Natural Language Engineering (2020), **26**, pp. 221–243
doi:10.1017/S1351324919000160

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

A new approach for textual feature selection based on N-composite isolated labels

Samir Elloumi

University of Tunis El Manar, Faculty of Sciences of Tunis, Computer Science Department, Tunis, Tunisia
*Corresponding author. Email: samir.elloumi@fst.utm.tn

(Received 9 May 2018; revised 23 November 2018; accepted 14 March 2019; first published online 29 April 2019)

Abstract

Textual Feature Selection (TFS) aims to extract relevant parts or segments from text as being the most relevant ones w.r.t. the information it expresses. The selected features are useful for automatic indexing, summarization, document categorization, knowledge discovery, so on. Regarding the huge amount of electronic textual data daily published, many challenges related to the semantic aspect as well as the processing efficiency are addressed. In this paper, we propose a new approach for TFS based on Formal Concept Analysis background. Mainly, we propose to extract textual features by exploring the regularities in a formal context where isolated points exist. We introduce the notion of N-composite isolated points as a set of N words to be considered as a unique textual feature. We show that a reduced value of N (between 1 and 3) allows extracting significant textual features compared with existing approaches even for non-completely covering an initial formal context.

Keywords: Textual features; Formal concept; Composite isolated point; Difunctional relation

1. Introduction

With the intense growth of online information presented mainly in a free-text format, there is an increasing need for powerful intelligent systems that can automatically discover useful knowledge from electronic text sources, commonly referred to as Text Mining (TM) (Agrawal and Batral 2013; Talib *et al.* 2016). TM, also known as intelligent text analysis, text mining, or knowledge discovery in text, refers generally to the process of extracting interesting, non-trivial information, and knowledge from unstructured text. As most information (over 90%) is stored as text (Berger 2012), TM is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information. Yet, unstructured text remains the largest readily available source of knowledge (Ronen and James 2006; Gupta 2009; Gharehchopogh 2010). Applications on TM can potentially yield tremendous benefits to both individuals and organizations. They involve information retrieval, categorization, opinion detection from social webs, topic tracking, information visualization, and so on (Stavrianou, Andritsos, and Nicoloyannis 2007; Rodriguez-Esteban 2019).

Textual Feature Selection (TFS) aims to describe the content of Textual Documents (TDs) into smaller space while handling efficiently a large number of words in TDs and providing a kind of semantic metadata that is useful for a variety of purposes. For context-based TM, three main techniques in the literature have been widely used: (i) Bag of Words (BOW), (ii) Bag of Phrases (BOP) (bag of prototypical-phrases or simply phrase extraction), and (iii) Bag of Concepts (BOC). The BOW technique deals with all document words (Witten *et al.* 1999). Hence, no correlation

© Cambridge University Press 2019

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

(x_0) The United Nations was formed at the end of the Second World War in October.
(x_1) It was formed to try to make sure that future generations would not suffer from horrible damage caused by wars.
(x_2) It was also intended to develop friendly relations between countries around the world and to help with any problems countries might have.
(x_3) The most important part of the United Nations has always been the Security Council.
(x_4) The Security Council was formed to try to keep peace and security around the world.
(x_5) In the beginning the security council consisted of eleven countries; now it has fifteen.
(x_6) Five of these are permanent members.

Figure 1. An example of a TD.

between them has been considered. For the BOP, the sequences of more occurring adjacent words are selected. These types of phrases known as *keyphrases* (Hasan and Ng 2014; Berend 2016) can be considered as highly relevant for describing a TD. The last technique, the BOC, focuses on the identification of surrounding words (i.e., word concept) for a given word present in a TD. This ensures the link between a word and its ontology concept (Bernotas *et al.* 2007).

Despite the widespread applications domains and techniques, TFS faces different challenges, mainly because of the richness and ambiguity of natural language processing. Also, the effectiveness of TFS is limited by the large volume of textual data, as well as its complex and noisy characteristics (Ma *et al.* 2007). These kinds of challenges can be addressed by considering semantic aspects in textual data pre-processing and by proposing more efficient algorithms for TFS based on well-founded mathematical concepts. As a matter of fact, in this paper, we propose to explore the links between Formal Concept Analysis (FCA), Relational Algebra (RA) and TM for TFS. In Elloumi, Ferjani, and Jaoua (2016), the isolated points were proposed as concept labels where the objective was to obtain the conceptual coverage of a Formal Context (FCT). For textual data, naturally, redundant information is usually present. That is why we can easily summarize large-size text with few words while preserving the most important content. We argue that the coverage criterion in TFS is not as important as the quality of extracted textual features. For instance, we could extract only few words or composite words that reflect the main topics in a set of documents, particularly, in the case of redundancy.

As a mathematical background, FCA and RA have been already combined and used to discover regularities in data (Kcherif, Gammoudi, and Jaoua 2000). In fact, a Formal Concept (FC) represents the atomic regular structure for decomposing a Binary Relation (BR). The Riguet's difunctional relation (fringe relation) (Riguet 1948), whose elements are defined as isolated points, describes invariant regular structures that could be used for database decomposition and TFS (Elloumi, Ferjani, and Jaoua 2016). In this paper, we propose to discover such regular structures for different levels $(l_1, l_2 \ldots, l_N)$ of i word combinations, where $i \in [1, N]$ and N is the maximal number of words to be combined together. In practice, $N = 2$ means that the words in the FCT are combined 2 by 2 in order to locate the 2-composite isolated points. For each level l_i, we show how the i-composite isolated points are located and selected as i-composite textual features.

As an introductive example, let us consider the textual document depicted in Figure 1. For level l_1, we can obtain {members, damage, council, world, security} as textual features. For level l_2, they can be {nation-war, nation-world, council-nation, nation-security, council-world}. In this paper, we explain why and how these textual features are selected. It is organized as follows. In Section 2, we present the work related to TFS and their usefulness in TM applications. In Section 3, we provide the basic definitions related to FCA, RA, and TM. In Section 4, we present the links between N-composite isolated points and discovering information regularity for different levels. In Section 5, we introduce our new approach for TFS. In Section 6, we present the experimental results, and finally, in Section 7, we draw some conclusions and give some proposals for further work.

2. Related work

Several papers have already discussed the TFS literature review (Kosala and Blockeel 2000; Dasgupta *et al.* 2007; Hasan and Ng 2014; Irfan *et al.* 2015). Basically, three main techniques for TFS have been widely used and categorized as follows: (i) BOW, (ii) BOP or bag of

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

prototypical-phrases, and (iii) BOC. The BOW technique deals with all words in TDs (Witten *et al.* 1999) without considering any correlation between them. In Yang and Pedersen (1997), a comparative study for five BOW selection methods concluded the strong correlation between document frequency, information gain, and χ^2 statistics. As existing approaches, we could cite the single BOW and the single Keywords. In the single BOW, we use all words except for empty ones. The single Keywords are provided by experts in the field, as cited in Elloumi, Ferjani, and Jaoua (2016). Despite of their easy representation, the major limitations are (Salton and McGill 1983):

- BOW models not respecting the word semantics
- Problem with collocations (juxtaposition of words)
- Word combination and order generally ignored
- Over-fitting and selecting a limited number of features

Our proposed approach extracts single BOW for 1-composite isolated points since we extract unique words. For BOP, the sequences of more occurring adjacent words are selected (Hasan and Ng 2014). These types of phrases, known as *keyphrases*, can be considered as highly relevant for describing a TD. A couple of examples of these approaches are KEA (Witten *et al.* 1999) and Frequent-Term-Set (Agrawal, Imielinski, and Swami 1993). The main characteristics of BOP are as follows:

- They group all words that are interrelated.
- Some are domain dependent, for instance, the approaches based on AGROVOC vocabulary (Rajbhandari and Keizer 2012).
- In the Frequent-Term-Set approach, the frequent term set is extracted based on the Apriori algorithm (Agrawal, Imielinski, and Swami 1993) where the parameters, support, and confidence are user predefined.
- Redundant and infrequent terms could mislead the analysis and increase the complexity.

