Egyptian Informatics Journal 19 (2018) 129–132

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

Feature selection for document classification based on topology

O.G. El Barbary *, A.S. Salama

Mathematics Department, Faculty of Science, Tanta University, Egypt

ARTICLE INFO

Article history:
Received 11 June 2017
Revised 3 November 2017
Accepted 5 January 2018
Available online 10 February 2018

Keywords:
Information retrieval system
Document classification
Topological space
Feature selection
Near open sets
Rough sets

ABSTRACT

Feature selection is the method of how to select the best subset of the document occurring in data core for using it in purposes of data mining or applications. In this paper, we introduced a new technique using topological spaces for developing Information Retrieval System (IRS). First, we introduced the definition of topological information retrieval systems (TIRS) as a generalization of the information retrieval system. Second, we applied some topological near open sets to these systems for feature selection. Indiscernibility of keywords in these systems are discussed and their applications are given. We suggested and examined the order relation that representing the relationships among documents of the document space.

1. Introduction

The beginning of the global network has greater than before, through the evolution of information retrieval technique. As a good replacement for going to the local library to look for information, people can search on the Web. Therefore, the virtual number of manual versus computer-assisted searches for information has shifted radically in the past few years. In addition, the automated information retrieval for many document collections helped in reading, understanding, indexing and tracking a large amount of data. For this cause, researchers in fields of document retrieval, computational linguistics, and textual data mining are working hard on development new methods to process these data [1–6].

This representation suffers from two major challenges the problem of feature selection, and the problem of high dimensionality. In the bag-of-words model, every word in the document can be selected as a feature, and the dimension of the feature space is equivalent to the number of different words in all of the documents. The occurrences of the word in the document, in the group, and in the whole gathering is very important for information retrieval process.

There are several methods for term selection, in this paper; we will present new method for term selection using topology.

The concepts of topological space (near open sets) are one of the recently mainly powerful tools of data analysis. Many researchers have appeared lately, they are applied the near open sets in their fields, for instance information analysis in chemistry and in physics. The principle of the present work is to put a starting point using topological structures for the applications in information retrieval. Rough set theory is a mathematical tool introduced by Pawlak in 1982 [7], it supports the uncertainty reasoning although qualitatively. The basic concepts and relations of this theory have been studied in [8,9].

Topology is the rich field of mathematics that exist in nearly all branches of mathematics; in addition, it is used in many real life applications. We consider that the topological near open sets are the central base for knowledge extraction from incomplete information tables and in data processing [10–15].

In this paper, we proposed the topological information retrieval system based on the notion of some topological near open sets. The knowledge used in these systems consist of an information retrieval system. In this system, each document is represented by its values on a finite set of keywords. We defined the topological bases on the set of keywords of this system. With the topological information retrieval system, we can able to perform approximate retrieval. We introduced the basic mathematical operations on topological systems based on general topology. We suggested

* Corresponding author.
E-mail addresses: ualbarbari@su.edu.sa (O.G. El Barbary), dr_salama75@yahoo.com (A.S. Salama).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2018.01.001

and examined the order relation that representing the relationships among documents of the document space. The approximate retrieval is carried out by the reduction of the unique query. This is finished using topological methods, such as topological near open sets and their generalizations.

2. Feature selection

Feature selection is the process that leads to the reduction of dimensionality of the original data set. The selection term set should contain enough or more reliable information about the original data set. To this end, many criteria are used [16–18]. For apply the feature selection there are two ways to select it. The first is forward selection starts with no terms and adds them one by one, at each one adding the one that reductions the mistakes. The second is the backward selection that starts with all the terms and eliminates them one by one. Hence, eliminate the one that reductions the most error, in hopes no further elimination up to the error.

3. Feature selection methods

Many of feature selection methods contain relied greatly on the analysis of the character of a particular data set through statistical or information-theoretical procedures. For text learning tasks, there are mainly calculation on the vocabulary-specific characteristics of given textual data set to spot excellent term features. Even though the statistics itself do not care concerning the meaning of the text, but these methods are useful for text learning tasks [19].

Many feature selection methods described a statistical feature selection algorithm call RELIEF that uses instance base learning to hand over a relevance weight to each feature [20].

