ICTAI2020 - 32th International Conference on Tools with Artificial Intelligence, November 09-11, 2020, Virtual Conference.

An Unsupervised Approach for Precise Context Identification from Unstructured Text Documents

Maha Mallek
LIS UMR CNRS 7020 - LARIA, ENSI
Aix Marseille University - University of Manouba
Marseille, France - Manouba, Tunisia
maha.mallek@lis-lab.fr

Sébastien Fournier
LIS UMR CNRS 7020
Aix Marseille University
Marseille, France
sebastien.fournier@lis-lab.fr

Ramzi Guetari
LIMTIC laboratory, ISI
University of Tunis El Manar
Ariana, Tunisia
ramzi.guetari@isi-utm.tn

Bernard Espinasse
LIS UMR CNRS 7020
Aix Marseille University
Marseille, France
bernard.espinasse@lis-lab.fr

Wided Lejouad Chaari
LARIA, ENSI
University of Manouba
Manouba, Tunisia
wided.chaari@ensi-uma.tn

***Abstract*— The majority of the documents produced and exchanged through medias and social networks are unstructured. Due to the amount of these unstructured documents on the Web, their exploitation represents a tedious or even impossible task for human beings without assistance by dedicated algorithms and specialized computer systems in document classification or information extraction. To be efficient and relevant, such systems have to understand the content of these unstructured documents. The context (or topic) of a document is one of the basic information essential for the understanding of its content, and the more precise the context of a document, the more relevant its understanding will be. This paper presents a precise context identification approach that is evaluated quantitatively and qualitatively on several reference corpora and compared to other context identification systems. The contexts identified by our model are much more precise than those identified by these others systems.**

Keywords—accurate context extraction, unstructured textual document, text mining, semantic analysis.

I. INTRODUCTION

Nowadays, and thanks to the Web, the use of information technology has led to an explosion of the volume of data exchanged between individuals, companies, etc. The majority of the data exchanged are unstructured: text written in natural language, images, sound and video. These data usually convey vital information whose exploitation requires deep reading and analysis by the user. The challenge is to understand what these data represent and how to extract valuable information from them.

Extracting new valuable information is easy for a human mind as long as the volume of data remains reasonable. However, human capacity becomes powerless when it comes to dealing with huge amounts of data. The process is expensive in terms of time and resources; the error rate is high, and the data need to be constantly updated. The challenge is to make the information extraction process automatic, accurate and relevant.

Our research concerns unstructured textual data. Automatically processing these data is a task that can vary widely in complexity, and it mainly depends on the structure of the text in the documents. This task is easy only when the textual data is structured, but for unstructured ones, the work is more tedious. A document is characterized by metadata allowing better use of its content. Some metadata are automatically assigned by the operating system (size, creation date, last modification date) and others are fixed by the authors (document name, title, etc.). Sometimes the titles and contents of different documents are similar but their semantics differ significantly. The difference then lies in the context of each of the documents. The context is uncluttered content, extracted from the document itself and allowing better understanding and use of its content. It allows, among other things, to better classify and better help to find it using an information search system, thus improving search performance by significantly reducing noise.

Informally, we can say that the context captures "what a document is about" [1]. Indeed, the context can be a group of words containing thematic information. It can also be thought of as a title or header representing a particular block of text. In the same way, the context can characterize a collection of textual documents dealing with the same theme. Once we understand the subject or topic of the unstructured text, we can easily categorize the text based on its content and then help improve the extraction of valuable information from them.

Various tools and techniques have been developed to extract the context of unstructured documents. The different techniques include in particular keywords-based approaches [2] [3] [4], Topic modeling-based approaches [5] [6] [7], and approaches based on metadata [8] [9] [10].

Keyword-based approaches use frequencies and co-occurrences of keywords in the text to extract the subject and title of a document. Approaches based on subject modeling, which consist of a subject, aim to identify subjects by exploring and organizing the content of a textual document. In fact, they aim to group information into manageable clusters, where each cluster represents a single subject. Context extraction methods using metadata mainly use formatting information such as font size to extract the title of a document.