The last technique, BOC, focuses on the identification of surrounding words (i.e., word concept) for a given word present in a TD (Bernotas *et al.* 2007). Some approaches have applied FCA on textual data and have extracted relevant textual features from FCs. They are considered BOC as well. In our approach, the set of N-composite isolated points constitutes one BOC. BOC are closely related to mathematical FCA settings. FCA has been shown to provide a theoretical framework for the efficient resolution of many practical problems from data mining (Rodriguez-Esteban 2019), conceptual reasoning (Elloumi *et al.* 2014), software engineering, and information retrieval (Ganter and Wille 1999; Weiss *et al.* 2005; Jaoua *et al.* 2010). An interesting problem is to find a representation of an FCT by minimal or an optimal set of concepts, particularly, in the case of textual data. Suppose that we have a set of documents (i.e., a corpus) that we need to explore and organize. The corpus could be represented by an FCT where the objects are documents and the properties are the words in the corpus, that is, BOW. An FC indicates a strong link between a subset of documents set_{doc} and a subset of words set_w since they are completely interrelated. Each document d in set_{doc} has all words in set_w. Conversely, each word w in set_w should exists in all documents in set_{doc}. Finding a minimal representation of an FCT means the election of a minimal set S_{cov} of concepts, called minimal coverage, where the union is the entire FCT xt, that is, the corpus. Finding the optimal coverage of a BR is known to be an NP-hard problem (Garey and Johnson 1979). Nevertheless, it could mean the election of optimal maximal rectangles (or equivalently FCs) that achieve maximal gain in storage space terms Jaoua *et al.* (2004). Kcherif, Gammoudi, and Jaoua (2000) introduced a rectangular decomposition approach based on the Riguet's difunctional relation (Riguet 1948). The computation of this difunctional was reduced to the determination of a set of isolated points allowing the determination of the minimal set of rectangles covering a given BR. An extended isolated point-based version, called *genCoverage*, was proposed in

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Elloumi, Ferjani, and Jaoua (2016) by considering the extended context notion. Belohlavek and Vychodil (2010) tackled the same issue by attempting to solve the *Boolean factor analysis* problem by proposing a new method of decomposing an $n \times m$ binary matrix I into a Boolean product $A \circ B$ of an $n \times k$ binary matrix A and a $k \times m$ binary matrix B, where k is as small as possible. Mouakher and Ben Yahia (2016) introduced a new approach, based on a greedy algorithm, for the extraction of an optimal covering of a BR. The latter approach relied on the FC lattice representation. The guiding idea of Mouakher and Ben Yahia's approach was that coverage should not be extracted regardless of the quality of knowledge that may be drawn from it. Therefore, the authors introduced a gain function based on the assessment of the correlation of the intent part of pertinent FCs. Generally, the number of FCs grows exponentially with the size of the matrix (i.e., BR)(Godin, Missaoui, and Alaoui 1995). Fortunately, the use of mandatory concepts allows selecting an information with a minimal loss set of few relevant concepts in polynomial time. In this paper, we apply some existing conceptual approaches for BOC extraction. In particular, we cite :

- **Difunctional** (Kcherif, Gammoudi, and Jaoua 2000): It is based on the exploration of the composed labels built on the Riguet difunctional relation (Riguet 1948).
- **Optimal factors** (Belohlavek and Vychodil 2010): It explores the textual features as concept properties discovered by the optimal factor approach.
- **MinGenCoverage** (Elloumi, Ferjani, and Jaoua 2016): It considers the minimal generators as isolated-point candidates. Their corresponding isolated concepts are considered as composite labels.

3. Key settings

In the recent years, a large amount of research has concurrently explored the fields of computer science, RA, FCA, and lattice theory. These researches have enormously contributed to finding original solutions for complex problems in the fields of knowledge engineering, data mining, and information retrieval (Osei-Bryson 2010). Thus, RA and FCA are considered as useful mathematical foundations that have unified data and knowledge in information retrieval systems. In the following, we recall some basic definitions borrowed from RA as well as FCA.

3.1 FCA and textual data

FCA is a mathematical tool for analyzing data and formally representing conceptual knowledge. FCA helps to form conceptual structures from data. These structures, such as closed itemsets, generic bases, and minimal generators, are very useful for TM (Bastide *et al.* 2000). In the remainder of the paper, we use the theoretical framework presented in Ganter and Wille (1999).

Definition 1. **Formal Context**: An FCT is a triplet $\mathcal{K} = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$, where $\mathcal{X}$ represents a finite set of objects or documents, $\mathcal{Y}$ is a finite set of attributes (or words), and $\mathcal{R}$ is a BR (i.e., $\mathcal{R} \subseteq \mathcal{X} \times \mathcal{Y}$). Each pair $(x, y) \in \mathcal{R}$ expresses that the document $x \in \mathcal{X}$ contains the word $y \in \mathcal{Y}$.

Example 1. Let $\mathcal{X} = \{d_0, d_1, d_2, d_3, d_4, d_5, d_6\}$ be a set of documents. Let $\mathcal{Y} = \{$united, nations, world, war, damage, security, council, members$\}$ be a set of words. In Table 1, we present an example of FCT relating the elements of $\mathcal{X}$ with those of $\mathcal{Y}$.

Definition 2. **Itemset**: An itemset is a set of items, that is, words, included in $\mathcal{Y}$ and representing an object, that is, a document. An itemset is also called a termset. In the sequel, we denote by $\mathcal{I}$ the power set of $\mathcal{Y}$.

Definition 3. **Galois Connection**: We define two functions f and g, summarizing the links between the subsets of documents and the subsets of words induced by $\mathcal{R}$. Thus,

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 1. FCT representing the text presented in Table 1

	united	nations	world	war	damage	security	council	members
d_0	1	1	1	1	0	0	0	0
d_1	0	0	0	1	1	0	0	0
d_2	0	0	1	0	0	0	0	0
d_3	1	1	0	0	0	1	1	0
d_4	0	0	1	0	0	1	1	0
d_5	0	0	0	0	0	1	1	0
d_6	0	0	0	0	0	0	0	1

- $f : \mathcal{P}(\mathcal{X}) \rightarrow (\mathcal{P}(\mathcal{Y})=\mathcal{I}), f(X) = \{y \in \mathcal{Y} | \forall x \in X, (x, y) \in \mathcal{R}, X \subseteq \mathcal{X}\}$,
- $g : (\mathcal{P}(\mathcal{Y}) = \mathcal{I}) \rightarrow \mathcal{P}(\mathcal{X}), g(Y) = \{x \in \mathcal{X} | \forall y \in Y, (x, y) \in \mathcal{R}, Y \subseteq \mathcal{Y}\}$.

Definition 4. **Closed Itemset** (Pasquier *et al.* 1999): An itemset $I \in \mathcal{I}$ is said to be closed *if* $f \circ g(I) = I$. $f \circ g(I)$ is called a closure operator.

Definition 5. **Formal Concept**: An FC $C = (A, B)$ is a pair where $A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}, A = g(B)$, and $B = f(A)$. Set A is called the extent of C and set B is called its intent (Ganter and Wille 1999). An FC is also called a *maximal rectangle* or a *non-enlargeable rectangle* (Jaoua *et al.* 2004).

Example 2. From the FCT presented in Table 1, the following pair $C_1 = (A, B) = (\{d_3, d_4, d_5\}, \{security, council\})$ is an FC since $f(A) = B$ and $g(B) = A$. Set B is a closed itemset. Note that all documents in A have all the words in B. Conversely, all words in B belong to all documents in A.

Remark 1. From any FCT, it is possible to locate several FCs covering it, that is, the union of FCs is equal to FCT.

Example 3. By considering the Table 1, we obtain the coverage $Coverage_1 = \{C_0, C_1, C_2, C_3, C_4, C_5, C_6\}$ where :

- $C_0 = (\{d_0\}, \{nation, united, war, world\})$
- $C_1 = (\{d_0, d_2, d_4\}, \{world\})$
- $C_2 = (\{d_1\}, \{damage, war\})$
- $C_3 = (\{d_3\}, \{council, nation, security, united\})$
- $C_4 = (\{d_3, d_4, d_5\}, \{security, council\})$
- $C_5 = (\{d_4\}, \{council, security, world\})$
- $C_6 = (\{d_6\}, \{members\})$

Example 4. From Table 1, we obtain $Coverage_2 = \{C'_0, C'_1, C'_2, C'_3, C'_4, C'_5\}$

- $C'_0 = (\{d_0, d_3\}, \{nation, united\})$
- $C'_1 = (\{d_0, d_2, d_4\}, \{world\})$
- $C'_2 = (\{d_0, d_1\}, \{war\})$
- $C'_3 = (\{d_1\}, \{war, damage\})$
- $C'_4 = (\{d_3, d_4, d_5\}, \{security, council\})$
- $C'_5 = (\{d_6\}, \{members\})$

Remark 2. In (Mouakher and Ben Yahia 2016) the authors discussed the coverage quality by considering different metrics applied to their FCs, such as coupling, cohesion, gain, stability, etc.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 2. Fringe relation corresponding to BR presented in Table

	united	nations	world	war	damage	security	council	members
d_0	0	0	0	0	0	0	0	0
d_1	0	0	0	0	**1**	0	0	0
d_2	0	0	**1**	0	0	0	0	0
d_3	0	0	0	0	0	0	0	0
d_4	0	0	0	0	0	0	0	0
d_5	0	0	0	0	0	**1**	**1**	0
d_6	0	0	0	0	0	0	0	**1**

From Examples 3 and 4, two types of distinct coverage were obtained. In general, varied coverage could be obtained while processing any FCT. Furthermore, in Belohlavek and Vychodil (2010), it was mentioned that certain FCs were mandatory in that they needed to be included in every coverage stage. For instance, the FCs $C_1 = C'_1$, $C_4 = C'_4$ and $C_6 = C'_5$ are mandatory. In this paper, we defend the idea that the labels of these mandatory FCs constitute the first set of textual features. In the following, we introduce the mathematical basis related to difunctionality for locating these concepts based on property composition.

