has beBn established by federal funding to provide grants to research.

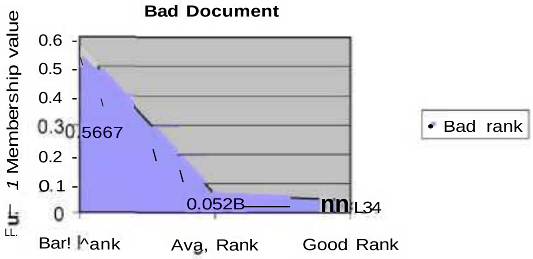


Fig. 5. Glimpse of url fifth of Table 5 and its corresponding fuzzy membership value determined by our system.

Accuracy of system evaluation is defined as

$$\text{Accuracy} = \frac{\text{No. of matches in system rating and user rating}}{\text{Total no. of documents rated by system}} \times 100 \tag{ð11Þ}$$

Table 7 summarizes the accuracy of the system in various domains. The average accuracy of the rating scheme is around 80% for most of the domains. This establishes the effectiveness of the grading scheme. Since bad documents can be identified by the system, these can be eliminated from the list presented to the user. The accuracy was low for the domain alternative medicine since documents were very varied on different forms of alternative medicine like aromatherapy, yoga, acupuncture etc. in this domain.

Table 6
De-fuzzified value and grades assigned by the system to documents of Table 5. The list has been rearranged according to their system ratings. Note that the first document of average class appears at the 23rd position

No.	list of top 10 retrieved url's using modified query	Defuzzified value	System grade	User rated
1	http://dmoz.org/Health/Addictions/Substance_Abuse/Treatment/Alternative/	0.8508	3	3
2	http://dmoz.org/Society/Issues/Health/Mental_Health/Substance_Abuse/	0.8508	3	3
3	http://uk.dir.yahoo.com/health/diseases_and_conditions/addiction_and_recovery/	0.8508	3	3
4	http://www.drug-rehabs.com/Top_Tool_Bar/Resources.htm	0.8508	3	3
5	http://www.leydenfamilyservice.org/alcoholdrug.html	0.8508	3	3
6	http://www.nethealth.com/links/addiction.htm	0.8508	3	3
7	http://mainseek.pl/ca/472620/Health/Addictions/Substance_Abuse/Resources/	0.8508	3	3
…	…	…	…	…
23	http://directory.google.com/Top/Health/Addictions/Substance_Abuse/Resources/	0.3308	2	2
…	…	…	…	…
39	http://www.aizan.net/families/links_alcohol_drug_abuse.htm	0.2264	1	2
…	…	…	…	…
45	http://www.nada.org.au/links.asp	0.1374	1	2

Table 7
Accuracy of system evaluation: comparing system grade vs user rating

Domain	Accuracy (%)
Alcohol addiction	80
Alternative medicine	67.74
Blood cancer	72.3
Air pollution	80
HIV	85
Indian tourism	85

9. Conclusion

In this paper we have presented the design of a complete client-side filtering system for general text documents. The system uses the rough set theoretic concept of discernibility to find words that can discern between good documents and bad ones by analysing a set of training documents rated by the user. This scheme is more powerful than the usual techniques of computing term frequency and inverse document frequency, since it takes into consideration the synonymous words very elegantly. A modified query is built with the discerning words. This query is found to fetch documents, which are more relevant to the user. However, since the documents are still fetched by a traditional search engine, the ordering of the returned documents is still not customized for the user. Hence, we have proposed a rough-fuzzy reasoning scheme which grades the documents. The system first learns the user's basis of rating by relating the grades to rough similarity measures between the training documents and the modified query. These associations are learnt using a decision tree and the classification knowledge is expressed as a set of rules. These rules are used to rate new documents and re-order them, on the basis of rough similarity measures between the new documents and the modified query. To obtain a performance analysis of the system, we requested the users to give their feedback about the retrieved documents also. The document grading scheme is found to work reasonably well.

The rough-set mechanism can be extended to automatic document classification also. The most discerning words for each category can be used as a signature for that category. Rough similarity measures can then be used for categorizing documents automatically. However, since the ranks in that case will not be a graded one, therefore the algorithm for finding words have to be modified. We are also exploring the possibility of building domain specific question answering systems using rough-fuzzy reasoning.

References

Allan, J., Ballesteros, L., Callar, J., & Croft, W. (1995). Recent experiments with INQUERY. In *Proceedings of the fourth text retrieval conference (TREC-4)* (pp. 49-63). NIST Special Publication.

