Egyptian Informatics Journal xxx (2018) xxx–xxx

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 xxxx

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
1110-8665/© 2018 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.

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 $\text{int}(X) = \cup\{Y \subseteq U : Y \subseteq X, Y \in \tau\}$ 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. $\text{TMIRS} = (\text{DS}, \text{SB}, \{\text{C}_s : \text{s} \in \text{SB}\}, \{\text{f}_s : \text{s} \in \text{SB}\}, \{\tau_s : \text{s} \in \text{SB}\})$ 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 $(\text{int}(C), cl(C))$ with respect to the topology τ_s.

For $C \subseteq C_s$ the set $\text{int}(C)$ 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 $\text{int}(C)$ are certainly belong to the document values of d. We say that $\text{int}(C)$ 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 $\text{int}(C) = \text{int}(C\prime)$ 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 $2^{C_s} / \approx_{t_s}$

Please cite this article in press as: El Barbary OG, Salama AS. Feature selection for document classification based on topology. Egyptian Informatics J (2018), https://doi.org/10.1016/j.eij.2018.01.001