Locality Mutual Clustering for Document Retrieval

Khu Phi Nguyen
University of Information Technology
Vietnam National University - HCMC
Q.6, Linh Trung, Thu Duc, HCMC, Vietnam
+84 8 372 52 002 +84 903 942 461
khunp@uit.edu.vn

Hong Tuyet Tu
University of Technical Education – HCMC
Faculty of Information Technology
1. Vo Van Ngan, Thu Duc, HCMC, Vietnam
+ 84 8 389 68 641 +84 908 379 610
hongttsp@gmail.com

ABSTRACT

Document retrieval is aimed at searching relevant documents in response to answer user query. To do this task, algorithms of document clustering play an important role. These algorithms are often based on frequency computation of key-phrases in both query and document, and focus on locality. It is dealt with this paper an algorithm based on locality mutual clustering is proposed to cluster documents and to find relevance documents in answer to user query. This proposed algorithm has been used for searching scientific papers in our institutions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering, Query formulation, Retrieval models, Search process.*

General Terms

Algorithms, Management, Design, Documentation.

Keywords

Document retrieval, mutual information, probability distribution, weighted graph.

1. INTRODUCTION

Document retrieval considered as a branch of information retrieval is a problem of matching of some user query against a set of text records such as articles, topics, etc. User queries can range from text descriptions to a few words such as keywords, key-phrases. A document retrieval system consists of a database of text-indexed documents and a user interface tool to access the database. Thence, the main tasks of the system are to find relevant documents to user query, to evaluate the matching results and to arrange them according to relevance.

In doing these tasks, the problem of document clustering plays an important role in partition a set of documents into a number of clusters, such that the documents in one cluster share the same topic. In other words, the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. After clustering, the problem of searching related documents to answer a user query is reduced to the one in locality, e.g. method of mutual subspace clustering [1], or local topology preserving indexing [2].

It is dealt with this paper a proposed algorithm, named locality mutual clustering, is aimed at clustering documents and response to user query based on graph representation and mutual information framework. Firstly, each document is characterized by topics that are keywords or key-phrases. Set of discriminative topics of a document space is represented by vertices or nodes in a graph. Hence, a vertex is labeled by its topic and a frequency of the topic in this topic in the document space. A couple of vertices may be connected together by an edge if there exists at least a document that contains both topics in these vertices. In such a case, weight of each edge is defined and equals to the mutual information between two its corresponding vertices.

Using the proposed algorithm, a user-queried content is partitioned into parts in conformity with topics. Each of the parts is located in a locality or a vertex with its nearest neighbor vertices. Based on mutual information between a vertex to its nearest neighbors and doing recursively the algorithm relevant documents are searched to answer user query.

2. RELATED WORKS

Nowadays electronic libraries deeply grow in size, querying documents and their contents will become unwieldy. Systems of searching assistance can provide hundreds and even thousands of matched documents. But users will only consider a handful of appropriate result documents with due attention. Many of matches that may best satisfy user query be buried in remainder documents that are never viewed. In such a case, the problem is how to do for making good these shortcomings, [3].

Document clustering is recently used as an important tool for document search engines and may be one remedy for information retrieval in this situation. In the context of information retrieval systems [4], documents have been clustered by the distributions of either keywords that co-appear in the documents or documents in which keywords appear. Michael Steinbach et al provided a comparison of document clustering techniques, [5]. And, it was recognized that the method of agglomerative hierarchical clustering and bisecting k-means are two main approaches to document clustering. Among these, the latter is better than the standard k-means approach and as good as, or better than the hierarchical approaches.

From viewpoint of grouping data, a global criterion function is used to optimize different aspects of intra-cluster similarity and inter-cluster dissimilarity so that similarities of elements in the

IMCOM (ICUIMC)'14, January 9–11, 2014, Siem Reap, Cambodia.
Copyright 2014 ACM 978-1-4503-2644-5 …$15.00.

intra-group are high and in the inter-group are low. A popular similarity measure is the cosine function to assign each document into a cluster with an effort to maximize the intra-cluster similarity, [5, 6]. Recently, Noam and Naftali [5] proposed a clustering method based on information theoretic formulation in which criterion function is determined by mutual information and members in the same cluster must have a maximum factor of mutual information. In the same aim, Patrick and Dekang put forward a clustering algorithm, called clustering by committee, to produce higher quality clusters in document clustering tasks, [7].