However, all of these approaches suffer from the same problem: they lack accuracy when identifying and extracting the context of a document, which dramatically degrades all uses related to the context such as the classification, information retrieval, understanding the content of a document, etc. let's consider as an example two different documents dealing with the theme of presidential elections. These are indeed two documents dealing with the same theme, but this theme remains vague and lacks details. The level of accuracy of such a subject must include the country, the year and, depending on the legislation and the country, can go as far as the round (as is the case in France). Therefore, these

documents may refer to two different presidential elections (for example, the 2016 US presidential elections and the second round of the 2017 French presidential elections). In order to improve the extraction of relevant information from the text, we seek to extract the most precise context possible describing a given document by analyzing the semantics using text mining techniques. The extraction of a precious and precise context is at the heart of our proposal.

The remainder of the paper is organized as follows: Section II presents related work on context extraction. Some of the fundamental concepts, and the proposed approach with its architecture and its main components are described in section III and IV. In Section V, our approach is evaluated quantitatively and qualitatively on several reference corpora and compared to other context identification systems. Finally, Section VI concludes this paper and outlines future work.

II. RELATED WORK

There is a large body of literature relevant to extract context from a document. In this section, we review the most recent and accurate works dealing with the document's context extraction problem. Three main approaches for context extraction were identified, including: Topic modelling-based approaches, Keywords-based approaches and metadata-based approaches.

A. Topic modeling based-approaches

Much research on topic modeling and keyword extraction-based approaches has been done. In the topic modeling approaches, the topic usually represents either a set of keywords describing the context or a single word that categorizes the topic of the text, for example (health, music, politics, etc…). In [11], authors have proposed an approach to extract the general theme of a document. For instance, if we consider the following two sentences:

"Emanuel Macron obtained 66.1% of the votes in the French presidential election of 2017" (s1)

"Barack Obama obtained 52.9% of the popular vote" (s2)

Based on the approach presented in [11], the two sentences above are categorized within the same predefined context "politics". However, the topic identified is not accurate enough to best characterize the real context of each of the two sentences. In fact, the first sentence concerns the "French presidential election of 2017" context, while the second one relates to the "American presidential election of 2012". It seems obvious to us that a single keyword is far from being able to characterize the context of a text concisely.

B. Keywords-based appraoches

The literature on keywords-based models shows a variety of approaches. Indeed, there are many works adopting this type of model in order to extract hot topic from the news and blogs. Recently, several authors [12], [13], [14] have proposed new models to find hot topics. In fact, they analyze news on the web and return words with the highest frequency during a predefined period of time. This is a very good approach except that it is dedicated to data from the web and microblogs only. Besides, the authors of [15] [16] have proposed an approach to automate the process of extracting topic and title from a single-document using key words-based techniques. However, the topic extracted may be irrelevant. Indeed, if we take as an example a document whose title is "Europe Gets Tough on Facebook", the result of extracting the topic is "Facebook Industry Company Data". This example shows that the set of words extracted from the document does not allow to construct a sentence or the sentence obtained is devoid of any meaning.

C. Metadata-based approaches

Great efforts were devoted to the study of the extraction of titles considering that they represent the main idea of a document. These researchers fall into two categories: Approaches for PDF documents and approaches for HTML web pages. These approaches are based on the style applied to the document (font size, alignment, margin, etc.) and some metadata to obtain this key phrase and ignore the semantics of the content.

1) Approaches for PDF documents: [17] and [18] have developed a simple rule-based heuristic, which considers style information (font size) to identify a PDF's title. To do this, they applied simple empirical rules reflecting the usual practices when presenting a text. Among the rules used, we can cite the most usual such as: "titles are usually located on the upper parts of the first pages", "Titles are usually in the largest font sizes", etc…

2) Approaches for HTML documents: These approaches are based on elements (tags) in the header and body of the document, etc. to extract the title. Among the most used tags in this sense we find for example: <Hn> (where $n \in \{1, 2, 3, 4, 5, 6\}$), <title>. In [19], authors have proposed a general scheme allowing to learn text titles according to style information.

