ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse71862019.pdf
https://doi.org/10.30534/ijatcse/2019/71862019

Formal Concept Analysis and Rough Set Theory for an Effective Representation of Arabic Short Text

Mohammed Bekkali[1], **Abdelmonaime Lachkar**[2]
[1]ENSA, USMBA, Fez, Morocco, bekkalimohammed@gmail.com
[2] ENSA, AEU, Tangier, Morocco, abdelmonaime_lachkar@yahoo.fr

ABSTRACT

With the increase use of mobile devices, such as smartphones and tablets, short text messages, tweets, comments and so on, have become a large portion of the online text data. Today, about one billion users interact daily on online social networks, where they share information and discuss a wide variety of topics. As in the rest of the world, users in Arab countries engage in social media applications. The representation of these short texts plays a vital role in language understanding. As a result, it may affect positively or negatively the performance of any Arabic text mining task. In this paper, we propose an efficient Arabic short text based on Formal Concept Analysis (FCA), which is a mathematical tool that offers conceptual data for knowledge representation, extraction and analysis with applications in different areas; combined with Rough Set Theory (RST) as a mathematical tool to deal with imprecise and vague data. The obtained results illustrate the interest of our contribution.

Key words: Arabic Language, Short Text, Rough Set Theory, Formal Concept Analysis, Formal Context, Concept Lattice.

1. INTRODUCTION

During the past decade and with the advent of the era of big data, a large number of online text data has been generated on the Web. The majority of this data is short text including (search snippets, micro-blog, products review, and tweets). These short texts are limited in length and different from traditional documents in their shortness and sparseness. Short text tend to be ambiguous without enough contextual information, and the degree of ambiguity is not the same for all languages. Since Arabic is a very high flexional language, where a single word can have multiple meaning, the short text representation becomes more and more difficult and there has been an increasing interest in language understanding of short text.

Two major approaches have been proposed to deal with the sparseness of short text. One aim to bring additional semantics from the dataset itself, this approach require Natural Language Processing techniques [14] [23]. The second try to enrich the representation of short text with additional information derived from existing large corpus or ontologies [16] [17] [20] [24] [25]. Both techniques have been shown a significant improvement by reducing the sparsity of the input data.

In this paper, we present an efficient representation to deal with Arabic short text based on Formal Concept Analysis (FCA) which is a mathematical tool mainly used for data analysis by identifying conceptual knowledge representation among data sets. FCA is developed based on formal context, which is usually defined as a binary relation between a set of objects and a set of attributes [21]. FCA and some derived structures have been successfully applied to the solution of problems from a wide range of tasks including clustering, categorization and information retrieval.

Recently, González and Hogan used FCA to compute data-driven schema for largescale heterogeneous knowledge graphs. First, they extract the sets of properties associated with individual entities; these property sets are annotated with cardinalities and used to induce a lattice based on set-containment relations, forming a natural hierarchical structure describing the knowledge graph. Then they propose an algebra over such schema lattices, which allows computing diffs between lattices [12]. Sahmoudi and Lachkar study how the FCA can be integrated and adapted as a new system for Arabic Web Search Results Clustering based on their hierarchical structure. In fact, browsing search results using a ranked for a specific request is time consuming. Web Search Results Clustering seems to be a good way to overcome this problem by groping similar documents in order to improve and facilitate browsing web pages in a more compact and thematic form [9]. Formica [1], d'Aquin, and Motta [13] used FAC for facilitating search and question answering applications over Semantic Web datasets. However, FCA is insufficient to process and analyze vague and imprecise data, such as short text. To overcome this problem, we propose to integrate the Rough Set Theory (RST) as a mathematical tool to deal with vagueness and uncertainty [19].

