Learn2Sum

Amal Beldi, Salma Sassi, Abedrazzek Jemai

Learn2Sum: A new approach to unsupervised text summarization based on topic modeling

Amal Beldi
Faculty of Mathematical Physical and Natural Sciences of Tunis, SERCOM Laboratory,
Tunis, Tunisia
amal.beldi@fst.utm.tn

Salma Sassi
FSJEGJ, University of Jendouba, VPNC Lab
Jendouba, Tunisia
salma.sassi@fsjegj.rnu.tn

Abedrazzek Jemai
Carthage University, Polytechnic School of Tunisia, SERCOM Laboratory, INSAT
Tunis, Tunisia
abderrazekjemai@yahoo.co.uk

ABSTRACT

Due to the enormous volume of data on the web, it is hard for the user to retrieve effective and useful information within the right time. Thus, it has become a need to generate a brief summary from a large amount of textual data according to the user profile. In this context, text summarization is used to identify important information within text documents. It aims to generate shorter versions of the source text, by including only the relevant and salient information. In recent years, the research on summarization techniques based on topic modeling techniques has become a hot topic among researchers thanks to their ability to classify, understand a large text corpora and extract important topics on the text. However, existing studies do not provide the support of personalization when generating summaries because they need to know not only which documents are most helpful to the users, but also which topics and keywords are more or less related to the user' interests. Thus, existing studies lack of the support of adaptive user modeling for user applications in the emerging areas of automatic summarization, topic modeling and visualization. In this context, we propose a new approach of automated text summarization based on topic modeling techniques and taking into account the user's profile which helps to semantically extract relevant topics of textual documents, summarizing information according to the user' topics interests and finally visualize them through a hyper-graph Experiments have been conducted to measure the effectiveness of our solution compared to existing summarizing approaches based on text content. The results show the superiority of our approach.

CCS CONCEPTS

• **Information systems → Summarization, topic modeling**.

KEYWORDS

Text transformation, classification, summarization, topic modeling, topics, user profile, graph

MEDES '22, October 19–21, 2022, Venice, Italy

ACM ISBN 978-1-4503-9219-8/22/10…$15.00
https://doi.org/10.1145/3508397.3564853

ACM Reference Format:
Amal Beldi, Salma Sassi, and Abedrazzek Jemai. 2022. Learn2Sum: A new approach to unsupervised text summarization based on topic modeling. In *International Conference on Management of Digital EcoSystems (MEDES '22), October 19–21, 2022, Venice, Italy*. ACM, New York, NY, USA, Article 39, 10 pages. https://doi.org/10.1145/3508397.3564853

1 INTRODUCTION

In recent years, text summarization has gained prime importance and the research in this area has become a hot topic among researchers. It is a core area of study under Natural Language Processing (NLP) and Computational Linguistics (CL) to generate coherent text summaries. Summarization is defined as the method of identifying the key topics and information from one or more document. So, it aims to reduce the content and size of the text to its important keywords. In this context, automatic text summarization [26] is also used to this aim, and there are mainly two approaches to generate automatic text summaries such as extractive and obstructive methods. Extractive summary technique [10] uses statistical methods and consists of producing summary based on the key features of the given text. Whereas Abstractive summary [18] uses linguistic methods to examine and interpret the text. Most of the current automated text summarization system use extractive methods to produce summary. For large scale of data, these methods are poorly applicable and they are sensitive to the problem of irrelevancy. Hence, the performance of these methods may be limited to single document summarization and the use of statistical techniques. So, there is a need of methods and tools to organize, analyze, search and discover the hidden insights in any large group of summary textual data. These methods are called topic modeling, which are a new powerful technique for automatic classification of documents, unsupervised analysis of big and variant document groups. They facilitate the understanding of vast quantities of information with any large group of unstructured textual data and allowing summarizing large collections of textual information [2]. Among topic modeling techniques, we cite Latent Semantic Analysis (LSA) [14], Probabilistic Latent Semantic Analysis (PLSA) [33], Latent Dirichlet Allocation[11] (LDA) [16], Hierarchical LDA (hLDA) [35] and the Correlated Topic Model (CTM) [1]. These techniques may improve the text summarizing system by the use of automatic topic extraction to speed up and improve the quality of search, reduce the problems caused by natural language and extract important topics to reduce textual corpus size. So, using topic modeling techniques to summarize textual documents should be able to greatly improve the quality of the topic selection. Many approaches [37] [38] [39]

are proposed in order to extract a set of understandable terms and to classify generate summaries according to topics. However, the major problems of existing automatic summarizing studies consist their disability to reflect the topic diversity of textual source, so as to ensure the summary quality. Also, existing approaches still enable to reduce the redundancy rate of a summary, so as to ensure the compression ratio. Finally, none of existing work allows to integrate the user profile when generating a topic base summary.
In order to overcome these challenges, we propose a new automatic summarizing approach based on topic modeling where a summary is created from a set of related documents and optionally satisfies a specific information need of a user. Our approach named Learn2Sum aims to summarize a large text corpora based on the generated topics as well as the user needs.