Furthermore, that feature selection should depend not only on the features and the goal concept, but also on the induction algorithm.

4. Basic concepts of topology and rough sets

A family τ of subset U is a topological space be topological space if it satisfying the following conditions:

1. $\varphi, U \in \tau$.
2. τ is closed under uninformed union.
3. τ is closed under limited intersection.

The subsets of U belong to τ are called the open sets.

It often happens that the open sets of a topological space can be complicated and yet they can have described using a selection of simple special ones. In addition, it is chance that many topological concepts can be characterized in terms of these simpler bases or subbase elements. Officially, $\beta \subseteq \tau$ is a base for (U, τ) if the non-empty open subbase of U represent a union of a subfamily of β. A family $\delta \subseteq \tau$ is a subbase if all finite intersections construct a base.

$cl(X) = \cap\{Y \subseteq U : X \subseteq Y, U - Y \in \tau\}$ and $\mathrm{int}(X) = \cup\{Y \subseteq U : Y \subseteq X, Y \in \tau\}$ are closure and interior of $X \subseteq U$, respectively.

The approximation space is a pair $A = (U, R)$, where R is called equivalence relation or indiscernibility relation. Furthermore, $[x]_R, x \in U$ is the equivalence class containing the element x.

5. Topological information retrieval systems

We define the information retrieval system as follows:

$IRS = (DS, KW, \{C_s : s \in KW\}, \{f_s : s \in KW\})$, where DS is the universe of documents. KW is the set of attribute where C_s is the set of attribute values. Finally, f_s is the information function of the system.

In multi information retrieval system (MIRS), each attribute $s \in KW$ defines a relation $R_s \subseteq DS \times C_s$ by $(d, c) \in R_s \Longleftrightarrow c \in f_s(x)$. By this way, each element of the document space can be description by means of a subset $SB \in KW$ of keywords, called the $SB-$ description and denoted by $SB(d)$. The $SB-$ description $SB(d)$ is defined as follows:

$SB(d) = \prod_{s \in SB} f_s(d)$, such that $SB \in 2^{C_s}$.

The extended information retrieval system to topological information retrieval system (TIRS) is done by familiarizing a general relation $R \subseteq C_s \times C_s$ on the range C_s.

By important a general relation on the set C_s we can make topological constructions on the keyword values. For every $R_s(C), C \subseteq C_s, s \in SB$, we define the topology τ_s which has $\{R_s(C), C \subseteq C_s\}$ as a sub-base. $\mathrm{TMIRS} = (\mathrm{DS}, \mathrm{SB}, \{\mathrm{C_s} : \mathrm{s} \in \mathrm{SB}\}, \{\mathrm{f_s} : \mathrm{s} \in \mathrm{SB}\}, \{\tau_s : \mathrm{s} \in \mathrm{SB}\})$ is a topological information retrieval system? Topology systems of attribute values $SB-$ describe the semantic closeness of attribute values and provide a simple and convenient instrument for telling a finite set of pamphlets by a finite and non-empty set of keywords.

In TMIRS, we can distinguish among attribute values $C \subseteq C_s$ topologically. If an element $d \subseteq DS$ of the universe has $f_s(d) \subseteq C$, we say that the keywords of the document d is discerned by the topological rough pair $(\mathrm{int}(C), cl(C))$ with respect to the topology τ_s.

For $C \subseteq C_s$ the set $\mathrm{int}(C)$ is the set of documents, which certainly belong to C. Also, $cl(C)$ is the set of documents, which possibly belong to C. The set $C_s - cl(C)$ is the negative region of those documents that certainly does not belong to C. This interpretation extends to elements of the universe as shown below.

If d is a document of the universe DS such as $f_s(d) = C$, then:

- The attribute values of the $\mathrm{int}(C)$ are certainly belong to the document values of d. We say that $\mathrm{int}(C)$ is the certain values of d.
- The attribute values of $cl(C)$ are possibly belong to the attribute values of d. We say that $cl(C)$ are the possible attribute values of d.
- The attribute values of $C_s - cl(C)$ are certainly do not belong to the attribute values of d.