RST has been introduced by Pawlak in the early 1980s [19], it has been integrated in many Text Mining tasks. The central point of this theory is that each object in a Universe is described by a pair of ordinary objects called lower and upper approximations, determined by an equivalence relation in the Universe. The lower approximation is the set of elements that belong certainly to an object; and the upper approximation is the set of elements that may belong to an object. By using the

RST, we enrich the short text representation with other terms with which there is semantic links in the same dataset; and subsequently we enrich the formal context constructed by FCA. Both FCA and RST will be described in detail in the following sections of this paper. The effectiveness of our proposition has been evaluated and tested as clustering system where a serie of experiment has been conducted. The obtained results show the interest of our contribution.

The remainder parts of this paper are organized as follows : the next we introduce the mathematical background of RST and FCA ; in section 3 we present in detail our proposition ; section 4 conducts the experiments results ; finally, section 5 concludes this paper and presents future work and some perspectives.

2. PRELIMINARIES

2.1 Rough Set Theory

Rough Set Theory is a mathematical tool to imperfect knowledge. It has been originally proposed for data analysis and classification [10] [19]. It have been applied for a very wide variety of applications, such as Artificial Intelligence, features selection/extraction, machine learning and pattern recognition.

The basic concept of RST is the notion of approximation space: any subset in U (a non-empty set of object called the Universe) can be approximated by its lower and upper approximation. These approximations can be defined with reference to an indiscernibility relation R (R can be any relation reflexive, symmetric and transitive). Let x_1, x_2 be two objects from U, if x_1Rx_2 then we say that x_1 and x_2 are indiscernible from each other. The indiscernibility relation R induces a complete partition of universe U into equivalent classes $[x]_R$, $x \in U$ [9]. The lower and the upper approximation of any subset $X \subseteq U$ can be determined as:

$$L_R(X) = \{x \in U \mid [x]_R \subseteq X\} \quad (1)$$

$$U_R(X) = \{x \in U \mid [x]_R \cap X \neq \Phi\} \quad (2)$$

RST is dedicated to any data type but when it comes with text representation, we use its Tolerance Model described in detail as follow.

Let $D = \{d_1, d_2\dots, d_n\}$ be a set of documents and $T= \{t_1, t_2\dots, t_m\}$ set of index terms for D. Each document di is represented by a weight vector $\{w_{i1}, w_{i2}\dots, w_{im}\}$ where w_{ij} denotes the weight of index term j in document i. The tolerance space is defined over a Universe of all index terms $U = T = \{t_1, t_2\dots, t_m\}$ [11] [15].

Let $f_{di}(t_i)$ denotes the number of index terms t_i in document d_i; $f_D(t_i, t_j)$ denotes the number of documents in D in which both index terms t_i and t_j occurs. The uncertainty function I with regards to threshold θ is defined as:

$$I_\theta = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \text{ U } \{t_i\} \quad (3)$$

So $I_\theta(I_i)$ is the tolerance class of index term t_i. Thus, we can define the membership function μ for $I_i \in T$, $X \subseteq T$ as [15] [18]:

$$\mu_X(t_i, X) = v(I_\theta (t_i), X) = |I_\theta(t_i) \cap X| / |I_\theta (t_i)| \quad (4)$$

Finally, the lower and the upper approximation of any document $d_i \subseteq T$ can be defined as:

$$L_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) = 1\} \quad (5)$$

$$U_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) > 0\} \quad (6)$$

The detailed algorithm for generating the upper approximation can be summarized and presented as bellow (Algorithm 1).

Algorithm 1 : generateUpperApproximation

```
Input (A document list dl)
Begin
Initialise the Term Set s
Initialise the co-occurrence matrix
for each term t in s do
  for each term w in s do
    if t and w occur together > θ then
      add w to tolerance Class of t
    end if
  end for
end for
for each document d in dl do
  for each term t in s do
    tc ← tolerance Class of t
    cof ← |tc ∩ d| / |tc|
    if cof > 0 then
      add term t to upper approximation of d
    end if
  end for
end for
end
```

In this section, we have presented the RST as an efficient tool to deal the vagueness of short text. This theory will be used combined the FCA in order to give an effective representation for Arabic short text. The basics of FCA will be discussed in the following section.