3.2 Difunctionality

In the sequel, we recall some important definitions related to difunctionality, as a closed notion to an uniformity aspect, which we may find or extract from BRs. We focus on difunctional and fringe relations. Let us consider two sets, $\mathcal{X}$ and $\mathcal{Y}$, a relation $\mathcal{R}$, as a subset of the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$, and two elements e, e' where $e \in \mathcal{X}$ and $e' \in \mathcal{Y}$. We start by presenting these basic definitions (Ganter and Wille 1999):

- The relative product or composition of two BRs $\mathcal{R}$ and $\mathcal{R}'$ is $\mathcal{R} \circ \mathcal{R}' = \{(e, e') | \exists t \in \mathcal{Y} : ((e, t) \in \mathcal{R})$ and $((t, e') \in \mathcal{R}')\}$.
- The inverse of relation $\mathcal{R}$ is $\mathcal{R}^{-1} = \{(e, e') | (e', e) \in \mathcal{R}\}$.
- The complement of relation $\mathcal{R}$ is $\overline{\mathcal{R}} = \{(e, e') | (e, e') \notin \mathcal{R}\}$.

Definition 6. **Difunctional Relation** (Ganter and Wille 1999): Let $\mathcal{R}$ be a BR. $\mathcal{R}$ is said to be difunctional *if* $\mathcal{R} \circ \mathcal{R}^{-1} \circ \mathcal{R} = \mathcal{R}$. $\mathcal{R}$ can be written in a block diagonal form by suitably rearranging rows and columns. This means that a difunctional relation presents the uniformity of the association of images and antecedents.

Even though a BR is not always uniform as difunctional, *Riguet* has proved that there is an interesting difunctional embedded in it, which is generally not empty (Riguet 1948). This relation is called a *fringe* relation that we define in the following.

Definition 7. **Fringe Relation** (Riguet 1948): A fringe relation is a difunctional relation embedded in a BR $\mathcal{R}$ and computed by $\mathcal{R} \circ \overline{\mathcal{R}^{-1}} \circ \mathcal{R} \cap \mathcal{R}$. It is denoted as $\mathcal{R}_d$. Moreover, a fringe relation is composed by a set of points contained in just one maximal rectangle inside relation $\mathcal{R}$ which can play an important role in different applications.

Example 5. In Table 2, we present the fringe relation corresponding to the $\mathcal{R}$ BR illustrated in Table 1. The difunctionality aspect is observed in the fringe relation as a sparse subrelation.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

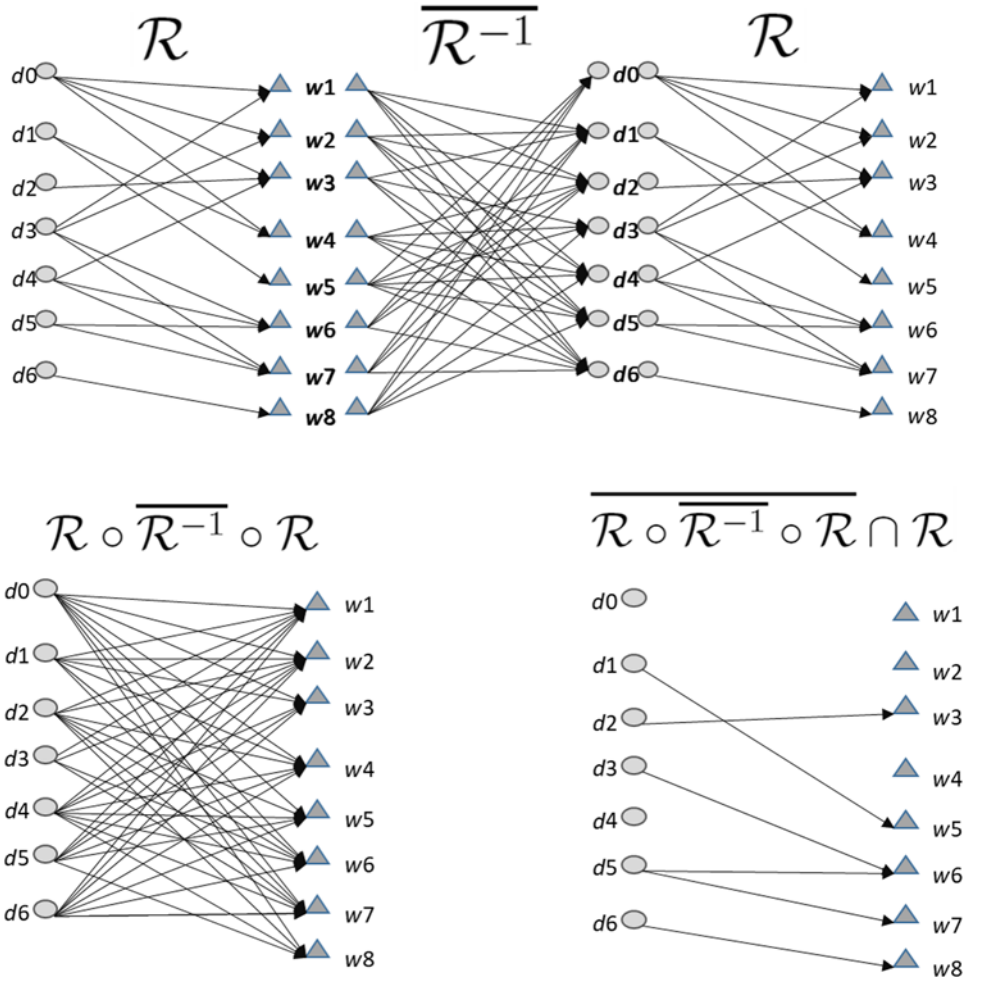


Figure 2. Steps for obtaining the fringe relation corresponding to the BR presented in Table 1.

Remark 3. To understand the formula $\overline{\mathcal{R} \circ \overline{\mathcal{R}^{-1}} \circ \mathcal{R} \cap \mathcal{R}}$, we present in Figure 2 the steps for obtaining the fringe relation corresponding to the BR presented in Table 1. The words w_t, $t = 1 \ldots 8$, correspond the BR words. For example, w_1=united, w_2= nations.

- Relation $\mathcal{R}$ is relating documents to their words. For instance, d_5 has two words: w_6 and w_7. Hence, the pairs (d_5, w_6) and (d_5, w_7) belong to $\mathcal{R}$.
- Relation $\overline{\mathcal{R}^{-1}}$ relates documents to words which are not related in $\mathcal{R}$. For instance, it contains the pairs (d_5, w_1), (d_5, w_2), (d_5, w_3), (d_5, w_4), (d_5, w_5), (d_5, w_8).
- Relation $\mathcal{R} \circ \overline{\mathcal{R}^{-1}}$ produces a relation between documents d_i and d_j if there exists at least one word in d_i and not in d_j. For instance, we get pairs (d_5, d_0), (s_5, d_1), (d_5, d_2) (d_5, d_3), (d_5, d_4),(d_5, d_6)
- Relation $\mathcal{R} \circ \overline{\mathcal{R}^{-1}} \circ \mathcal{R}$ relates documents to words, say (d_i, w_t), for w_t not in d_i if there exists at least a document d_j having w_t while having other words not existing in d_i. For example, we get the pairs (d_5, w_1), (d_5, w_2), (d_5, w_3), (d_5, w_4), (d_5, w_5), (d_5, w_8).

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

- Relation $\overline{\mathcal{R} \circ \overline{\mathcal{R}^{-1}} \circ \mathcal{R} \cap \mathcal{R}}$ expresses the negation of the previous condition. It means that if d_i is connected to w_t, then for all documents having w_t d_i will have no connection with any of the other words, for example, we get the pair (d_5, w_6) and (d_5, w_7).

The elements belonging to a fringe relation are called *isolated points* and their usefulness are shown in extracting concept's labels and TFS (Elloumi, Ferjani, and Jaoua 2016).

4. Isolated points and *N*-composite isolated points

In this section, we present the necessary definitions related to isolated points. We recall some of their useful properties and we present novel issues for obtaining N-composite ones as relevant textual features.

Definition 8. **Isolated Point** Kcherif *et al.* (2000): Let $\mathcal{R}$ be a BR and $\mathcal{R}_d$ be its associated fringe relation. If $p = (x, y) \in \mathcal{R}_d$, then p is called an isolated point. An isolated point belongs to only one FC and the latter is called an isolated concept.