Let us classify topological multi-valued information retrieval systems into single– granular topological information retrieval systems and multi-granular topological information retrieval systems. In single–granular topological systems, we restrict elements of the universe to have their documents in one and only one class $R_s(C)$ i.e., $f_s(d) = R_s(C), C \subseteq C_s$ is a singleton for every document and for each keyword s .We do not have such a restriction in multi–granular topological information retrieval systems.

6. Indiscernibility of keywords in topological information retrieval systems

Approximately, of the keywords may not be apparent from each other in the wisdom that they may have identical interior and closure approximations in the topological space (C_s, τ_s). More formally, two attributes values $C, C\prime \subseteq C_s$ are indiscernible from each other, denoted $C \approx_{t_s} C\prime$ if and only if $\mathrm{int}(C) = \mathrm{int}(C\prime)$ and $cl(C) = cl(C\prime)$.

It is easy to see that $\approx_{t_s}$ is an equivalence relation on the set 2^{C_s}. The equivalence class of a subset $C \subseteq C_s$ in $\approx_{t_s}$ is denoted $[C]_{\approx_{t_s}}$ and is an element of the quotient set $2^{C_s}/\approx_{t_s}$

Example 6.1. Suppose a set of documents (*DS*) such that each document has a number of keywords (KW). Hence, a given document may characterize by several keywords.

Suppose, DS= {d1, d2, d3, d4} and KW = {{KW11, KW12, KW13}, {KW21, KW22}, {KW31}, {KW41}, {KW51}} such thatd1= {KW13, KW21, KW31}, d2= {KW12, KW21, KW41}, d3= {KW12, KW22, KW31, KW51} and d4= {KW11, KW22, KW41, KW51}.

As shown above the R_w blocks are named $R_s(d1)$, $R_s(d2)$, $R_s(d3)$ and $R_s(d4)$. To obtain the topological elementary sets which discernible by R_w, we construct the topological space (C_s, τ_s) where τ_s is the topology generated by the subbase $S = \{R_s(di) : i = 1, 2, \ldots, 4\}$.

The base of this topology is given by $\beta = \{Rs(di) \cap Rs(dj) : i, j = 1, 2, 3, 4\}$, hence we have β = {d1, d2, d3, d4, {KW21}, {KW13}, φ,{KW12}, {KW41}, {KW22, KW51}}.

Now if we consider a document *di* with the keyword $C\prime = \{KW11, KW51\}$, then the interior approximation $\text{int}(C\prime) = \varphi$ and $cl(C\prime) = d4$ hence, $C\prime$ is not definable in our topological space. But $C\prime$ can be approximated using the topological rough pair $(\varphi, d4)$. Hence, we can only state that the attribute values "*KW11*", "*KW21*", "*KW41*" and "*KW51*" are possibly keywords of d_i according to the interpretation of topological rough pairs.

7. Document classification in topological information retrieval systems

A TIRC has a purpose to group documents of a set into classes or categories according to some knowledge.

Given a topological single–granular information retrieval system S = (*DS*, $\{\tau_s : s \in KW\}$, $\{C_{\boldsymbol{s}}: \boldsymbol{s} \in KW\}$, f_s) and an attribute *w* of *KW*, a topological classification over C_s, or class topological classification, represents a type of knowledge about the information retrieval system S. This specific type of knowledge about the information retrieval system S with respect to an attribute $s \in KW$ is defined to be a set of subsets of C_s, denoted L_I, and defined as:

$$L_I = \{C_s : s \in I, \forall s \neq m, \text{int}(C_s) \cap cl(C_m) = \varphi\}$$

By forming the subsets C_s we are able to express that some documents are reclassified or assigned to a category *S* in the index set *I*.

We defined the set L_S as:

$$L_S = \{E \cap C_s : C_s \in L_I, E \in C_a/\equiv \tau_s\}, \text{ where } E \cap C_s = \cup\{C \cap C_s : C \in E\}.$$

The set L_S is the set of classes that are assigned to category *n* in our topological single–granular information retrieval system.