2.2 Formal Concept Analysis

The basic idea of using FCA in knowledge representation and discovery is to generate the formal context between a set of documents and then construct the concept lattice as a new document's representation [21]. In this section, we present the FCA Theory by giving some important definitions and some illustrative examples.

2.2.1 Formal Context

Formal context (G, M, I) consists of a set of objects G, a set of attributes M, and I is defined by a binary relation between objects G and attributes M in a dataset that relates objects with values of the attributes. Table 1 shows an example of formal context.

Table 1: Example of formal context

Attributes Objects	**A**	**B**	**C**	**D**
Obj 1	.	X	.	.
Obj 2	X	.	X	X
Obj 3	.	X	X	.

2.2.2 Formal Concept of Formal Context

Formal concept of a formal context (G, M, I) is a set of objects that share similar characteristics. Using the mathematical definition given by Rudolf Wille [21], the formal concept is defined as a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A = B^I$ and $B = A^I$. A and B are called respectively the extent and the intent of the formal concept (A, B) [21], where:

$$A^I = \{m \in M \mid gIm \ \forall\ g \in A\} \qquad (7)$$

$$B^I = \{g \in G \mid gIm \ \forall\ g \in B\} \qquad (8)$$

A^I is the derivation operator of A and B^I is the derivation operator of B.

2.2.3 Concept Lattice

The concept lattice (G, M, I) is an ordered hierarchy of all formal concepts of the formal context (G, M, I). Many algorithms have been proposed to construct the concept lattice from the formal context. They may be divided in two categories: the first category, the proposed algorithms are developed to enhance the performance in generating the set of concepts such as [7], while in the second category, the used algorithms are developed to enhance performance in building the entire lattice such as [3] [8]. Figure 1 shows the concept lattice corresponding to the formal context presented in Table 1.

In the next section, we present our proposed method for enhancing the formal context representation by using the RST and therefore improving the corresponding concept lattice for a given Arabic short text dataset.

3. PROPOSED METHOD

In this section, we present our proposed method for short text conceptualization based on FCA and RST down to the smallest detail. The main steps are described by the flowchart presented in Figure 2. The following steps can summarize our proposed method:

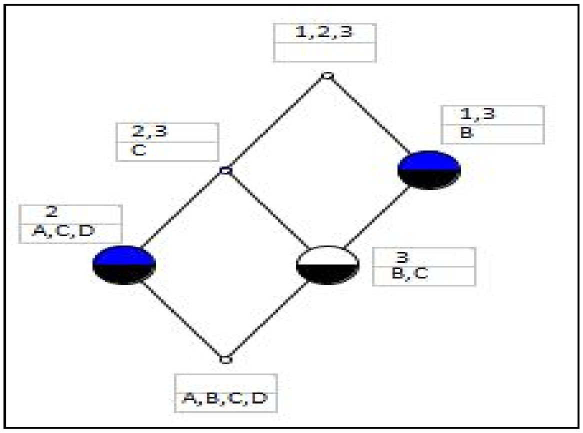


Figure 1: The concept lattice corresponding to the formal context presented in Table 1

Text Pre-processing

Each short text will be transformed to text brut, cleaned by removing Arabic stop words, Latin words and special characters like (/, #, $, etc…) and stemmed to find the corresponding stem. The experiments might lead us to confirm that many text-mining applications are based on the extraction of noun terms as a feature selection. In fact, the noun terms might be considered as the most descriptive terms for document content. We have used Al-Khalil morphologic Analysis System for Arabic Texts [4], which is included in the SAFAR platform for both stemming and noun terms extraction. Figure 3 presents an example of noun terms extraction.

Formal context construction

For the formal context construction, the obtained stems represent the set of attributes, short text represent the set of objects in the formal context and the binary relation defined as follows:
a. True “X”: if the word is part of the document.
b. False “.”: otherwise.