Example 6. With respect to the BR presented in Table 1, the FC $C = (A, B) = (\{d_3, d_4, d_5\}, \{security, council\})$ is an isolated one since $(d_5, council)$ is an isolated point.

Proposition 1. (Ferjani *et al.* 2012). Let $p = (x, y)$ be an isolated point in an FC $C = (A, B)$. The following properties are satisfied:

$$\begin{aligned} g(\{y\}) &= A \\ f(\{x\}) &= B \\ f \circ g(\{y\}) &= B \\ g \circ f(\{x\}) &= A \end{aligned}$$

Remark 4. From proposition 1, particularly, the expression $f \circ g(\{y\}) = B$ means that if a document d in $\mathcal{X}$ contains the word y, then it contains necessarily all words in B. Thus, we have the implication $y \rightarrow B \backslash \{y\}$. The word y can be selected as an interesting candidate representing all words in B since they are implied by y.

Definition 9. **Concept Label**: If $p = (x, y) \in \mathcal{R}$ is an isolated point in an isolated concept C, then y is called a ***concept label*** of C.

Example 7. The word *council* is selected as a concept label of C_4, and the implication *council* $\rightarrow$ *security* is true.

Remark 5. For textual data, the labels can reflect the indexing words or text features (Ferjani *et al.* 2012). In many cases, no single label can be associated with an FC since it does not contain single isolated points. To overcome this situation, we suggest composite properties as candidates for concept labeling. In this paper, we propose to consider N-composite properties to discover the composite labels. We formalize these ideas in the following.

Definition 10. **Composite Properties** (Ferjani *et al.* 2012): Let $(\mathcal{X}, \mathcal{Y}, \mathcal{R})$ be an FCT, and let $SP \subseteq \mathcal{Y}$, that is, a subset of properties such that $|SP| \geq 2$. A composition of the properties of SP defines a new composite property, denoted by $CP_{SP} = p_1.p_2 \ldots p_n$, $n = |SP|$, $p_i \in SP$, $i = 1 \ldots n$, which can be added to set $\mathcal{Y}$. This new property CP_{SP} can be used to define an extended context as well as to select a composite label for an FC when a single label cannot be assigned.

Definition 11. ***N*-Composite Property**: Let N be a positive integer and $SP \subseteq \mathcal{Y}$. An N-composite property denoted by NCP_{SP} is a CP_{SP} such that $|SP| = N$.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 3. Composite words extracted for different levels

***N*-composite properties**	**Examples**
1-composite	{united}, {nation},{security},{council}
2-composite	{united, nation} ,{security, council} {war, damage}
3-composite	{united,nation,world}, {security,council,member}

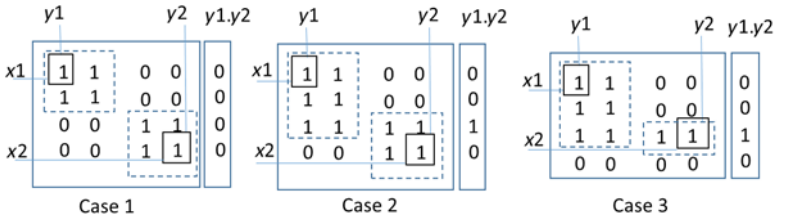


Figure 3. Three cases for isolated-point compositions.

Example 8. Let us consider the FCT presented in Table 1. In Table 3, we present examples of 1-composite, 2-composite, and 3-composite properties or words.

Definition 12. **Expanded Context**: Let $FC = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$ be an FCT and let NCP_{SP} be an N-composite property. NCP_{SP} can be added to set $\mathcal{Y}$ to define an expanded context: $EXP_{(FC,SP)} = (\mathcal{X}, \mathcal{Y} \cup NCP_{SP}, \mathcal{R}')$ where $R' = R \cup \{(x, NCP_{SP}) | (x, y) \in R, \forall y \in SP\}$.

Definition 13. ***N*-Composite Property Isolated Point**: Let $FC = (\mathcal{X}, \mathcal{Y}, \mathcal{R})$ and NCP_{SP} be an N-composite property. NCP_{SP} is an N-composite isolated point if there exists $x \in \mathcal{X} | (x, NCP_{SP})$ which is an isolated point in the expanded context $EXP_{(FC,SP)}$.

Remark 6. Why is it important to compose properties?

There are two main reasons:

- Discover the hidden regularity in the data implied by the N-composite isolated points. Therefore, it would be selected as a textual feature.
- Discover additional isolated concepts in order to cover more documents in a corpus. Note that for TFS, the coverage criteria is not the most important aspect. We are more focusing on extracting representative textual features without having necessarily 100% of coverage percentage.

Remark 7. How can we compose properties?

The isolated points have some important properties that will prevent any randomized word combination. In particular, we will check whether some word combinations are useless or not, mainly when isolated points are involved. Let us consider Figure 3 in which we distinguish three cases for isolated-point combination. In each case, we have two isolated points (x_1, y_1),(x_2, y_2) and their corresponding FC marked by dotted rectangles. $y_1.y_2$ denotes the 2-composite property obtained from $\{y_1, y_2\}$. We have :

- Case 1: The isolated points belong to two distinct concepts. Therefore, the intersection between $g(\{y_1\})$ and $g(\{y_2\})$ is empty, and it is useless to make the isolated-point composition. Proposition 2 formalizes case 1.
- Case 2: The intersection between $g(\{y_1\})$ and $g(\{y_2\})$ is not empty and we will check if some additional isolated points might be derived when considering $y_1.y_2$.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

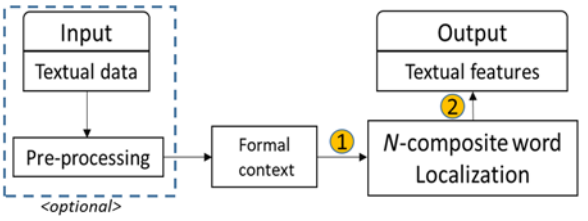


Figure 4. Features extraction process.

- Case 3: Even if the intersection of $g(\{y_1\})$ and $g(\{y_2\})$ is not empty, their composition is useless. In fact, the isolated point (x_2, y_2) belongs to an already covered concept. Here, the proof is trivial since an isolated point belongs to a unique concept.

Proposition 2. Let $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ be two isolated points in two distinct concepts $C_1 = (A_1, B_1)$ and $C_2 = (A_2, B_2)$. Let $SP = \{y_1, y_2\}$ and $2CP_{SP}$ its corresponding 2-composite property. The following property is satisfied:

- If $A_1 \cap A_2 = \emptyset$, then $2CP_{SP}$ is empty; that is, $g(\{SP\}) = \emptyset$. In other words, no 2-composite isolated point is derivable.

Proof. $g(SP)=\{x \in \mathcal{X} | \forall y \in \{y_1, y_2\}, (x, y) \in \mathcal{R}, Y \subseteq \mathcal{Y}\}$
$= \{x \in \mathcal{X} | (x, y_1) \in \mathcal{R} \wedge (x, y_2) \in \mathcal{R}\}$
$= g(y_1) \cap g(y_2)$ by definition
$= A_1 \cap A_2 = \emptyset$ by hypothesis. □

Remark 8. This property is important in the property composition algorithm. It shows that it is useless to make a composition between two isolated points when they belong to two distinct concepts. Moreover, the composition would give isolated points where all related concept's elements were covered.

Remark 9. As long as we increase the value of N, for the N-composite properties, the possibly isolated N-composite points and their associated FCs are covering additional elements in the BR. Generally, for TFS, when the size of data is big, it is recommended to limit the value of N to a small value (e.g., 2 or 3) even for non-covering the whole relation. Hence, combining non-isolated properties will give newly isolated points.

5. Using *N*-composite isolated points for textual feature selection

This section represents our contribution in this paper. We start by presenting the system architecture, as depicted in Figure 4, and we develop the N-composite word localization algorithm for TFS.

5.1 System architecture

We propose to extract the textual features by the steps (1) and (2) as illustrated in Figure 4. From the processed textual data, we obtain an FCT relating documents to words. For a level i from 1 to N, we iteratively locate the isolated i-composite words as indicated in Algorithm 1. Their union is the set of textual features.

In the case of new collected data sets, for example, the data collected in the scope of the Financial Watch Project (FW-Data)(Jaoua *et al.* 2010), we recommend to apply a pre-processing phase before starting the TFS as indicated in the $< optional >$ step. It consists in applying the natural language processing steps (e.g., stopword removal, synonymy processing) to prepare the FCT construction. This step is important before transforming the textual data into an FCT.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

In several benchmark data sets, this step has been already done for some versions. For instance, it is the case of the 20-newsgroup data sets (Rennie 2008) which are easy to read into Matlab/Octave as a sparse matrix. Hence, we just need to rearrange data into our algorithm input format.