Balabanovic, M. (2000). An adaptive web page recommendation service. In *1st international conference on autonomous agents* (pp. 378-385). ACM Press.

Bodoff, D., Enache, D., Kambil, A., Simon, G., & Yukhimets, A. (2001). A unified maximum likelihood approach to document retrieval. *Journal of the American Society for Information Science and Technology,* 52(10), 785-796.

Bao, Y., Aoyama, S., Du, X., Yamada, K., & Ishii, N. (2001). A rough set based hybrid method to text categorization. In *Second international conference on web information systems engineering (WISE'01), 1,* (pp. 0294).

Crestani, F. (1993). Learning strategies for an adaptive information retrieval system using neural networks. In *Proceedings of the IEEE international conference on neural networks.* San Francisco, CA.

Chakrabarti, S., Dom, B., Gibson, D., Keinberg, J., Raghavan, P., & Rajagopalan, S. (1998). Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceeding of the 7th international world wide web conference.*

Chouchoulas, A., & Shen, Q. (2001). Rough set-aided keyword reduction for text categorisation. *Journal of Applied Artificial Intelligence, 15(9),* 843-873.

Das-Gupta, P. (1988). Rough sets and information retrieval. In *Proceedings of the eleventh annual international ACM SIGIR conference on research and development in information retrieval, set oriented models* (pp. 567-581).

Fuhr, N., & Buckley, C. (1993). Optimizing document indexing and search term weighting based on probabilistic models. In *The first retrieval conference (TREC-1)* (pp. 89-99). NIST Special Publication.

Fuzzy Logic Toolbox. The MathWoks, Incorporation. Available: http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy/fuzzy. shtml?BB = 1.

Google Search Engine Optimization. Available: http://www.internet-advertising-marketing-manual.com/google-optimization.htm. Intarka, Inc. (1999).

Intarka, Inc. (1999). Intarka announces ProspectMiner 1.2—a powerful web mining solution for business. Sun Microsystems, Inc. Available: http://industry.java.sun.com/javanews/stories/story2/0,1072,18628,00.html.

Jochem, H., Ralph, B., & Frank, W. (1999). WebPlan: dynamic planning for domain specific search in the internet. Available: http://wwwagr.informatik.uni-kl.de/~webplan/PAPER/Paper.html.

Jensen, R., & Shen, Q. (2001). A rough set-aided system for sorting WWW bookmarks. In *Proceedings of the 1st Asia-Pacific conference* (pp. 95-105). Web Intelligence.

Korfhage, R. R. (1997). *Information storage and retrieval.* Wiley Computer Publishing (pp. 221-232).

Komorowski, J., Polkowski, L., & Andrzej, S. (1999). Rough sets: a tutorial. *The 11th European summer school in logic, language and information,* Utrecht University, Netherlands. Available: http://www.let.uu.nl/esslli/Courses/skowron/skowron.ps.

Mitchel, T. (1997). *Machine learning.* McGraw Hill.

Menasalvas, E., Millan, S., & Hochsztain, E. (2002). A granular approach for analyzing the degree of afability of a website. In *International conference on rough sets and current trends in computing, (RSCTC2002).*

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11(5),* 341-356.

Pazzani, M., Muramatsu, J., & Billsus, D. (1996). Syskill and webert: identifying interesting web sites. In *Proceedings of the national conference on artificial intelligence (AAAI-96)* (pp. 54-61). Portland, OR.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: experiments in automatic document processing* (pp. 313-323). Englewood Cliffs, NJ: Prentice Hall.

Roldano, C. (1999). Userprofiling with bayesian belief networks. Available: http://www.labs.bt.com/profsoc/facts/workshop/abstract/BBN.html.

Salton, G. (1971). *The SMART retrieval system: relevance feedback and the optimization of retrieval effectiveness.* Prentice-Hall.

Salton, G., Fox, E. A., & Voorhees (1985). Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science, 36(3),* 200-210.

Stefanowski, J., & Tsoukias, A. (2001). Incomplete information tables and rough classification. *Computational Intelligence, 17,* 454^66.

Srinivasan, P., Ruiz, M. E., Kraft, D. H., & Chen, J. (2001). Vocabulary mining for information retrieval: Rough sets and fuzzy sets. *Information Processing and Management, 37,* 15-38.

Szczepaniak, P. S., & Gil, M. (2003). Practical evaluation of textual fuzzy similarity as a tool for information retrieval. In *Advances in web intelligence* (pp. 250-257). LNAI 2663.

WordNet. Available: http://www.cogsci.princeton.edu/~wn/.