Table 2: Formal Context before applying RST

Attributes Documents	عربية	مسلسلات	وثائقية	أجنبية
1	X	.	X	X
2	X	.	.	X
3	.	X	.	.
4	X	X	.	X

The Rough Set Theory to enrich the Formal Context

The main objective of this step is to enrich the formal context by additional relation using the existing terms. In fact, our contribution is to enrich the formal context without adding new attributes or objects. Thus, the reason behind using the RST is to find a new relation between unrelated attributes and objects. The following steps can summarize the enrichment process:

— Calculate the frequency of each term in the document and in the whole corpus.
— Determine the tolerance class of all terms; for a given term, the tolerance class contains all the terms that occur with this term number of times upper than a threshold θ. This tolerance class is defined using the formula (3).
— Deduce the Upper Approximation of the document using the formula (6) then for each term a weight is calculated by replacing the original TF-IDF formula, which combines the definitions of term frequency and inverse document frequency by the following formula:

$$w_{ij} = \begin{cases} (1 + \ln(f_{di}(t_i))) * \ln\frac{N}{f_D(t_j)} & ; t_j \in d_i \\ \min t_k \in w_{ij} * \frac{\ln\left(\frac{N}{f_D(t_j)}\right)}{1+\ln\left(\frac{N}{f_D(t_j)}\right)} & ; t_j \in U_R(d_i) / d_i \\ 0 & ; t_j \notin U_R(d_i) \end{cases} \qquad (9)$$

where:

- w_{ij} is the weight of the term j in document d_i.
- $f_{di}(t_j)$ is the frequency of the term t_j in the document d_i.
- $f_D(t_j)$ is the total frequency of the term t_j in the whole corpus.
- $U_R(d_i)$ is the upper approximation of the document d_i.
- N is the number of the document in the corpus

This formula ensures that each term occurring in the upper approximation of di but not in di, has a weight smaller than the weight of any terms in di. Normalization by vector's length is applied to all weights of document vectors w_{ij} [18].

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_{t_k \in U_R(d_i)} (w_{ij})^2}} \qquad (10)$$

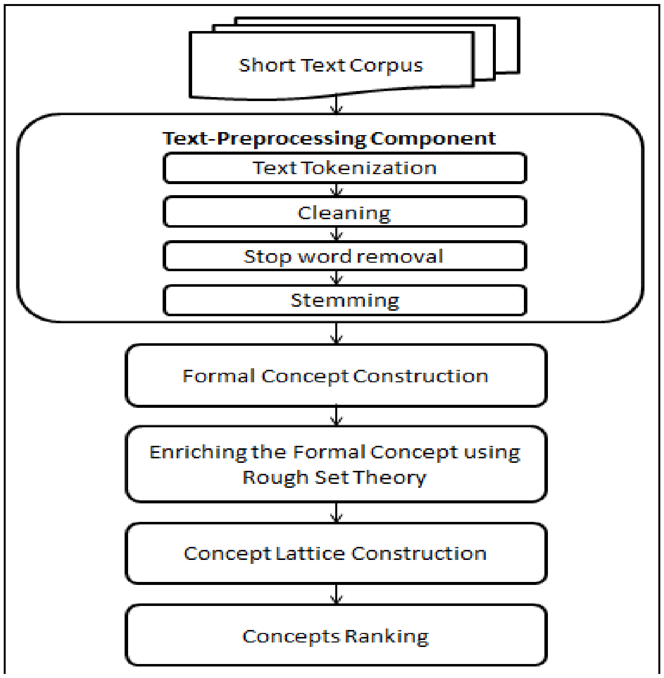
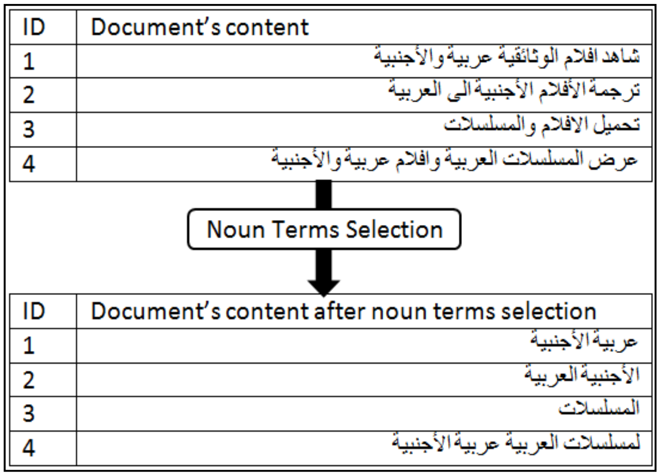