These approaches are effective when structured documents (PDF, HTML) are available. They suffer from several shortcomings and inconsistencies. Indeed, metadata are generally entered by the authors of the documents and are therefore subjective. The styles and rules on which these methods of title extraction are based are not always reliable, especially since the styles can be modified by the authors. While PDF documents suffer from a lack of structure, the structure of HTML documents is also questionable and lacks reliability. Finally, these methods do not work at all when the title is not mentioned in the text.

The challenge of our work is to make the context extraction process automatic, accurate and relevant. Indeed, our objective is to extract a precise context that accurately describes a given document by performing semantic analysis using text mining techniques.

III. PROPOSED APPROACH OVERVIEW

In this section, a new approach for extracting the context of a document is presented. We first describe problem background. We then present an overview of the different components of our approach.

A. Definitions

We define preliminary concepts for our context extraction approach.

1) What is a topic? A topic is a particular subject that you discuss or write about. In our approach, a Topic "T" is defined as an important textual unit characterizing the document "D". Thus, a set of Topics associated to a document

"D" allows to understand its content and to determine its subject.

2) What is a document? A document is a written piece of text serving as information, as a proof, or as a testimony. With the emergence of new media, a document is defined as "an assembly formed by a medium and information, generally persistant, and as it can be read by humans and machines." In our model, a document "D" is a text written in natural language. It is defined by an identifier "Id", its important topics <Tdk=1..m> and "locator" such as an URI/URL. A document is formalized as:

$$D = < Id, Td_k=1..m, locator >$$

3) What is a context? The simple definition of context is the background information surrounding a subject. It is very important since it provides an overall idea of what a document is about. A context "ctx", in our approach, is the minimum information that characterizes a document "D". It is defined by a label "L" and a set of Topics "Tc". The label "L" is a nominal group or a keyphrase that provides a brief idea of the text content. A context is formalized as:

$$ctx = < Id_c, L, Tc_i=1..n >$$

B. The context extraction approach overview

In this section, we propose an unsupervised approach for context extraction. We have a database listing all the contexts that have already been determined. During the analysis of a new document, the deduced context can either be already known, in which case, the document is labeled by this context. If the context is not known, the database is enriched by this new context which is used to label the document processed.

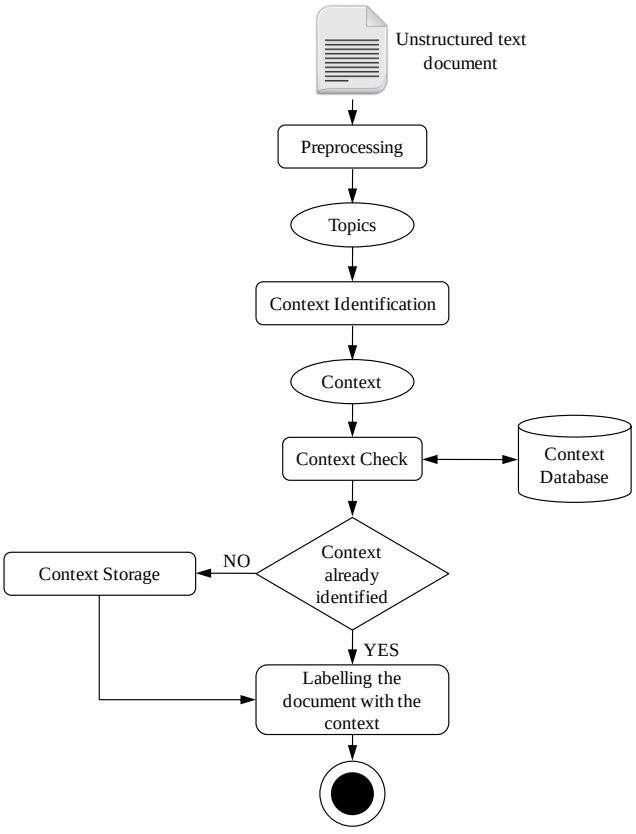

Figure 1: Overview of the components in our approach for context extraction.