5.2 Main algorithm

The pseudo-code presented in Algorithm 1 reflects the steps we follow to compute the N-composite isolated properties and to extract the textual features. First, we identify the set of isolated points and their corresponding concepts. If the FCT is not covered and the composition level N is not reached, then we iterate on the remaining set of composite words and keep only those representing the isolated points.

- In lines 2–10, we prepare the closure sets related to documents and words, respectively. The routine *BuildObjectsList* constructs *OL* and *WordsList* as follows:
  - *OL*: It has an hMap structure or a dictionary (as defined in Python), where, $OL \leftarrow \{\text{obj} : f(obj)$ **for** obj **in** $TDs\}$. $f(obj)$ is the list of words in obj.
  - *WordsList* is the set of all words in TDs.
- For each word w in WordsList, we compute $g(w)$ and its closure $f \circ g(w)$. This result is added to the *levelClosures* list defined within *WordClosure*. As long as the level increases, we get level-composite words. Hence, the variable *Wordclosure* stores the composition level and the closures of the level-composite words in levelClosures.
- In lines 12–18, we locate the current level-composite isolated point w.r.t. the properties presented in Proposition 1. We update the set of features *SF* and we remove the covered elements from the set *ListNonCovered*. The *ExtractCompositeIsolated* routine considers Proposition 1 and makes a traversal of *Wordclosure.levelClosures* in order to identify the isolated composite points. Note that for level 1, *composites* represent the 1-composite isolated points. The *Feature* routine extracts the words of isolated points. The routine *Concepts* returns a set of pairs $setP = \{(object, word)\}$ forming the FC associated with composites. The elements of *setP* are removed from *ListNonCovered* to check the coverage condition.
- In lines 19–23, the routine *combine* combines (the new level value) words as new candidates of isolated composite points based on the previous level-composite elements. For instance, suppose for level 2, we have the composite properties $\{w_1.w_2, w_2.w_3, w_3.w_4\}$, then, for level 3, we get $\{w_1.w_2.w_3, w_1.w_2.w_4, w_2.w_3.w_4\}$. In order to reduce the combinatory number, we consider Remark 7 and Proposition 2.

5.2.1 Complexity analysis

Let $n = |\mathcal{X}|$ and $m = |\mathcal{Y}|$ be, respectively, the number of documents and the number of words in the TDs.

- In phase 1, the procedures *BuildObjectsList* and the construction of *WordClosure* require $n \times m$ operations.
- In phase 2, the algorithm performs the *WordClosure.levelClosures* traversal and locates the level-composite isolated points by applying the *ExtractCompositesIsolated* function. In the worst case, we have $M = max(N, MaxLevel)$ where N is the number of traversals and *MaxLevel* is the max level specified by a user. Let us consider a composition level i for which $|WordClosure.levelClosures| = m_i$. The *WordClosure.levelClosures* property composition requires $m_i + (m_i - 1) + (m_i - 2) + \cdots + 1 = \frac{m_i(m_i+1)}{2}$ operations. Hence, there are $\sum_{i=1}^{M} \frac{m_i(m_i+1)}{2}$ operations.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

```
   Input: - TDs: textual documents represented as an FCT, N: Max Level
   Output: - SF: Set of Features
 1 begin
 2     SF ← NULL
 3     OL ← NULL Objects list
 4     //Phase 1: Prepare the closures
 5     BuildObjectsList(TDs,OL,WordsList)
 6     // WordClosure initialization WordClosure.level← 1
 7     WordClosure.listComposites ← WordsList
 8     foreach word w ∈ WordsList do
 9         // for each w we build its closure in a levelClosures list
10         WordClosure.levelClosures.add(w,f ∘ g(w))
11     //Phase 2:
12     level ← 1
13     ListNonCovered ← pairsIn(OL)
14     while (level ≤ N) and NotEmpty(ListNonCovered) do
15         composites ← ExtractCompositeIsolated(wordClosure,level)
16         SF ← SF ∪ Features(composites)
17         ListNonCovered.Remove(Concepts(composites))
18         level← level+1
19         wordClosure.addNewLevel(level)
20         i ← 1
21         while (i ≤ wordClosure.level) do
22             LC ← Combine(i,wordClosure,level)
23             wordClosure.listComposites.add(LC)
24     return SF
```

Algorithm 1: TFS algorithm

5.2.2 Illustrative example

Let us reconsider the FCT given by Table 1. Figure 5 illustrates the different phases of Algorithm 1. In phase 1, the *wordClosure* list is constructed. It contains the words, their closures, and their associated documents obtained by the Galois connection. In phase 2, we have two levels. At level 1, the *ExtractCompositeIsolated* routine returns the set of isolated 1-composite points. At level 2, the *combine* routine combines two words w.r.t. Remark 7 and Proposition 2 and returns *ListComposites*. The latter represents the new candidates of isolated composite points. In the second iteration, the *ExtractCompositeIsolated* returns the isolated 2-composite points and the FCT is entirely covered.

5.2.3 Discussion

The TFS algorithm output represents the set of isolated concepts with their labels. Hence, we have two results: (i) It is partial-total FCT coverage, so it informs about the coverage percentage. (ii) It represents single and possibly composite textual features reducing the TDs dimensionality. Consequently, we need to be informed about their relevance by using appropriate metrics.

- (i) Partial-total FCT coverage: As we can see in Example 5, at level 1, the isolated 1-composite points are associated with the concepts $\{C_1, C_2, C_4, C_6\}$ as listed in Example 3. They cover

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

```
Input: WC: WordClosure , L:Level, OL: Objects List
Output: CP: composites isolated words for level L
1  begin
2      Locate the listComposites for level L.
3      listComposites ← WC.getListComposites(L).
4      foreach word w ∈ listComposites do
5          OL ← g(word)
6          foreach O ∈ OL do
7              // check if word is an isolated
8              if f ∘ g(word) = f(O) then
9                  CP.add(word)
10     return CP;
```

Algorithm 2: EXTRACTCOMPOSITESISOLATED function

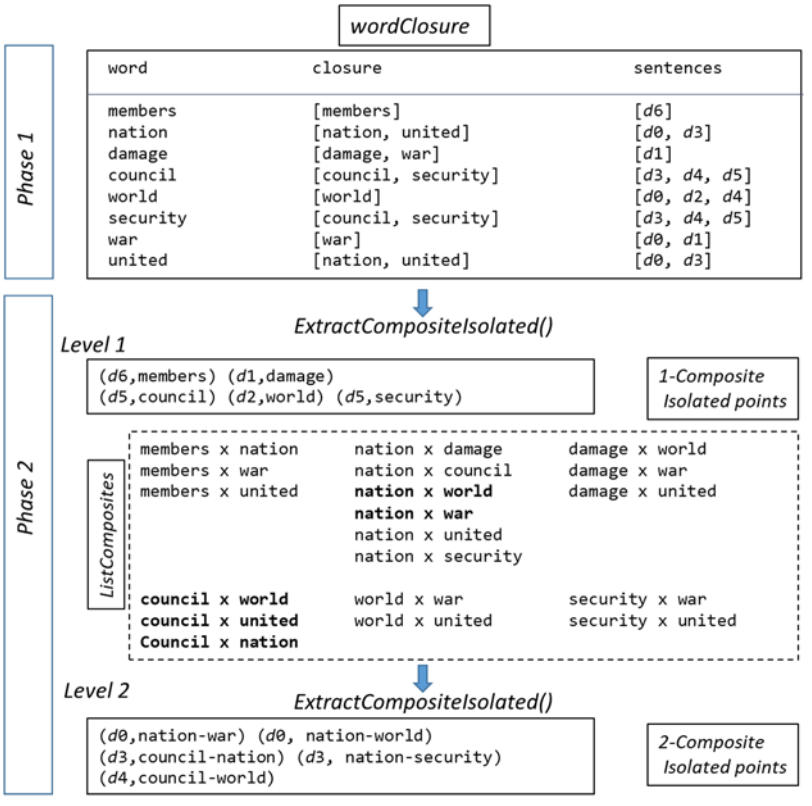


Figure 5. Illustrative example.

12/17 or (70.58%) elements in the FCT given by Table 1. At level 2, the isolated 2-composite points are associated with the additional concepts $\{C_0, C_3, C_5\}$. As a result, we get 100% coverage.

- (ii) Single and composite properties: With respect to Example 5, at level 1, we get five 1-composite isolated points, that is, five selected features. Thus, we have 3/8 or 37.50% as a reduction rate. The words **council** and **security** belong to the same concept C_4 so, we can

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

consider the option to keep only one of them and reach a 50.00% reduction rate while covering (70.58%) elements in the FCT. At level 2, the coverage percentage is 100%. Even if the number of isolated 2-composite points is 5, some of them are equivalent as they belong to the same concept. For instance, it is the case of (d_0, nation-war) and (d_0, nation-world). In total, we have at least three additional features.