In general, the document space is of high dimensionality and high volume of data, these make clustering in such a space is often more difficulty, even unfeasible. The authors in [8] proposed a novel document clustering algorithm which aims to cluster the documents into semantic classes by locality preserving indexing. The proposed algorithm projects the document space into a lower dimensional semantic space and gives the intuitive motivation due to unsupervised approximations. In addition, this algorithm also used a p-nearest neighbor graph to discover the local manifold structure and then the global one. Experiments on data corpora, namely Reuters or Topic Detection and Tracking, showed that this proposed algorithm performed more applicable and much better than traditional clustering algorithm when the data set is large.

A document may contain multiple topics, a large set of independent keywords or key-phrases, sequences and a few core-words called features. Document clustering algorithms are usually based on topic features and on counting frequency of features to perform cluster. These approaches cluster documents independent of their context and do not cater for the meaning behind feature texts whereas feature space can be very challenging for document clustering. So, these algorithms are sometimes less efficient in producing results with high cluster quality. From this point, a new approach for document clustering based on the topic map representation of the documents was proposed by using the inferred information through topic maps, [9]. The comparative experiments on standard document datasets, e.g. Reuters-21578, NEWS20 available at [10-12] respectively, reveal that this approach is effective in improving the cluster quality.

Beside existing document clustering techniques using similarity criteria, semantic classes, feature of synonymy between related documents are also interested in increasing the effectiveness of document clustering. In clustering documents, citation contexts are used to provide relevant synonymous and related vocabulary to improve text-based clustering techniques. Thence, by using link-based method to determine the similarity following number of co-citations some clustering algorithms become more available. Experiments in [13] shown that the use of citation contexts is an effective and feasible means of clustering scientific documents. A detailed survey of the problem of text clustering is also presented in [14] with the challenges of the clustering problem in social networks and linked data. The most significance is large amounts of text data being created by dynamic applications and text applications increasingly arise in heterogeneous applications or multimedia data.

In document clustering, user supervision can also be provided in forms of labeling features. Various types of semi-supervised clustering algorithms were explored with feature supervision [15], and the instance-level supervision with feature-level supervision can improve significantly document clustering performance. Experimental results demonstrate that all types of semi-supervised clustering algorithms with feature supervision improved clustering performance, but the evaluation of the effectiveness through user studies is still in progress.

In 2012, Christos and Vassilis proposed an improved clustering using the external knowledge from WordNet hypernyms, [16]. Study on a corpus of news articles from news portals shown that this proposed algorithm improves possibly standard k-means, especially in enriching the clustering process by utilizing hypernyms and generating useful labels. Up to now, many clustering algorithms have been proposed. However, most of them suffer from challenges in dealing with problems such as high dimensionality, scalability, accuracy, and meaningful cluster labels. For instance in [17], many related studies of the hierarchical document clustering algorithm are illustrated and shown that this method performs well but still having a scope to improve.

3. GRAPHIC REPRESENTATION

Let D be a set of documents and assume that a document is specified by a group of distinguished topics, keywords or key-phrases and T is the set of these ones.

Let G = (V, E) be a graph representing T and relations between topics in D. Each vertex v∈V is assigned by a unique natural number i with respect to a topic t occurring with a frequency f in D. These fields of v can be specified using v.i, v.t and v.f. An edge e = (s, h)∈E, starting from s∈V and reaching to h∈V, is taken shape if there exists at least a document in D containing both topics s.t and h.t, respectively. Frequency f of the couple of topics s.t and h.t in D is assigned for e, denoted by e.f .