Let L_S be a topological classification and $C \in L_S$ for a fixed set *S*. We use the usual rough set interpretation; hence, given an attribute value $c_s \in C$ the following interpretation is used:

- If $c_s \in \text{int}(C)$ then we say that c_s is certainly belong to the category *S*, denoted by the topological classification rule $c_s \xrightarrow{Ok} n$.
- If $c_s \in cl(C)$ then we say that c_s is possibly belong to the category *n*, denoted by the topological classification rule $c_s \xrightarrow{?} n$.

Note that the condition $\text{int}(C_s) \cap cl(C_m) = \varphi$, Used in the definition of L_S, ensures that an attribute value cannot certainly assign to category *S* and at the same time, possibly assigned to another category *m*. Consequently, we cannot certainly assign an attribute value to two different categories. Topological classification rules allow us to represent knowledge without the need to discern all keywords of an information retrieval system.

8. Experimental

8.1. Preprocess

Before extracting features of testing data, some preprocessing in the text being performed. All the experiments are performed after normalizing the text. In normalization process the text is converted to UTF-8 encoded and punctuations and non-letters are removed.

They are very common words that appear in the text that carry little meaning; they serve only a syntactic function but do not indicate the subject matter. These stop words have two different impacts on information retrieval process. They can affect the retrieval effectiveness because they have a very high frequency and tend to diminish the impact of the frequency difference among less common words. Identifying a stop words list or a stop-list that contains such words in order to eliminate them from text processing is essential to an information retrieval system. We explore the use of stop words and their effect on Arabic information retrieval. A general stop-list is created, based on the Arabic language structure and characteristics without any additions.

8.2. Experimental evaluation

8.2.1. Simulation results for classification

- Input data: (keywords, text),
- Output: classified documents according to feature selection by topology.

We have used about 130 keywords selected by a human from the corpus.

In information retrieval content, precision and recall are defined in expressions of a set of retrieved documents (e.g. the list of documents created by a web search engine for a query) and a set of related documents (e.g. the list of all documents on the internet that are relevant for a certain topic), cf. relevance.

Precision is the number of correct outcomes divided by the number of all returned outcomes. Namely, precision ($P(di)$) of documents *di* is the division of the cardinality of the intersection of retrieved documents ($Ret(di)$) with the relevant documents ($Rel(di)$) that are relevant to the query by the cardinality of retrieved documents:

$$P(di) = |Rel(di) \cap Ret(di)|/|Ret(di)|.$$

Precision is second-hand with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together to present a single measurement for a system called F measure.

Recall is the number of correct outcomes divided by the number of outcomes that should have been returned. Recall ($R(di)$) of documents *di* is the division of the cardinality of the intersection of retrieved documents ($\text{Ret}(di)$) with the relevant documents ($\text{Rel}(di)$) that are relevant to the query by the cardinality of relevant documents:

$$R(di) = |Rel(di) \cap Ret(di)|/|Rel(di)|.$$

The F measure ($F(di)$) is the division of twice precision and recall by the sum of precision and recall.

$$F(di) = \frac{2 \times P(di) \times R(di)}{P(di) + R(di)}.$$

8.3. Experimental data

For evaluating the performance of our approach, we take on the performance measures Precision (P) and Recall (R) in our system.

Table 1
Precision, recall and F- measure for document classification using TIRC method.

Name of field	Precision	Recall	F-measure
medicine	0.67	0.91	0.77
Arts	0.74	0.66	0.69
Commercial	0.72	0.84	0.77
Politics	0.98	0.79	0.87
Technology	0.57	0.86	0.68
Science	0.44	0.75	**0.55**
History	0.67	0.84	0.75
Sport	0.59	0.64	0.73
Health	0.81	0.95	0.87
Economics	0.59	0.79	0.69
Biology	0.78	0.98	**0.88**

Our experiments, trained the system using documents collected from the Internet. Our data are composed of Al-Jazeera news, Al-Ahram broadsheet, Al-Watan paper, Al Akhbar, Al Arabiya and Wikipedia the free encyclopedia and more. The number of files in our corpus is 1819 files and it is about 26.4 megabytes. There were various topics related to Sports, Computers, Politics, Economics, etc. The corpus partitioned into 11 super fields and 24 subfields. Precision, Recall, and F-measure are measured extracted from the evaluation results, it turns out as the results given in Table 1. Effectiveness of document retrieval system is evaluated by using pair wise document comparison. Comparison of Recall and Precision allows ranking of the retrieval efficiency.