Remark 10. In this paper, we process all selected features without considering the equivalence between some of them.

6. Experiments

We start by introducing the used textual data sets and the evaluations metrics we used for evaluating the extracted-textual-feature relevance for different composition levels. After that, we present the experimental results.

6.1 Textual data sets

We have considered two data sets: the FW-Data (Jaoua *et al.* 2010) and the 20-newsgroup data sets (Rennie 2008) (Matlab/Octave version).

- *FW-Data*

FW-Data is a collection of approximately 400 TDs collected in the scope of the Financial Watch Project (Jaoua *et al.* 2010) with the collaboration of the CEA-LIST institute.[a] The documents have been already classified in three categories according to the predefined classes in the financial domain which are management change, performance, and transaction. The collected data have been analyzed and annotated by experts in order to build a multilingual platform for knowledge discovery, information extraction, and automatic translation in the financial domain. The pre-processing step is performed as it is new collected data. More particularly, we apply the Standford POS tagger tool (Manning *et al.* 2014) for annotating text with part-of-speech information. Furthermore, we apply a transformation process for getting a uniform representation for some phrases. For example, the phrases "Qatar National Bank," "National Bank," and "QNB" designate the same bank and should be represented by a unique way. Also, we apply the Porter Stemmer algorithm to reduce a token to its root form (Porter 1980). For synonym processing, we create a synonym table and feed it with processed noun entries, captured from Wordnet3 (Miller 1995). Our synonym table contains around 427,000 entries. It is composed of three attributes: IdWord, word, and IdR, where:

- IdWord is the word identifier. It is incrementally obtained while processing the entries.
- word is the stem of the processed word entry.
- IdR is the identifier of its representative word.

In the synonym table, each word has a unique representative one. Thus, the synonym table construction algorithm considers three cases:

- Case 1: Add non existing sets of synonyms to the table of synonyms.
- Case 2: Enlarge the set of synonyms having the same representative word. Hence, we will get new rows in the table.

[a] http://www-list.cea.fr/en/.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 4. Characteristics of FW-Data

Class	# Documents	# Sentences	# Words
Management change	126	1268	9794
Transaction	103	1900	13,668
Performance	147	721	6521
Total	376	3889	29,983

Table 5. Statistical significance of the differences between FW-Data groups

Group 1	Group 2	*t*-stat	Degree of freedom	Critical value	*P*-value
Transaction	Performance	2.381	5078	1.645	0.017
Transaction	M. Change	−2.277	5078	1.645	0.022
Performance	M. Change	−4.168	5078	1.645	0.000

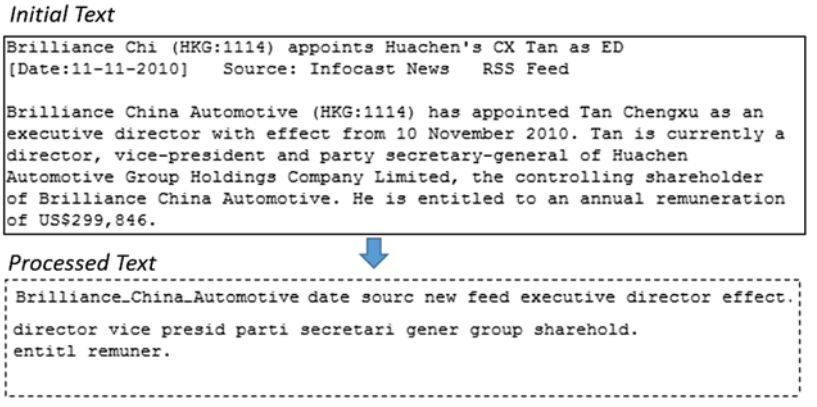


Figure 6. Example of management change text and its transformation.

- Case 3: Ignore the conflictual cases. These cases correspond to a new entry from Wordnet3 having at least two words already saved in the synonym table and having two distinct representative words.

As an illustration, we present in Figure 6 an example of a processed text for the management change class.

There exist also other techniques for finding the similarity between words by considering embedding models such as the GloVe approach (Pennington, Socher, and Manning 2014) and the one estimating the word representation (Mikolov *et al.* 2013). In Table 4, we present the main characteristics of the FW-Data. For each class, we indicate the number of documents, the number of sentences, and the number of words. In addition, we present in Table 5 the statistical significance of the differences between the three groups in FW-Data. The *t*-test, or student's *t*-test (Gosset 1908), is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. The significance level (also called alpha) is set to 0.05. It allows to decide to either reject or accept the null hypothesis, that is, that the population means from the two unrelated groups are equal. Form Table 5, we remark that, in all cases, the

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

absolute value of t-stat is greater than the critical value. Also, we have P is less than alpha. These observations converge to conclude to reject the null hypothesis that the means are equal.

- *20-newsgroups data set*

We have considered the (Matlab/Octave) version which has been already a processed data set. It contains both training and testing sets. We convert the training file to our TFS-algorithm input without considering the class attribute. It is a BR composed of 53,975 words and 11,269 objects.

6.2 Evaluation metrics

We propose to evaluate the extracted-textual-feature relevance by considering the standard measures used for supervised classification. For textual data, the classes are the documents categories. Hence, the metrics are Precision, Recall, F-measure, micro-average accuracy, and macro-average accuracy (Gordon and Kochen 1998). Let us consider N the number of categories in a textual data set and $i \in N$ a given category. The metrics use the parameters TP_i, FP_i, and FN_i standing for true positive, false positive, and false negative for a class i, respectively. They are computed as following:

- $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$
- $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$
- $F_\text{measure}_i = \frac{2\times\text{Precision}_i\times\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i}$
- Micro-average accuracy=$\sum_{i=1}^{N} \frac{TP_i}{TP_i+FP_i}$
- Macro-average accuracy=$\sum_{i=1}^{N} \frac{\text{Precision}_i}{N}$

6.3 Evaluation of extracted-textual-feature relevance

The evaluation of the extracted-textual-feature relevance is based on their number w.r.t. composition levels and on the evaluation metrics as presented in Section 6.2. To evaluate our approaches compared to others, LIBSVM is used (Chang and Lin 2011) as an integrated approach to the Rapidminer software tool (KDnuggets 2012) installed on a personal computer with a 2.3 GHz CPU, 8 GB of RAM, and Windows 7.

6.3.1 Results for FW-Data

We start by presenting the level impact for the coverage percentage for FW-Data. For each level, we obtain a set of features whose efficiency is to be experimented. Furthermore, we present the evaluation of the approach of isolated N-composite points compared to other ones, as cited in the related work. In Table 6, we present the evaluation of the approach of isolated N-composite points for different levels and classes applied on the FW-Data. As we could remark, the composite properties have a positive impact on the precision–recall values for all classes. Hence, combining two words is very significant to reach acceptable classification accuracy. Regarding the total number of extracted features for all classes, we have (790) 1-composite and (1570) 2-composites covering 46.82% and 98.50% of FCT, respectively. Obviously, these values do not correspond necessarily to the summation of different features in the different classes, since the redundancies are eliminated. For level 3, only six additional features were obtained. Hence, for all classes, their total number is 2366. The management change class has the best values compared with the other classes, whereas the transaction class has the lowest values. The micro-average precisions are 67.08% and 83.05%

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 6. Evaluation for different levels w.r.t. FW-Data classes

Levels	Class	# Features	Precision (%)	Recall (%)	*F*-measure (%)
	M. change	1171	61.10	93.20	73.80
Level 1	Performance	622	61.20	36.70	45.90
	Transaction	584	75.60	32.30	45.30
	Average	792	65.96	54.06	54.98
	M. change	1823	85.33	94.70	89.77
Level 2-3	Performance	834	80.20	66.60	72.77
	Transaction	944	85.80	52.30	64.98
	Average	1201	83.77	71.20	75.84

Table 7. Results of different approaches for FW-Data

	Approach name	Features	Precision (%)	Recall (%)	F-measure (%)
BOW	Single of BOW	2537	87.71	21.97	35.11
	Single of Keywords	203	75.49	30.15	43.90
BOP	KEA approach	108,744	67.77	27.49	39.12
	Frequent term-Set	822	75.44	63.64	69.04
BOC	Difunctional approach	114	53.94	59.05	56.24
	FC as optimal factors	218	82.36	67.70	74.31
	MinGenCoverage	601	86.36	60.00	70.36
Our new approach	***N*-composites ($N = 2,3$)**	2366	83.77	71.20	75.84

for level 1 and levels 2 and 3, respectively. The macro-average precision for level 1 is 65.96% and 83.77% for levels 2 and 3.