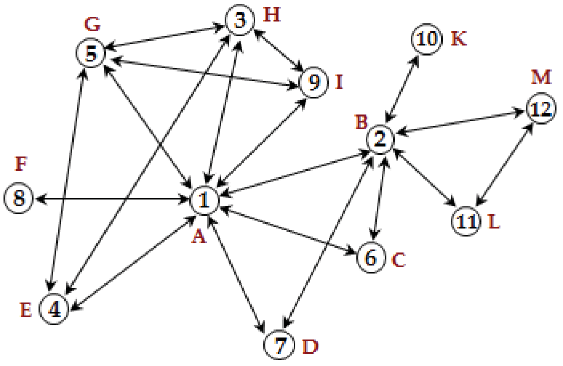


Figure 1. Representation of document space

Let X and Y are random variables associated with vertices u,v∈V. Probabilities P(X) and P(Y) specifying how frequent topics u.t and v.t occur in D can be estimated by u.f and v.f. The joint distribution P(X, Y) with respect to e = (u,v) is estimated by e.f, a frequency to occur simultaneously of u.t and v.t in D. Hence, mutual information between X and Y is determined by:

$$I(X;Y)=\sum_{x,y} P(X,Y)\log_2\left(\frac{P(X,Y)}{P(X)\times P(Y)}\right) \qquad (1)$$

$$\approx \sum_{u,v\in V,\ e=(u,v)} e.f\times\log_2\left(\frac{e.f}{u.f\times v.f}\right) \qquad (2)$$

Therefore, the mutual weight of an edge e = (u, v) denoted by e.w can be defined as follows:

$$e.w = e.f \times \log_2\left(\frac{e.f}{u.f \times v.f}\right) \qquad (3)$$

The following example is to demonstrate the above definitions. A document space D and set of topics T are shown in Table 1. Frequencies of vertices, edges, and mutual weights are listed in Table 2. Figure 1 illustrates a graphic representation of these establishments.

Table 1. Document space and set of topics

D	Topics
1	B, A, C
2	B, D, A
3	A, B
4	B, K
5	B, L, M
6	E, A, G, H
7	A, E, H
8	H,G, A, I
9	A, F

Where:

|D| = 9

|V| = |T| = 12

i	t	f
1	A	7
2	B	5
3	H	3
4	E	2
5	G	2
6	C	1
7	D	1
8	F	1
9	I	1
10	K	1
11	L	1
12	M	1

Table 2. Frequencies and mutual weights

u.i	v.i	u.t	v.t	u.f	v.f	e.f	e.w
1	3	A	H	7	3	3	0,120857
1	4	A	E	7	2	2	0,080571
1	5	A	G	7	2	2	0,080571
1	6	A	C	7	1	1	0,040286
1	7	A	D	7	1	1	0,040286
1	8	A	F	7	1	1	0,040286
1	9	A	I	7	1	1	0,040286
1	2	A	B	7	5	3	-0,124799
2	6	B	C	5	1	1	0,094222
2	7	B	D	5	1	1	0,094222
2	10	B	K	5	1	1	0,094222
2	11	B	L	5	1	1	0,094222
2	12	B	M	5	1	1	0,094222
3	4	H	E	3	2	2	0,352214
3	5	H	G	3	2	2	0,352214
3	9	H	I	3	1	1	0,176107
4	5	E	G	2	2	1	0,129992
5	9	G	I	2	1	1	0,241103

4. LOCALITY MUTUAL CLUSTERING

To query about documents containing simultaneously of topics or key-phrases s_0, s_2, s_3,... , s_{n-1} as input query, a locality mutual clustering algorithm or LMCA is designed recursively with the following steps:

- Preprocessing input query to split the query into separated key-phrases, to exclude key-phrases not included in T, and to arrange the legitimate key-phrases to an array in descending of their frequencies.
- Taking the first element of the array to find a cluster of key-phrases related to this element, and sorting the found these related key-phrases in decreasing order of their mutual weights with it.
- Processing the first element to find a group of key-phrases related to it, and then clustering them to make a part of solution.
- Calling recursively the algorithm to process the remainder part of the key-phrase array and to find additional part of solution.
- The recursive call in the algorithm is terminated if there is no element to be processed.