Short Text Corpus
Text-Preprocessing Component
Text Tokenization
Cleaning
Stop word removal
Stemming
Formal Concept Construction
Enriching the Formal Concept using Rough Set Theory
Concept Lattice Construction
Concepts Ranking

Figure 2: Flowchart of the proposed method



ID	Document's content
1	شاهد افلام الوثائقية عربية والأجنبية
2	ترجمة الأفلام الأجنبية الى العربية
3	تحميل الافلام و المسلسلات
4	عرض المسلسلات العربية وافلام عربية والأجنبية

Noun Terms Selection

ID	Document's content after noun terms selection
1	عربية الأجنبية
2	الأجنبية العربية
3	المسلسلات
4	لمسلسلات العربية عربية الأجنبية

Figure 3: Example of Noun Terms Selection

After we take back the same process for all others documents in the corpus, we have a new enriched context formal. Figure 4 presents an illustrative example of the use of rough set in order to enrich the context formal. The letter X with red color are new discovered relation using rough set theory.

Table 3: Formal Context after applying RST

Attributes / Documents	عربية	مسلسلات	وثائقية	أجنبية
1	X	X	X	X
2	X	X	X	X
3	X	X	X	X
4	X	X	X	X

Concept lattice Construction

The obtained formal context will be used to construct the concept lattice. Figure 5 shows the generated concept lattice before and after using Rough Set. In our case, we use the free Java API named ToscanaJ2, which uses Ganter's algorithm [2] to generate the set of Formal Concepts and the corresponding concept lattice.

This latter represents a set of concepts organized in hierarchical structure, each concept regroups a set of documents (Objects represented by Document's IDs in formal context rows) that represent the Extent sharing a set of terms (Attributes represented by terms in formal context columns) that represent the Intent.

Concept Ranking

Generally, the problem with the concept relevance ranking is to estimate the relevance of a concept. To overcome this problem, Zhang et al. have proposed a new method to construct the two reduced levels hierarchy from the concept lattice. This method is based on two mathematical measures [22]: the first one is the concept importance measure used to indicate how important the concept is. This measure is relevant to both the number of documents in the extent and the number of descendant concepts of this concept. The second measure is the concept similarity, which is based on Jaccard's similarity coefficient and it will be used in the merging process to construct a two level hierarchy. In this work, we propose our new concept relevance measure that takes into consideration the two following components:

- The number of documents in the Extent
- The weight of each word in the Intent

We define our proposed relevance S(C$_i$) as a measure of Concept C$_i$ as follows:

$$\text{Extent\_Weight} = |\text{Extent}(C_i)| \,/\, \text{Nbr\_Total\_Docs} \quad (11)$$

$$\text{Intent\_Weight} = \textstyle\sum(\text{WRST}(\text{Intent}(C_i)) \,/\, |\text{Intent}(C_i)| \quad (12)$$

Where:

- |Extent(C$_i$)|: The number of documents in the Extent.
- Nbr_Total_Docs: The total number of document in the Corpus.
- $\sum$(W$_{RST}$(Intent(C$_i$)) / |Intent(C$_i$)| : The Average of weight (based on RST) of all words of the Intent in the corresponding concept.0

To assess the performance of our proposed method a serie of experiments have been conducted; in the next section, we present the obtained results.