In Table 7, we present the results of the evaluation of the approach of isolated *N*-composite points compared to the other ones as cited in Ferjani *et al.* (2012) and Elloumi *et al.* (2016) based on the precision/recall metrics. In this paper, we are particularly interested in comparing between conceptual approaches as we put forward a new version for extracting the isolated concepts.

In this study, we use Rapidminer (KDnuggets 2012) and the linear support vector machine (SVM) (Chang and Lin 2011) as a classifier with 10-folds in order to evaluate the different approaches. We calculate the precision, recall, and *F*-measure for the different approaches, as depicted in Table 7. It can be clearly seen that the precision varies from 53.94% to 87.71% (with a gap approximately equal to 33.77%) and the recall varies from 21.97% to 71.26% (with a gap around 49.29%) where the highest recall value is observed for the proposed new approach for N=2,3. The most striking feature is that (single of keywords) the keywords have been manually extracted by experts which does not give the highest precision 87.71%. In addition, the recall is very low (30.15%). This could be explained by the limited number of features suggested by the expert (218 labels are extracted) which are not as representatives as expected. Note that the recall, which describes what proportion of TDs that should be classified as **yes** and in fact classified as **yes** is very low in the BOW (21.97%). Regarding the conceptual approaches, the MinGenCoverage presents the best precision value (86.33%). In the new proposed approach, both precision/recall measurements are greater than 70.00%. Hence, the selected features composed of 790 single words and 1576 2–3-composite words are significant features w.r.t. the whole textual data.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 8. Characteristics of 20-newsgroup data set

Class	# Documents	# Words
alt_atheism	481	10,029
comp_graphics	582	7995
comp_os_ms-windows_misc	573	10,041
comp_sys_ibm_pc_hardware	588	11,710
comp_sys_mac_hardware	576	9082
comp_windows_x	593	13,647
misc_forsale	583	11,660
rec_autos	593	11,790
rec_motorcycles	597	11,426
rec_sport_baseball	595	13,563
rec_sport_hockey	599	11,784
sci_crypt	595	10,182
sci_electronics	592	9537
sci_med	595	8251
sci_space	594	7981
soc_religion_christian	600	8901
talk_politics_guns	546	9709
talk_politics_mideast	565	8230
talk_politics_misc	465	9891
talk_religion_misc	357	9198

6.3.2 Results for 20-newsgroup

We consider the (Matlab/Octave) version which is already a processed data set. It contains both training and testing sets. We have converted the training file to our TFS-algorithm input without considering the class attribute. It is a BR composed of 53,975 words and 11,269 objects. As results, we have 7382 1-isolated points covering 100% of the BR. This informs about the sparseness of the 20-newsgroup data set since the BR is completely covered at level 1. In Table 8, we present some characteristics of the 20-news data sets in terms of number of documents and number of words per document. We apply the *nltk.PlaintextCorpusReader* (Bird and Loper 2004) to extract the BOW for each class. Note that the requested time in average to extract the composite features is around 3 min per class. We remark that the number of 1-composites represents 4.69% of the entire number of words which could be interesting for the dimensionality reduction problem. However, the number of 2-composites goes up by more than 112%, which constitutes a drawback from the dimensionality reduction perspective. Nevertheless, it could be useful for constituting BOP where words are grouped w.r.t. the isolated composite aspect as a means of discovering more regularities in a corpus. Regarding the total number of extracted features for all classes we have 7382 distinct features for level 1 and 402,245 distinct ones for level 2.

As illustrated in Table 9, we can remark that the precision levels range from 42.9% to 100% where 9/20 classes have more than 84%. However, the recall values are low except for the basketball class. In average, the macro-average precision for level 1 is 77.20% and the micro-average precision is 66.35%. These results indicate that the isolated 1-composite points produce acceptable precision

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 9. Evaluation for different levels w.r.t. 20-newsgroup classes

	Class	# Features	Precision	Recall	*F*-measure
	alt_atheism	410	100	1.3	2.5
	comp_graphics	467	91.7	5.7	10.7
	comp_os_ms-windows_misc	471	76.9	2.6	5
	comp_sys_ibm_pc_hardware	512	55.6	21.4	30.9
	comp_sys_mac_hardware	475	94.1	12.5	22.1
	comp_windows_x	488	89.5	17.4	29.2
	misc_forsale	513	90.2	12.1	21.3
level1	rec_autos	496	79.7	36.7	50.3
	rec_motorcycles	494	72.1	41.6	52.7
	rec_sport_baseball	468	24.7	89.7	38.7
	rec_sport_hockey	470	84.5	40.9	55.1
	sci_crypt	474	77.4	26.9	39.9
	sci_electronics	513	73.5	25.4	37.8
	sci_med	497	85.5	18.1	29.8
	sci_space	494	86.6	21.4	34.4
	soc_religion_christian	518	53.5	48.2	50.7
	talk_politics_guns	462	92.9	3.6	6.9
	talk_politics_mideast	439	75	1	1.9
	talk_politics_misc	394	42.9	1.2	2.3
	talk_religion_misc	316	97.7	11.2	20
	Average	467	77.20%	21.94%	27.11%
	alt_atheism	6464	94.1	80	86.5
	comp_graphics	2776	73.2	71.3	72.2
	comp_os_ms-windows_mis	12,355	82	79.1	80.5
	comp_sys_ibm_pc_hardware	3385	73.1	73.7	73.4
	comp_sys_mac_hardware	5827	79.6	78.9	79.3
	comp_windows_x	7962	84.8	93.3	88.9
	misc_forsale	758	92.8	83.89	88.1
level2	rec_autos	26,152	81.9	83.3	82.6
	rec_motorcycles	20,522	87.5	87.5	87.5
	rec_sport_baseball	21,763	89.6	94.1	91.8
	rec_sport_hockey	69,627	96.6	96.6	96.6
	sci_crypt	107,600	95.6	93.1	94.3
	sci_electronics	4406	79.5	79.5	79.5
	sci_med	66,992	92.9	83.9	88.1

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Table 9. Continued

Class	# Features	Precision	Recall	*F*-measure
sci_space	10,379	92.6	93.4	93
soc_religion_christian	6369	82.9	89.9	86.3
talk_politics_guns	35,722	95	90.6	92.8
talk_politics_mideast	44,346	97.2	93.7	95.4
talk_politics_misc	26,061	94.6	87.9	91.1
talk_religion_misc	1204	83.1	73.1	77.8
Average	24,034	87.43%	85.33%	86.285%

Table 10. Results of different approaches for 20-newsgroup data set

Approach name	# Features	Micro-average accuracy (%)
Relevance frequency	>5000	80.81
Term frequency	>5000	80.83
TF.IDF	>5000	80.12
Relevance frequency	2000	68.79
Term frequency	2000	69.13
1-Composites	7382	69.13
2–3-Composites	402,245	85.29

levels but they are not discriminating the classes. For level 2, we got 87.43% for macro-average precision and 85.29% for micro-average precision.

In Table 10, we present the obtained results for different approaches, as cited in Lan, Tan, and Low (2006) in terms of micro-average accuracy. For relevance frequency and term frequency, we present two results related to the considered number of features (i.e., $>$ 5000 and 2000). This result indicates the positive impact of using a greater number of features as an increase accuracy from 69.13% up to 80.83%. The two last rows in Table 10 concern the results obtained by our new approach for levels 1 and 2. The number of features at level 2 is approximately 50 times that of level 1. This could be disadvantageous in terms of dimensionality aspect, but accuracy rises from 69.13% up to 85.29%. When we analyze the training set, we observe that many objects have a zero value for all features. Accordingly, the isolated 1-composite points do not exist for these objects. Moreover, some features are present in different objects even if they are not within the same class. These features are misleading and an additional filtering step could be considered to remove them. By considering the isolated 2-composite points, the accuracy is improved which witnesses the usefulness of considering the additional features during the learning stage. We are confident that it is worth studying many alternatives in property composition algorithm in order to extract more relevant features.

7. Conclusion

Textual data processing requires well-found approaches to derive relevant textual features. In this paper, we have put forward a new approach for TFS based on FCA background. Mainly, we have proposed to extract textual features by exploring the Riguet-difunctional relation. It allows discovering the regularities in an FCT as a set of isolated points. These regularities are also expressed

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

as logical implications between words within the same concept. As a main contribution, we have introduced the notion of isolated N-composite points as a set of N words to be considered as a unique textual feature. In addition, we have suggested some key elements for avoiding the composition of useless words in the TFS algorithm. From the experimental results, we have concluded that a reduced value of N (between 1 and 3) permits extracting significant textual features compared with existing approaches even for not completely covering the initial FCT. Nevertheless, we need to explore the relation between the coverage quality and the relevance of textual features. As for further work, we are confident that the proposed approach can contribute to analyzing large textual corpora more efficiently and to be extended for a multi-language aspect.