Pseudo-code of the LMCA is proposed as follows:

```
Algorithm LMCA(G,Q)
Input: G = (V,E) // Graph of document space
          Q = $s_0$,$s_2$,$s_3$,… ,$s_{n-1}$ // Query
Output: Ans // String to answer query
Begin
// Preprocess input query
string t[n];
for i := 0 to n-1 do t[i] := $s_i$ endfor
for i := 0 to n-1 do
 if ( notFound v∈V | v.t = t[i] ) then
  for j := i to n-2 do t[j] := t[j+1] endfor
  n := n-1; i := i-1;
 endif
endfor
for i := 0 to n-2 do
 for j := i+1 to n-1 do
  find v∈V | v.t = t[i] ;
  find v'∈V | v'.t = t[j] ;
  if ( v.f < v'.f ) then swap(t[i], t[j]) ;
 endfor
endfor
// Sort topics and select the 1st element
string Ans, q1, q2 ;
Ans := t[0]; q1 := q2 := "" ;
find u∈V | u.t = t[0] ;
find V(u) := { v∈V | ∃e∈E e.s = u and e.h = v} ;
find E(u) := { e∈E | e.s = u, e.h = w, w∈V(u)∪{u} };
sort E(u) by descending order of e.w;
// Store remainder topics for the next step
for i := 1 to n-1 do
 if ( notFound v∈V(u) | v.t = t[i] ) then
  q2 := concat(q2, ",", t[i] ); // stored array
  for j := i to n-2 do t[j] := t[j+1] endfor
  n := n-1; i := i-1;
 endif
endfor
```

```
// Find clusters of the selected element
Let V(u-) := V(u); E(u-) := E(u) ;
for i := 1 to n-1 do
 if ( notFound $v \in V(u)$ | v.t = t[i] ) then
  if (q1= "") then q1 := concat(t[0], ",", t[i])
  else q1 := concat(q1, ",", t[i] )
  endif
 else
  find $v \in V(u)$ | v.t = t[i] ;
  for $e \in E(u-)$ do
   if ( v = e.h ) then
    Ans := concat(Ans, ",", t[i]) ;
    $V(v) := \{ w \in V \mid \exists e \in E \; e.s = v$ and $e.h = w \}$;
    $V(u-) := V(u-) \cap V(v)$ ;
    $E(u-) := \{ e \in E \mid e.s = v, \; e.h = w, \; w \in V(u-) \cup \{u\} \}$;
    sort E(u-) by descending order of e.w ;
    for j := i to n-2 do t[j] := t[j+1] endfor
    n := n-1; i := i-1 ;
    break;
   endif
  endfor
 endif
endfor

// Clustering related topics
initiate $C := \{ \{v\} \mid v \in V(u-) \}$;
for $v \in V(u-)$ do
 find $L(v) := \{ w \in V(u-) \cup \{u\} \mid \exists e \in E \; e.s = v, e.h = w \}$;
endfor
for $v \in V(u)$ do
 for $w \in V(u)$ do
  find cluster $c \in C \mid v \in c$ ;
  find cluster $c' \in C \mid w \in c'$ ;
  if $w \notin c$ and $w \in L(v)$ and $v \in L(w)$ then
   $c := c \cup c'$ ;
   $C := C \setminus c'$ ;
  endif
 endfor
endfor

// Recursive calls to answer
if ( $q1 \neq$ "" ) LMCA(G, q1) ;
if ( $q2 \neq$ "" ) LMCA(G, q2) ;
end.
```

5. CASE STUDY

In case of Table 1, to a query "A" the LMCA firstly outputs a set of topics conjoined with A listed in decreasing order of their mutual weights at the first column of Table 3.a. Next to each of these topics in the first column is a set of its related topics arranged in the same descending order of mutual weights with it. Finally, all of the searched topics are clustered to obtain the combinations [H, E, G, I], [C, D, B], [F], in which a notation [x, y, z] is to show a collection of x, y, z. Thus, in answer to the query, the LMCA returns a set of key-phrases in the form: "A and ([H, E, G, I] or [C, D, B] or [F])". It is verified that such an answer corresponds to the documents numbered 1, 2, 3; 6, 7, 8, 9 in Table 1.