4. EXPERIMENTS AND RESULTS

In this section, we propose to present a comparative study to assess the performance of our proposed method for Arabic short text representation. To this end, we use our approach as clustering system. Note that it is well known that the concept represents a cluster when using FCA for a clustering process [5] [22]. The first dataset used in our experiments is a subset of the Open Directory Project (ODP) which is a searchable web-based multi-language directory consisting of few million web pages pre-classified and organized as tree. For Arabic language, the ODP includes 4781 document pre-classified into 459 categories by a group of human experts. Consequently, the ODP present to us a good ground truth for our comparative study.

The quality of results of any clustering system can be measured by the degree to which its ability to correctly reclassify a set of pre-classified document into exactly the same categories without knowing the original category assignment. It can be measured by two metrics: Normalized Mutual Information (NMI) and Normalized Complementary Entropy (NCE). These metrics are introduced by Geraci et al. to compare the effectiveness of different Web Search Result Clustering (WSRC) algorithms [6]. For a given a set S of N documents pre-classified under C= {c$_1$, c$_2$ ... c$_n$} of categories and a set C'= {c'$_1$, c'$_2$... c'$_m$} as the clustering results, the NMI and NCE are defined as follow:

$$NMI(C,C') = \frac{2}{\log|C||C\prime|}\sum_{c\in C}\sum_{c\prime\in C\prime} P(c,c')\log\frac{P(c,c\prime)}{P(c)P(c\prime)} \quad (13)$$

Where:

$$P(c) = \frac{|c|}{N}, P(c') = \frac{|c\prime|}{N}, P(c,c') = \frac{|c|\cap|c\prime|}{N} \quad (14)$$

And

$$NEC(C,C') = \sum_{i=1}^{m}\frac{|ci\prime|}{N\prime} NCE(C,ci') \quad (15)$$

Where:

$$NEC(C,ci') = 1 - \frac{2}{\log|C|}\sum_{j=1}^{n} - \frac{P(cj,ci\prime)}{P(cj)}\log\frac{P(cj,ci\prime)}{P(cj)} \quad (16)$$

$$N' = \sum_{i=1}^{m}|ci'| \quad (17)$$

NMI is designed for non-overlapping clustering, therefore higher values NMI means better clustering quality. NCE range in the interval [0, 1] and it is designed for considering overlap, greater value of NCE mean better clustering result. Zhang and Feng [22] show that these metrics suffer from many biases due to:

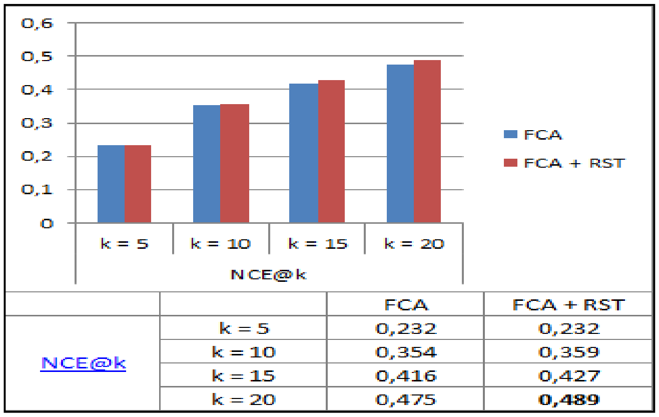
1. The value of NMI and NCE may become higher when the number of clusters generated becomes higher.
2. If the clusters that need to be compared are fixed, the more groups in the original categories, the higher value of the NMI obtained.
3. The performance of two different clustering algorithms may be influenced if using different original categories.

To overcome the above biases of the two metrics, they propose two improved metrics: A-NMI@K and A-NEC@K. A means the average and K means the number of clusters used in the experiment. In this comparative study, we set K as 5, 10, 15 and 20.