The extraction process is illustrated in Figure 1 and is presented in five main stages: (1) preprocessing, (2) context identification, (3) searching for context in database, (4) labeling the document with the corresponding context and (5) Adding the new context in database.

1) Pre-processing: Preprocessing is the essential step in any text mining process. It is the first step that prepares the document in order to extract context. There are two main tasks:

a) The document cleaning phase where the stop words removal, the text segmentation and stemming are applied: Before extracting the different topics, all of the stop words, which are the common words without significant influence on the understanding and automatic processing of a text, are removed. The segmentation phase consists in dividing the text into sentences where each sentence is represented by a set of words (known as tokens). Finally, a stemming phase is applied to the remaining words.

b) Topics extraction: In this step, it is a matter of determining the most important words that can best characterize the content of the document. To do this, we have evaluated three different techniques: TF-IDF [20], TextRank [21] and 'Latent Dirichlet Allocation' LDA [22]. Experiments allowed us to use the LDA algorithm which achieved the best results compared to TF-IDF and TextRank. LDA allows to associate somehow a simple general context to a document from some of its words (known as topics in our approach). These topics with the predefined set of words can be used as factors to best describe the entire document.

2) Context identification: This process allows to extract the label "*L*" which covers what a document is about and associated it to the different topics. The context identification process is based on three main steps that are described in section IV.

3) Checking the identified context: The context produced by the process of our approach is used to label the document (to allow it to be classified). A similarity calculation is applied to the identified context and those of the Contexts' Database to determine if the context has already been identified before or if it is a context representing a new document theme. The Topics of the document as well as the Topics associated with each context of the Contexts' Database are transformed by semantic vectors using "universal TF-Hub integrations". The vector associated with the processed document is compared to each vector associated with a Contexts' Database context by applying the cosine similarity, described by formula (1). Let Di be the semantic vector associated to the document and Cj the semantic vector associated with a given context in database, the similarity between them is defined as follows:

$$\cos\alpha = \frac{\overrightarrow{D_i}.\overrightarrow{C_j}}{\|D_i\|.\|C_j\|} \quad (1)$$

4) Labelling the document by the corresponding context: When the context is well identified, an update of the database is performed by adding the locator of the document to the corresponding context. The database is composed of two tables: The first table contains the "Label" that accurate describes the context (e.g., U.S. Presidential Election of 2016) with the identifier "*ID*" of each context. The second table

contains a list of contexts identified by their identifier "*ID*", the set of Topics "*Tc*" that characterizes each context as well as the different locators of documents that are associated with each of the identified contexts.

5) Updating Contexts' database with a new Item: Once a document is processed and its context identified, two cases can occur: either the context is already listed in the context database or it is not. In the second case, the context database is updated by inserting the new item identified.

IV. UNSUPERVISED CONTEXT IDENTIFICATION PROCESS

Context identification is a fundamental step of the presented approach which follows the preprocessing of a document. After a pre-processing step, the context identification task requires three main steps as shown in figure 2: (A) Extraction of different candidate labels, (B) Extraction of the final label, and finally (C) Extraction of the final context.