Similarly, if a query is to find documents contained a key-phrase like "A and H and G" in D of Table 1, the LMCA outputs conjoined topics [E], [I] and their related topics in Table 3.b. Hence, the answer is of the form: "A and H and G and (E or I)". This means that there exists such a key-phrase in D and this key-phrase also includes other topics named E or I. The answer to the query is that there are documents numbered 6, 8 in Table 1 with the key-phrases "A and H and G and E" and "A and H and G and I" satisfying the query.

Table 3. Results of the query

(a) about "A"

V(u)	L(v)
H	E, G, I, A
E	H, G, A
G	H, I, E, A
C	B, A
D	B, A
F	A
I	G, H, A
B	C, D, A

Ans	Combinations
A	H, E, G, I
	C, D, B
	F

(b). "A and H and G"

V(u)	L(v)
E	H, G, A
I	G, H, A

Ans	Combinations
A,H,G	E
	I

b. In a document database of tens of thousands of e-papers at the library of our Dept. a computer program containing the LMCA was setup. Users can search their papers using queries. For instance, if user need papers that contain "granular computing, knowledge discovery, information retrieval" in all, this phrase is input as a query to the program. Then LMCA responds to the query with an answer, like that "granular computing and knowledge discovery", "granular computing and information retrieval" arranged in descending order of their mutual weights with the query. In the second column of Table 4, clusters of other key-phrases co-appeared in the corresponding papers are also included. For example, the following key-phrases: "rough sets", "data mining", "rough set theory", "classification", "concept formation", "data analysis" are possibly co-appeared with the "granular computing and knowledge discovery" key-phrase in the searched documents.

Table 4. Results of the query "granular computing, knowledge discovery, information retrieval"

Ans	Combinations
granular computing, knowledge discovery	rough sets, data mining, rough set theory, classification, concept formation, data analysis
granular computing, information retrieval	information granulation, information retrieval support systems

In case of a query to find papers that contain "granular computing, knowledge discovery, information retrieval, ontology, rough sets, data mining". The LMCA responds that such a query in all is not found in the document database. But, two key-phrases named "granular computing and knowledge discovery" and "granular computing and information retrieval" have appeared. And also, there are other key-phrases co-appeared in the documents that contain these two searched key-phrases in the document database. For instance, in Table 5, it is obtained documents with the key-phrase "rough sets, granular computing, data mining, knowledge discovery" and these documents also contain "classification" key-phrase, etc.

Table 5. Results of "granular computing, knowledge discovery, information retrieval, ontology, rough sets, data mining"

Ans	Combinations
rough sets, granular computing, data mining, knowledge discovery	Classification
granular computing, information retrieval	information granulation, information retrieval support systems
granular computing, ontology	rough set theory, description logic, ontology editor, granular information

This way of the LMCA answer creates favorable conditions for users in concretizing precisely their queries and re-querying to meet their documents more accordingly. Documents that contain user-specified key-phrase, e.g. "granular computing, knowledge discovery, information retrieval, ontology, rough sets, data mining" in Table 5, is entered in the edit-box of the LMC computer program shown in Figure 2. After clicking on the next button, corresponding results are illustrated in the Found Documents windows of the LMC program. Users can view, print, save their found documents by choosing the Download buttons.

6. CONCLUSIONS

In this paper, a new algorithm named LMCA for document clustering based on graphic representation of mutual information between topics of document space is proposed. The LMCA performs clustering in the nearest neighbor locality of user-queried topic components that have highest mutual information between topics in the neighbor, and using the recursive technique to integrate gradually results, then output answer to input query. The experiments for document retrieval on the real data set show that the proposed LMCA outperforms others and users meet favorable conditions to concretize their queries. Problems of determining document characteristics and analyzing complexity when working with large-scale data are still open and would be future works.

7. ACKNOWLEDGMENT

This work is funded by Vietnam National University VNU-HCM under grant number C2013-26-03.