Author ORCID. S. Elloumi, 0000-0002-1822-5334

References

Agrawal R. and Batra M. (2013). A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN* **2**(6), 2231–2307.

Agrawal R., Imielinski T. and Swami A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, USA, pp. 207–216.

Bastide Y., Pasquier N., Taouil R., Lakhal L. and Stumme G. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. *Proceedings of the International Conference DOOD'2000*, LNCS, Springer-Verlag, pp. 972–986.

Belohlavek R. and Vychodil V. (2010). Discovery of optimal factors in binary data via a novel method of matrix decomposition. *Journal of Computer and System Sciences* **76**(1), 3–10.

Berend G. (2016). Exploiting extra-textual and linguistic information in keyphrase extraction. *Natural Language Engineering* **22**(1), 73–95.

Berger C. (2012). Big data analytics with oracle advanced analytics in-database option. *Oracle and/or its affiliates: Data Mining and Advanced Analytics*. Available at http://www.oracle.com/technetwork/database/options/advanced-analytics/oaa12cpreso-1964644.pdf. Last visited in November 2018.

Bernotas M., Karklius K. Laurutis R. and Asta Slotkien A. (2007). The peculiarities of the text document representation, using ontology and tagging-based clustering technique. *124x Information Technology and Control* **36**(2), 217–220.

Besanon R., De Chalendar G., Ferret O., Gara F., Mesnard O., Laïb M. and Semmar N. (2010). LIMA: *A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation*. *Proceedings of LREC 2010*, pp. 3697–3704.

Bird S. and Loper E. (2004). NLTK: the natural language toolkit. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, pp. 69–72.

Brank J., Grobelnik M., Frayling N. and Mladenic D. (2002). Interaction of Feature Selection Methods and Linear Classification Models. *Proceedings of the 19th Conference on Machine Learning (ICML-02), Workshop on Text Learning*.

Chang C. and Lin C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 1–27. Available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Dasgupta A., Drineas P., Harb B., Josifovski V. and Mahoney M. W. (2007). Inductive learning algorithms and representations for text categorization. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 230–239.

Dumais S., Platt J., Heckerman D. and Sahami M. (1998). Feature selection methods for text classification. *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management* **2**, 148–55.

Elloumi S., Boulifa B., Jaoua A., Saleh M., Al Otaibi J. and Frias M. (2014). Inference engine based on closure and join operators over truth table BRs. *Journal of Logical and Algebraic Methods in Programming* **83**(2), 180–193.

Elloumi S., Ferjani F. and Jaoua A. (2016). Using minimal generators for composite isolated point extraction and conceptual binary relation coverage: Application for extracting relevant textual features. *Information Sciences* **336**, 129–144.

Ferjani F., Elloumi S., Jaoua A., Ben Yahia S., Ismail S. and Ravan S. (2012). Formal context coverage based on isolated labels: An efficient solution for text feature extraction. *Information Sciences* **188**, 198–214.

Financial Keywords (2017). A Collection of Financial Keywords and Phrases. Software available at http://home.ubalt.edu/ntsbarsh/stat-data/keysphrasfinance.htm. Last visited in march 2017.

Ganter B. and Wille R. (1999). *Formal Concept Analysis*. Berlin: Springer-Verlag.

Garey M.R. and Johnson D.S. (1979). *Computers and Intractability : A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. San Francisco: W. H. Freeman and Company, p. 340.

Gharehchopogh F.S. (2010). Approach and review of user oriented interactive data mining. *IEEE the 4th International Conference on Application of Information and Communication Technologies, AICT2010*. IEEE, Tashkent, Uzbekistan.

Godin R., Missaoui R. and Alaoui H. (1995). Incremental concept formation algorithms based on Galois (concept) lattices. *Computational Intelligence* **11**(2), 246–267.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Gordon M.D. and Kochen M. (1998). Recall-precision trade-ok: a derivation. *Journal of the American Society for Information Science* **40**, 145–151.

Gosset W.S. (1908). Student. In *The Probable Error of a Mean*, Biometrikam, vol. 6, pp. 1–25.

Gupta V. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence* **1**, 60–76.

Harish B.S., Manjunath S. and Guru D.S. (2012). Text document classification: An approach based on indexing. *International Journal of Data Mining and Knowledge Management Process (IJDKP)* **2**, 43–62.

Hasan K. S. and Ng V. (2014). *Automatic Keyphrase Extraction: A Survey of the State of the Art*. ACL, vol. 1, pp. 1262–1273.

Irfan R., King C., Grages D., Ewen S., Khan S., Madani S. and Li H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review* **30**(2), 157–170. doi:10.1017/S0269888914000277

Jaoua A., AlJa'am J., Hammami H., Ferjani F., Laban F., Semmar N., Essafi H. and Elloumi S. (2010). *IEEE International Conference on Progress in Informatics and Computing (PIC)* **1**, 652–655.

Jaoua A., Elloumi S., Hasnah A., Jaam J. and Nafkha I. (2004). Discovering regularities in databases using canonical decomposition of binary relations. *Journal on Relational Methods in Computer Science (JoRMiCS)* **1**, 217–234.

Jones K.S. and Willet P. (1997) *Readings in Information Retrieval: Porter Stemmer*. San Francisco: Morgan Kaufmann. ISBN 1- 55860-454-4.

Kcherif R., Gammoudi M.M. and Jaoua A. (2000). Using difunctional relations in information organization. *Information Science* **125**, 53–166.

KDnuggets (2012). *Data Integration, Analytical* ETL, *Data Analysis, and Reporting*. Software available at http://sourceforge.net/projects/rapidminer/.

Kosala R. and Blockeel H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter* **2**, 1–15.

Lan M., Tan C.L. and Low H.B. (2006). Proposing a new term weighting scheme for text categorization. *AAAI* **6**, 763–768.

Ma W., Fang W., Wang G. and Liu J. (2007). Concept index for document retrieval with peer-to-peer network. IEEE Computer Society. *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, Qingdao, China, **3**, 1119–1123.

Manning C.D., Surdeanu M., Bauer J., Finkel J.R., Bethard S. and McClosky D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland USA, pp. 55–60.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*: 1301–3781.

Miller G.A. (1995). Wordnet: A lexical database for ENGLISH. *Communications of the ACM* **38**(11), 39–41.

Mouakher A. and Ben Yahia S. (2016). QualityCover: Efficient binary relation coverage guided by induced knowledge quality. *Information Sciences* **355**, 58–73.

Osei-Bryson K. (2010). Towards supporting expert evaluation of clustering results using a data mining process model. *Information Sciences* **180**, 414–431.

Pasquier N., Bastide Y., Taouil R. and Lakhal L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems Journal* **24**(1), 25–46.

Passalis N. and Tefas A. (2016). Entropy optimized feature-based bag-of-words representation for information retrieval. *IEEE Transactions on Knowledge and Data Engineering* **28**(7), 1664–1677.

Pennington J., Socher R. and Manning C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543.

Porter M.F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems* **14**(3), 130–137.

Porter M.F. (2006). Stemming algorithms for various European languages. Available at http://www.snowball.tartarus.org/texts/stemmersoverview.html. Last visited in November 2018.

Rajbhandari S. and Keizer J. (2012). The Agrovoc concept scheme-a walkthrough, food and agriculture organization of the united nations, Rome 00153, Italy. *Journal of Integrative Agriculture* **11**(5), 694–699.

Rennie J. (2008). The 20-newsgroups Dataset. Available at http://qwone.com/ jason/20Newsgroups/. Last visited in November 2018.

Riguet (1948). Relations binaires, fermetures et correspondances de Galois. *Bull. Soc. Math. France* **78**, 114–155.

Rodriguez-Esteban R. (2019). Text mining applications. In *Encyclopedia of Bioinformatics and Computational Biology*, Academic Press, pp. 996–1000.

Ronen F. and James S. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Co.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Stavrianou A., Andritsos P. and Nicoloyannis N. (2007). Overview and semantic issues of text mining. *ACM Sigmod Record* **36**(3), 22–34.

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

Talib R., Hanif M., Ayesha S. and Fatima F. (2016). Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications* **7**(11), 414–418.
Weiss S.M. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer-Verlag.
Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. (1999). KEA: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, 254–255.
WordNet (2017). Wordnet 3.0. http://wordnet.princeton.edu/wordnet/documentation. Last visited in march 2017.
Yang Y. and Pedersen J.O. (1997). A comparative study on feature selection in text categorization. *ICML* **97**, 412–420.

Cite this article: Elloumi S (2020). A new approach for textual feature selection based on N-composite isolated labels. *Natural Language Engineering* **26**, 221–243. https://doi.org/10.1017/S1351324919000160

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324919000160 Published online by Cambridge University Press