5. CONCLUSION

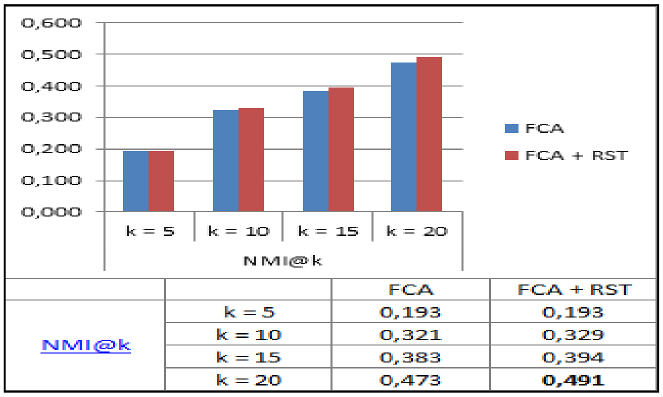
FCA has been successfully used as graphical organization and visualization tool for knowledge representation and processing in various fields. In this paper, we have suggested to use FCA as a representation model for Arabic short text by proposing two major contributions. On the one hand, we have proposed to select just noun terms as the most descriptive terms of a document. In the other hand, and in order to enhance semantics links between documents, we have used RST as a mathematical tool to deal with vagueness and imprecise data.

To illustrate the interest of our contribution, a series of experiments have been conducted; the obtained results were very encouraging and illustrate the efficiency of our proposed method. In future works, we believe that it could possible to improve the performance of our proposed method by integrating some external knowledge resources such as ontologies and or thesaurus.



		FCA	FCA + RST
NCE@k	k = 5	0,232	0,232
	k = 10	0,354	0,359
	k = 15	0,416	0,427
	k = 20	0,475	**0,489**

Figure 4: The obtained results for NCE@K



		FCA	FCA + RST
NMI@k	k = 5	0,193	0,193
	k = 10	0,321	0,329
	k = 15	0,383	0,394
	k = 20	0,473	**0,491**