8. REFERENCES

[1] Ming Hua, Jian Pei. 2012. Clustering in applications with multiple data sources-A mutual subspace clustering approach. *Elservier B.V, Neurocomputing* 92, (Feb. 2012), 133-144. doi:10.1016/j.neucom.2011.08.032.

[2] Jieqing Xing, Chuanyi Fu.2013. LTPI: A Spectral Clustering Method Based on Local Topology Preserving Indexing and its Application for Document Clustering. *Intl' Jour. of Hybrid Information Technology*, Vol. 6, No. 1, Jan. 2013.

[3] C.R. Palmer, J. Pesenti, R.E Valdes-Perez, M.G. Christel, A.G. Hauptmann, D. Ng.H.D. 2001. Wactlar. Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results. In *E.Fox and C.L. Borgman (Eds.), Proc. of the First ACM/IEEE JCDL'01,* (June 24-28, 2001), New York, ACM 1-58113-345-6/01/0006.

[4] Noam Slonim, Naftali Tishby. 2000. Document Clustering using Word Clusters via the Information Bottleneck Method. In *ACM SIGIR*, 208–215. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3062 .

[5] M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. *Proc. of Text Mining Workshop*, KDD 2000.

[6] Congnan Luo, Yanjun Li, Soon M. Chung. 2009. Text Document Clustering based on neighbors. *Elsevier B.V., Data & Knowledge Engineering* 68, (Jun, 2009), 1271–1288. doi:10.1016/j.datak.2009.06.007.

[7] Patrick Pantel and Dekang Lin. 2005. Document Clustering with Committees. In *Proc. of SIGIR'02*, (August, 2002), 11-15, ACM 1-58113-561-0/02/0008, Tampere, Finland.

[8] Deng Cai, Xiaofei He, Jiawei Han. 2005. Document Clustering Using Locality Preserving Indexing. *IEEE Trans. Knowledge and Data Eng.*,Vol. 17, No. 12, (Dec. 2005), 1624-1637.

[9] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq. 2010. Document Clustering based on Topic Maps. *Intl' Jour. Of Computer Applications* (0975-8887), (December, 2010), Vol.12– No.1.

[10] Reuter's newswire. http://www.daviddlewis.com/resources/testcollections/reuters21578/

[11] Data set among text mining community. http://people.csail.mit.edu/jrennie/20Newsgroups/

[12] Medical literature database of National Library of Medicine. http://davis.wpi.edu/xmdv/datasets/ohsumed.html

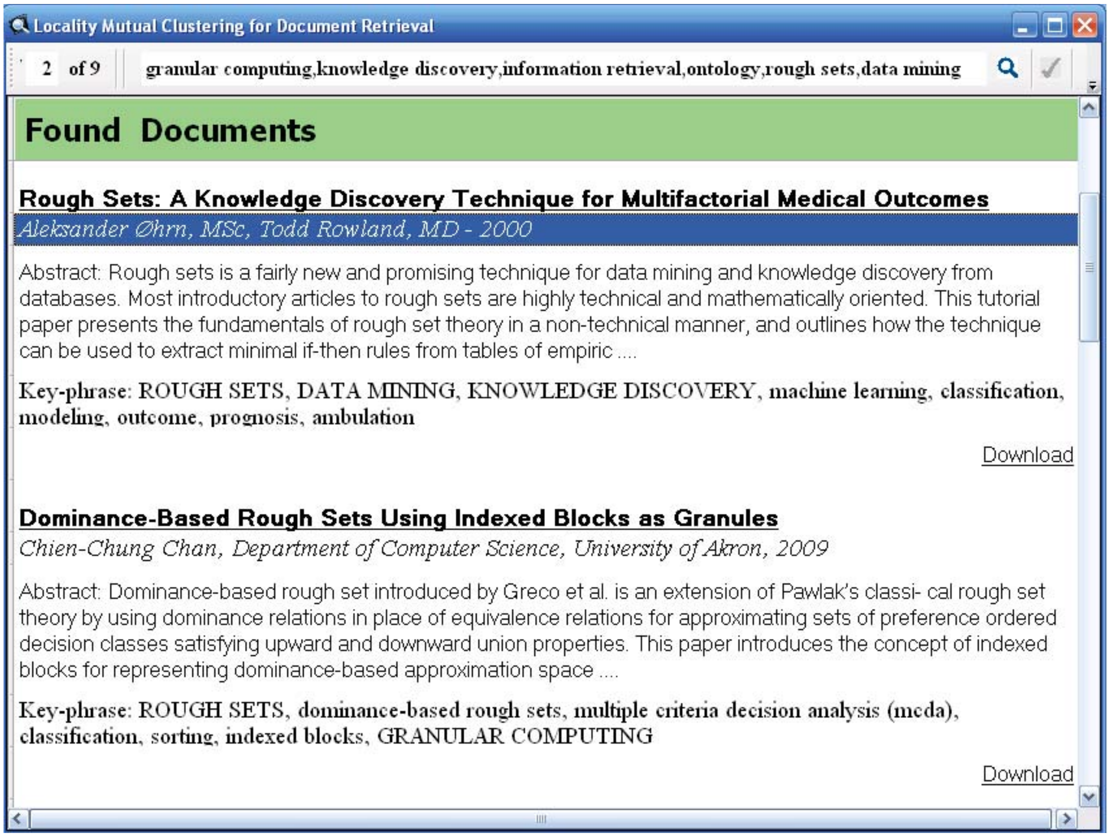
[13] Bader Aljaber, Nicola Stokes, James Bailey, Jian Pei. 2009. Document clustering of scientific texts using citation contexts. *Springer Science and Business Media*, LLC 2009.