The F measure accuracy of Table 1 record the maximum result in the super field Biology, its reach to 0.88. This is due to the high recall that is 0.98. If we calculate the average mean for all experiment, it will be about 0.75 accuracy for the F measure. This is consider a very good result for document classification.

9. Conclusion and future work

We suggested a new method depending on the neighborhoods for improving information retrieval. The proposed model has many advantages such as topological information retrieval systems that afford approximate retrieval may play an important role in the wild development of existing information.

In addition, we can view feature selection as a technique for replacing a complex classifier (using all features) with a simpler one (using a subset of the features). We present a new technique for term selection using topological space. Moreover, introduced the topological classification method.

References

[1] Abd El-Monsef ME, El-Sayed Atlam, Amin M, El-Barbary O. Arabic document classification: a comparative study. J Comput 2011; 3(4).
[2] Abd El-Monsef ME, El-Sayed Atlam, Amin M, El-Barbary O. Field association words with Naive Bayes classifier base Arabic document classification. IJCSI 2011; 8(3).
[3] El-Sayed Atlam, El-Barbary O. Combining FA words with vector space models for Arabic text categorization. Inf-Int Interdisciplinary J 2013; 16(6) (A): 3517–28.
[4] Atlam El-Sayed, El-Barbary O. Arabic document summarization using fuzzy ontology. Int J Innovative Comput 2014;10(4):1351–67.
[5] El-Barbary OG. Using field association words and K-means cluster for Arabic document classification. Ciniea ea Tecnia J 2015;30(3):287–99.
[6] El-Barbary OG. Using Arabic Skelton morphology and maximum entropy for Arabic document classification. Br J Math Comput Sci 2016;14(3).
[7] Pawlak Z. Rough sets. IJCIS 1982;11:341–56.
[8] Pawlak Z, Skowron A. Rough sets: some extensions. Inf Sci 2007;177:28–40.
[9] Pawlak Z, Skowron A. Rough sets and Boolean reasoning information sciences 2007;177:41–73.
[10] Rosario SF, Thangadurai K. Karur and Tamil Nadu, RELIEF: Feature Selection Approach. ijird, 2015; 4(11).
[11] Salama AS, El-Barbary OG. New topological approaches for data granulation. J Software Eng Appl 2013. doi: https://doi.org/10.4236/jsea. 2013.67B001.1-6.
[12] Salama AS, El-Barbary OG. Future applications of topology by computer programming. Life Sci J 2014; 11(4): 168–72.
[13] Salama AS, El-Barbary OG. New approach to vocabulary mining and document classification. Life Sci J 2014: 84–91.
[14] Salama AS, El-Barbary OG. Multi topological approximations of rough set theory. Int J Granular Comput, Rough Sets Intell Syst 2012: 1–19. http://doi.org/10.1504/IJGCRSIS.2013.054120.
[15] Salama AS, El-Barbary OG. Fuzzy-rough set and fuzzy ID3 decision approaches to knowledge discovery in datasets. J Fuzzy Set Valued Anal 2012: 1–25. http://doi.org/10.5899/2012/jfsva-00118.
[16] Jannik Strtgen, Leon Derczynski, Ricardo Campos, Omar Alonso. Time and information retrieval: introduction to the special issue. Inf Process Manage 2015; 51(6): 786–90.
[17] Silvestri Fabrizio, De Francisci Gianmarco, Morales Roi Blanco. Into the news: online news retrieval using closed captions. Inf Process Manage 2015;51 (1):148–62.
[18] Shakery Azadeh, Rahimi Razieh, King Irwin. Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework. Inf Process Manage 2016;52(2):299–318.
[19] Zhu William. Topological approaches to covering rough sets. Inf Sci, Inf Comput Sci Intell Syst Appl 2007;177:1499–508.
[20] Zhenjun Zhang, Bo Jiang, Feiyue Qiu, Liping Wang, Bi-level weighted multi-view clustering via hybrid particle swarm optimization. Inf Process Manage 2016; 52(3): 387–398.