Figure 5: The obtained results for NMI@K

REFERENCES

1. Anna Formica. 2012. Semantic Web search based on rough sets and Fuzzy Formal Concept Analysis. Knowl.-Based Syst. 26 (2012), 40–47. https://doi.org/10.1016/j. knosys.2011.06.018
2. B. Ganter, “Two basic algorithms in concept analysis,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5986 LNAI, pp. 312–340, 2010. https://doi.org/10.1007/978-3-642-11928-6_22
3. Bordat, J., 1986. Calcul pratique du treillis de Galois d’une correspondance. Math. Sci. Hum. Math. Soc. Sci. 96, 31–47
4. Boudchiche, M.; Mazroui, A.; Ould Abdallahi Ould Bebah, M.; Lakhouaja, A.; Boudlal, A.; 2017. “AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer”, Journal of King Saud University – Computer and Information Sciences. 29(2). pp. 141-146. https://doi.org/10.1016/j.jksuci.2016.05.002
5. C. Carpineto and G. Romano, “Exploiting the potential of concept lattices for information retrieval with CREDO,” J. Univers. Comput. Sci., vol. 10, no. 8, pp. 985–1013, 2004.
6. F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani, “Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution,” String Process. Inf. Retr., vol. 13, no. 10, pp. 25–36, 2006. https://doi.org/10.1007/11880561_3
7. Ganter, B., 2003. Ch1 & Ch2: Contexts, concepts, and concept lattices. Form. Concept Anal. Methods Appl. Comput. Sci.
8. Godin, R., Missaoui, R., Alaoui, H., 1995. Incremental concept formation algorithms based on Galois (concept) lattices. Comput. Intell. 11, 246–267. http://dx.doi.org/10.1111/j.1467-8640.1995. tb00031.x.
9. Issam Sahmoudi, Abdelmonaime Lachkar, “Formal Concept Analysis for Arabic Web Search Results Clustering”, Journal of King Saud University - Computer and Information, Volume 29 Issue 2, April 2017, pp 196-203 https://doi.org/10.1016/j.jksuci.2016.09.004
10. Jan Komorowski, Lech Polkowski, Andrzej Skowron “Rough Sets: A Tutorial”, 1998
11. Jin Zhang and Shuxuan Chen “A study on clustering algorithm of Web search results based on rough set”, Software Engineering and Service Science (ICSESS), 2013 https://doi.org/10.1109/ICSESS.2013.6615308
12. Larry González, Aidan Hogan, “Modelling Dynamics in Semantic Web Knowledge Graphs with Formal Concept Analysis”, WWW 2018, April 23-27, 2018, Lyon, France https://doi.org/10.1145/3178876.3186016
13. Mathieu d’Aquin and Enrico Motta. 2011. Extracting relevant questions to an RDF dataset using formal concept analysis. In International Conference on Knowledge Capture (K-CAP). ACM, 121–128. https://doi.org/10.1145/1999676.1999698
14. Mohammed Bekkali, Abdelmonaime Lachkar, “Arabic Tweets Categorization based on Rough Set Theory”, David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal – 2014, pp. 83–96, 2014. © CS & IT-CSCP 2014 https://doi.org/10.5121/csit.2014.41109
15. Mohammed Bekkali, Abdelmonaime Lachkar, “Web Search Engine-Based Representation for Arabic Tweets Categorization”, M. Kaya et al. (eds.), From Social Data Mining and Analysis to Prediction and Community Detection, Lecture Notes in Social Networks, 2017, DOI : 10.1007/978-3-319-51367-6_4
16. Bekkali M, Sahmoudi. I, Lachkar. A (2015) Enriching Arabic tweets representation based on web search engine and the rough set theory. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining, pp 1573–1574 https://doi.org/10.1145/2808797.2809339
17. Navigli R, Ponzetto S (2012) BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, Elsevier, pp217–250
18. Ngo Chi Lang, “A tolerance rough set approach to clustering web search results”, Poland: Warsaw University, 2003. https://doi.org/10.1007/978-3-540-30116-5_51
19. Pawlak, Z 1991. Rough sets: Theoretical aspects of reasoning about data, Kluwer Dordrecht.
20. Phan X-H, Nguyen L-M, Horiguchi S (2008). Learning to classify short and sparse text & web with hidden topics

from large-scale data collections. In: Proceedings of 17th international conference on World Wide Web, pp 91–100 https://doi.org/10.1145/1367497.1367510

21. Wille, R., 2005. Formal concept analysis as mathematical theory of concepts and concept hierarchies. Form. Concept Anal. 1–33. http://dx.doi.org/10.1007/11528784_1.
22. Y. Zhang and B. Feng, “Clustering search results based on formal concept analysis,” Information Technology Journal. pp. 746–753, 2008 https://doi.org/10.3923/itj.2008.746.753
23. R. Nandhakumar, Antony Selvadoss Thanamani. A Clustering Technique for Reducing Noise in High Dimensional Non-Linear Data Using M-DENCLUE Algorithm. International Journal of Advanced Trends in Computer Science and Engineering, 8(5),September - October 2019, 2414- 2417 https://doi.org/10.30534/ijatcse/2019/83852019
24. Sushil Kumar Trisal, Ajay Kaul. Dynamic Behavior Extraction from Social Interactions Using Machine Learning and Study of Over Fitting Problem. International Journal of Advanced Trends in Computer Science and Engineering, 8(5),September - October 2019, 2205 – 2214 https://doi.org/10.30534/ijatcse/2019/54852019
25. Sheetal S. Pandya, Nilesh B. Kalani. Review on Text Sequence Processing with use of different Deep Neural Network Model. International Journal of Advanced Trends in Computer Science and Engineering, 8(5),September - October 2019, 2224 – 2230 https://doi.org/10.30534/ijatcse/2019/56852019