[14] Charu C. Aggarwal, ChengXiang Zhai. 2012. A Survey of Text Clustering Algorithms. *Mining Text Data*, Springer, 77–128.

[15] Yeming Hu, Evangelos E. Milios, James Blustein. 2012. Enhancing Semi-Supervised Document Clustering with Feature Supervision. *SAC'12*, March 25-29, 2012, © 2011 ACM 978-1-4503-0857-1/12/03. Italy.

[16] Christos Bouras , Vassilis Tsogkas. 2012. A clustering technique for news articles using WordNet. *Elsevier B.V., Knowledge-Based Systems*, (Jun 2012), http://dx.doi.org/10.1016/j.knosys.2012.06.015 .

[17] Deshmukh D.B. and Pandey Y. 2012. A Review on Hierarchical Document Clustering. *Journal of Data Mining and Knowledge Discovery* ISSN: 2229–6662 & ISSN: 2229–6670, Vol.3, Issue 2, 65-68. Available online at http://www.bioinfo.in/contents.php?id=42.



Locality Mutual Clustering for Document Retrieval
2 of 9
granular computing,knowledge discovery,information retrieval,ontology,rough sets,data mining

Found Documents

Rough Sets: A Knowledge Discovery Technique for Multifactorial Medical Outcomes

Aleksander Øhrn, MSc, Todd Rowland, MD - 2000

Abstract: Rough sets is a fairly new and promising technique for data mining and knowledge discovery from databases. Most introductory articles to rough sets are highly technical and mathematically oriented. This tutorial paper presents the fundamentals of rough set theory in a non-technical manner, and outlines how the technique can be used to extract minimal if-then rules from tables of empiric ....

Key-phrase: ROUGH SETS, DATA MINING, KNOWLEDGE DISCOVERY, machine learning, classification, modeling, outcome, prognosis, ambulation

Download

Dominance-Based Rough Sets Using Indexed Blocks as Granules

Chien-Chung Chan, Department of Computer Science, University of Akron, 2009

Abstract: Dominance-based rough set introduced by Greco et al. is an extension of Pawlak's classi- cal rough set theory by using dominance relations in place of equivalence relations for approximating sets of preference ordered decision classes satisfying upward and downward union properties. This paper introduces the concept of indexed blocks for representing dominance-based approximation space ....

Key-phrase: ROUGH SETS, dominance-based rough sets, multiple criteria decision analysis (mcda), classification, sorting, indexed blocks, GRANULAR COMPUTING

Download

Figure 2. Found documents with respect to user-entered key-phrase in the LMC computer program