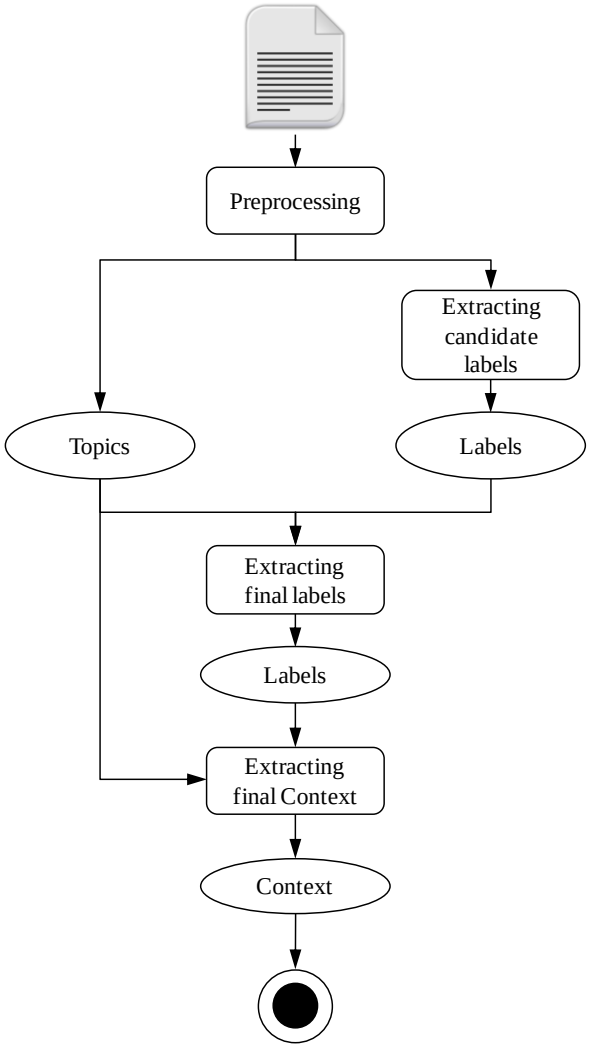

Figure 2: Unsupervised context identification process

A. Extracting different candidate labels

Candidate labels are derived from the units that frequently occur in a document. These units are in the form of unigram words, bigram words or trigram words [24]. To find these units from a text document we are adopting the FP-growth algorithm [25], which is an improvement of the APriori algorithm [26]. Indeed, FP-Growth is more scalable due to its linear running time.

B. Extracting the final label

Among the different candidate labels, we consider maximum five to identify the final context of the document. The principle consists in calculating a similarity rate between each candidate label with the different topics that have already been identified in the preprocessing phase. The five that have the highest similarity rate with the different topics are defined as the most important labels and can represent the minimal information that gives a brief idea of the text content.

The semantic similarity calculation was carried out by two different methods: the Jiang and Conrath similarity [27] and the Universal Sentence Encoder [28]. The result of this calculation is in the form of a matrix R [m, n] where m is the number of candidate labels, n is the number of Topics in the document and R[i, j] is the semantic similarity between the i candidate label and the j Topic. The performances of the two calculations are presented in the section V.

The Universal Sentence Encoder (USE) is used to calculate the semantic similarity between two words. The universal-sentence-encoder model is trained with a deep averaging network (DAN) encoder. It is based on transfer learning which is to exploit the knowledge of a neural model trained on a data set to apply and enrich it in the context of learning about a different set of data.

The five representative labels thus obtained are used for the specification of the final label of the document. We look for the minimal Key phrase which contains the maximum of these five labels, the latter which will be the final label of the document.

C. Extracting the final context

The process of identifying the context of a document is not limited to the key phrase identified. The obtained label may contain noise and irrelevant information for the context that needs to be removed. For this reason, while extracting the final label that gives a brief idea of the document content, two cases can be distinguished. If the extracted label is a nominal group, in this case, no treatment will be performed and the final label remains the same. Otherwise, when the extracted label is a key phrase, we need to perform a simplification process with removing irrelevant words and complex grammatical structures. In order to perform the latter, we use the measure of TF-IDF in order to remove nouns with low TF-IDF weight. For instance, if we consider as a label "British authorities accused three members of Ghana's Parliament and a former lawmaker of visa fraud", the TF-IDF weight gives the terms "members" and "lawmaker" a very low weight score. Therefore, the final label became "British authorities accused Ghana's Parliament of visa fraud".

Sometimes additional information needs to be added to improve the quality of the context identified. For example, Suppose that the identified label is "football world cup", the additional information is the year in which the world cup will take place. Annotating the label with temporal and spatial information can give more accuracy when identifying context. At the present time, temporal and the spatial tags are not considered in our application, but work was started to improve the process used. The association between the final label and the different topics of the document defines the new context (Label + Topics).

V. EXPERIMENT AND EVALUATION OF OUR MODEL

In order to prove the practical usefulness of our approach, the results of our research works have been implemented as an application allowing to extract context from document with greater accuracy and relevance compared with other research works. First, we briefly present the dataset and the metric used for experiments. Afterwards, we present quantitative evaluation of our proposed approach. Finally, we demonstrate qualitative results containing high-quality contexts identified by our model.

A. Dataset

To evaluate context extraction model, we consider three real-world datasets as a reference. One of them is created manually by collecting 600 documents from Wikipedia, which composed of 30 contexts. This choice is motivated by the fact that the prepared collection is based on full Wikipedia texts. The others were obtained from previous works in the literature such as BBC News and NewYork Times.

B. Evaluation metric

In order to evaluate the performance of our system, we adopt the official evaluation metric, which is based on macro-averaged F1-score. The F1 score can be interpreted as a weighted average of the precision and recall: *Precision* measures the system's ability to reject documents that are irrelevant to a query. *Recall* measures the system's ability to find all the relevant documents.

C. Quantitative evaluation

The proposed approaches for automatic context extraction are tested on the texts which are taken from Wikipedia. The output contexts extracted from the tool for each document is evaluated against reference contexts (original title provided by Wikipedia) of these documents to check whether the contexts that were extracted from the proposed approach are relevant when compared with actual contexts. The results of this evaluation are shown in Table I:

TABLE I. FINAL LABEL EXTRACTION PERFORMANCE IN TERMS OF F1-MESURE

Corpus	Algorithm used for similarity	Algorithm used for topics extraction	Precision (%)	Recall (%)	F1-score
Dataset from Wikipedia	Jiang and Conrath [27]	TF-IDF [20]	49%	65%	56%
		TextRank[21]	65%	73%	68%
		LDA [22]	68%	78%	73%
	Tenserflow Hub [28]	TF-IDF [20]	56%	70%	62%
		TextRank[21]	67%	71%	69%
		LDA [22]	**78%**	**91%**	**84%**

The best result is based on combining LDA, FPgrowth and Tenserflow hub, and it achieves 84%, which is higher than the use of Jiang and conrath similarity 73%.

D. Qualitative evaluation

In order to compare and analyze the results qualitatively with the existing method of context extraction that were available in [15] and [11], we carried out our experiments on New York Times and BBC News datasets. The context extraction results provided by our approach compared with the results obtained by [15] and [11], of the 5 randomly selected documents from New York Times and BBC News datasets respectively, are shown in tables II and III.

Our context extraction results show that they are content-rich and easy to understand compared with results obtained by [15]. The words associated with each context for our proposed approach are strongly semantically related than that of [15]. For instance, reference context 2 is "Apple Settles Legal Dispute with Nokia" and the context found by [15] is "Apple Nokia Company". This implies that the context obtained by [15] is devoid of any meaning. We notice that our model develops a far more well-understanding result "Apple and Nokia settled a legal dispute" which covers what a document is about.

TABLE II. RESULTS PROVIDED BY OUR APPROACH COMPARED WITH THE RESULTS OBTAINED BY [15] ON NEWYORK TIMES DATASET

Titles provided by New York Times	**Context extracted by [15]**	**Context extracted by our model**
1. 2015 Was Hottest Year in Historical Record, Scientists Say	Record Year Heat	The hottest year in the historical record
2. Apple Settles Legal Dispute With Nokia	Apple Nokia Company	Apple and Nokia settled a legal dispute
3. Atlantic Hurricane Season Is Expected to Be Busy	Storm Hurricane Season Forecast	Atlantic hurricane season
4. Britain Accuses Ghana Lawmakers of Visa Fraud	Visa Ghana Parliament Britain	British authorities accused Ghana's Parliament of visa fraud
5. Trump Will Withdraw U.S. From Paris Climate Agreement	Trump Agreement Paris President	the Paris climate accord

TABLE III. RESULTS PROVIDED BY OUR APPROACH COMPARED WITH THE RESULTS OBTAINED BY [11] ON BBC NEWS DATASET

Titles provided by BBC News	**Titles extracted by [11]**	**Titles extracted by our model**
1. India students caught 'cheating' in exams in Bihar	Society	Cheating in exams in the Indian state
2. Glasgow School of Art: 'One of the great buildings'	Art	Glasgow school of art
3. Martin Guptill hits highest World Cup score in New Zealand victory	Sport	Martin Guptill made the highest score in World Cup
4. Possible fatty acid detected on Mars	Science	A fatty acid discovered on Mars.
5. US winemakers reject arsenic claim	Society	California winemakers rejecting claims

Table III show the results on BBC News dataset. The article "India students caught 'cheating' in exams in Bihar" has the general context "Society", but the content of this article is about education in India and more accurate about cheating in exams in Bihar. Compared to the original context of the document, our model is able to extract a more detailed context "Cheating in exams is fairly common in the Indian state of Bihar". In conclusion, our evaluation shows that our approach

achieves a good performance while improving the quality of the context extracted compared with other systems.

In conclusion, our evaluation shows that our approach achieves a good performance while improving the quality of the context extracted compared with other systems. Adding semantic capabilities and more accurately contextualization to unstructured documents could be integrated to different applications as decision-making. Particularly, when dealing with search engines, the classification of documents according to a precious context improves information search effectiveness. Therefore, the relevant context should be clearly identified, and specified to obtain more performing and robust systems.

VI. CONCLUSION

Due to the amount of these unstructured documents on the Web, their exploitation represents a tedious or even impossible task for human beings without assistance by dedicated algorithms and specialized computer systems in document classification or information extraction. To be efficient and relevant, such systems have to understand the content of these unstructured documents. The context (or topic) of a document is one of the basic information essential for the understanding of its content, and the more precise the context of a document, the more relevant its understanding will be. In this paper we have presented a precise context identification approach based on a specific approach. This system has been evaluated quantitatively and qualitatively on several reference corpora and compared to other context identification systems.

Despite our approach encouraging results, there is still room for improvement: (i) the proposed method assume that a document is assigned to only one context. In our future work, we will focus our efforts on improving the efficiency of this method by considering that a document could be assigned to multiple contexts; (ii) Similarly, we now consider the only English language. It is important in our opinion to generalize this work for complex languages such as Chinese.; finally (iii) we also encountered some gaps when comparing the final label with the title provided by Wikipedia since we do not consider synonyms, but work was started to address these gaps and improve the process used.

REFERENCES

[1] M.R. Brett, "Topic Modeling: Basic Introduction," Available: http://journalofdigitalhumanities.org/2-1/topicmodeling-a-basic-introduction-by-megan-r-brett/.

[2] M. Yinghua, S. Guiyang, L. Jianhua and L. Shenghong, "A novel text subject extraction method," IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2003.

[3] F.Z. Lahlou, A. Mountassir, H. Benbrahim and Ismail Kassou, "A Text Classification based method for context extraction from online reviews," Intelligent Systems: Theories and Applications (SITA), 2013.

[4] Z. Wang, K. Hahn, Y. Kim, S. Song and J.M Seo, "A news-topic recommender system based on keywords extraction," Multimedia Tools and Applications, vol. 77, pp. 4339–4353, 2018.

[5] F. Viegas, W. Cunha and Ch. Gomes, "Semantically-Enhanced Topic Modeling," Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 893-902, 2018.

[6] X. Zhang and R. He, "Topic Extraction of Events on Social Media Using Reinforced Knowledge," International Conference on Knowledge Science, Engineering and Management, pp. 465-476, 2018.

[7] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou and J. Huang, "A Topic Detection Method Based on KeyGraph and Community Partition," Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, pp. 30-34, March 2018.

[8] Y. Hu, H. Li, Y. Cao, D. Meyerzon and Q. Zheng, "Automatic Extraction of Titles from General Documents using Machine Learning," Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, pp. 145-154, June 2005.

[9] Y. Wu, X.J. Zhang, Q. Li and J. Chen, "Title extraction from Loosely Structured Data Records," Machine Learning and Cybernetics, Vol. 5, 2008.

[10] S. Changuel, N. Labroche and B. Bouchon-Meunier, "A General Learning Method for Automatic Title Extraction from HTML Pages," International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 704-718.

[11] N. Guo, Y. He ; C. Yan ; L. Liu and C. Wang, "Multi-Level Topical Text Categorization with Wikipedia," IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC), Shanghai, China 2016

[12] Y. Jahnavi and R. Yalavarthi, "Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents," 15th International Conference on Advanced Computing Technologies (ICACT), 2013.

[13] H. Ma, "Hot topic extraction using time window," International Conference on Machine Learning and Cybernetics, China, July 10-13, 2011.

[14] Y.P. Zhang and H. Zhang, "Social Topic Detection for Web Forum," International Conference on Computer Science and Service System, 2012.

[15] A. Sajid, S. Jan and I.A. Shah, "Automatic Topic Modeling for Single Document Short Texts," International Conference on Frontiers of Information Technology (FIT), 2017.

[16] J. Yun, L. Jing and Y. Zhang, "Document Topic Extraction Based on Wikipedia Category," Fourth International Joint Conference on Computational Sciences and Optimization, 2011.

[17] J. Beel, B. Gipp, A. Shaker and N. Friedrich, "Extracting Titles from Scientific PDF Documents by Analyzing Style Information," International Conference on Theory and Practice of Digital Libraries, pp 413-416.

[18] J. Beel, S. Langer, M. Genzmehr and C. Mueller, "Docear's PDF inspector: Title extraction from PDF files," Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries.

[19] S. Gupta and K.K Bhatia, "Domain Identification and Classification of Web Pages Using Artificial Neural Network," International Conference on Advances in Computing, Communication and Control,

[20] G. Salton G and C. Buckley, "Term-weighing approaches in automatic text retrieval," In Information Processing & Management, pp. 513-523, 1988

[21] R. Mihalcea R and P. Tarau, "TextRank: Bringing Order into Texts," Conference on Empirical Methods in Natural Language Processing, 2004

[22] X. Hu, "News Hotspots Detection and Tracking Based on LDA Topic Model," International Conference on Progress in Informatics and Computing, pp. 248-252, 2016.

[23] L. Nguyen, "Some Novel Algorithms for Global Optimization and Relevant Subjects," Applied and Computational Mathematics, Special Issue, Some Novel Algorithms for Global Optimization and Relevant Subjects, 2017

[24] KA. Dhand, JS. Umale and PA. Kulkarni, "Context Based Text Document Sharing System Using Association Rule Mining," Annual IEEE India Conference (INDICON), Pune, India, pp. 11-13, Dec 2014.

[25] C. Borgelt, "An Implementation of the FP-growth Algorithm," the 1st international workshop on open source data mining, Chicago, Illinois, pp. 21 - 21, 2005.

[26] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules," In VLDB'94, pp. 487-499.

[27] J. Jiang, and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," International Conference Research on Computational Linguistics (ROCLING X), pp. 19–33, 1997.

[28] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Abrego , S. Yuan, Ch.Tar and R. Kurzweil, "Multilingual Universal Sentence Encoder for Semantic Retrieval," July 2019.