Our approach was represented through a labled graph which helps to improve the summarization process accuracy and make the summaries more accurate and efficient.
To achieve our objectives, two main challenges have been addressed in our study:

- **Challenge1**: How to automatically extract topics, identify relationships between them in order to reorganize the summary according to the extracted topics?
- **Challenge2**: How to represent extracted data/terminologies toward a graphical model based on user profile?

The rest of this paper is organized as follows: we provide in Section 2 our motivation and some background. Next, we study in section 3 the related works and we illustrate a comparative study between existing approaches. Section 4, explains the methodology of our approach which consist of summarising a text corpora according to their dominant topics using H-LDA and the user profile model. In section 5, we describe the experiments conducted to validate our approach. Finally, section 6 concludes the paper and discusses some future work.

2 TOPIC MODELING TECHNIQUES

Topic modeling plays a vital role in the field of text summarization. The main idea is to consider a textual document as a set of topics [16]. Topic modeling techniques identify the topics in the document. These topics are then used to generate text clusters. The clusters include salient sentences from the source document. Each cluster would be labeled to the relevant identified topics. There are various topic modeling techniques such as Latent semantic Analysis (LSA)[32] which is the earliest attempt of topic modeling, although there is no explicit topic concept in LSA. Probabilistic Latent Semantic Analysis (PLSA) [33] is a proper probabilistic generative model in which each document is a mixture of topics, and each topic is a distribution of vocabulary. Blei et al.[31] propose a technique Similar to PLSA, named Latent Dirichlet Allocation (LDA) excepting that topic parameters in LDA are assumed to have Dirichlet priors, which makes LDA more effective. Since then, the researchers have proposed various models based on LDA. Also another technique named Dynamic Topic Model (DTM)[34] is introduced to obtain the evolution of topics over time in a sequentially organized corpus. Finally, Correlated Topic Model (CTM)[1] can represent pairwise topic correlations. The most existing works for summarization based on topic modeling used LDA since it has many characteristics that excels at feature reduction. It is also employed as a preprocessing step for other models, such as machine learning algorithms[31]. It can also be used to augment the inputs to machine learning and clustering algorithms by producing additional features from documents. However, LDA has many limits, it has a shortcoming that it cannot deal with various changes of data set well, which has become a limitation for its applications. In this context, HLDA[12] is a generalization of LDA and it can adapt itself to the growing data set automatically. HLDA can mine latent topics from a large amount of discrete data and organize these topics into a hierarchy, in which the topics of higher level are more abstractive while the topics of lower level are more specific. This hierarchy could achieve a deeper semantic model which is similar to a human mind.

3 RELATED WORK

3.1 Summarization approaches based on topic modeling

Presently, there have been a number of studies related to automatic summarization using extractive technique [10] ,[23], abstractive technique [24],[18] and hybrid technique. However, there are few studies related to text summarization based on graph [19], [20]. Topic modeling is used to divide the document into topics-words clusters and enabling researchers to understand the statistical relationships among topics. Here, we review the text summarization works based topic modeling techniques. [4] proposes a heuristic method which uses the LDA technique to identify the optimum number of independent topics present in the corpus. Some of the sentences are identified as the important sentences from each independent topic using a set of word and sentence level features. The authors of [5] reduced the semantic redundancy using LSA and agglomerative hierarchical clustering followed by document dimension reduction by selecting highly weighted sentences. Na et al.[6] mixed the topics of title and content of the document into a new topic in which the summary generation algorithm learns information entropy based weights in an adaptive asymmetric manner. In 2017,[7] proposed a merged approach of hierarchical topic modeling and the Minimal Description Length (MDL) principle. The authors presented the former used to describe topics while later generated news articles summary. In [8], authors proposed a topic modeling based approach to extractive automatic summarization, so as to achieve a good balance between compression ratio, summarization quality and machine readability.They extracted sentences associated with topic words from a preprocessed novel document. Second, they design an evaluation function to select the most important sentences from the candidate sentences and thus generate an initial novel summary. In [9], a document is represented as a combination of topics, and each topic is a probability distribution over words. [10] performed sentiment analysis from the students' comments toward a university, in this case the Universitas Diponegoro, using LDA and topic polarity word cloud visualization.The purpose of this study was to generate the topic polarity word cloud of the students' comments by using the best combination of parameters. 08 studied, implemented and analyzed the most suitable techniques in the case of sentiment analysis for tourism review in Indonesia. In this study, the authors proposed an unsupervised technique using

probabilistic topic models to classify online review based on the sentiment behind those reviews. Haung et al in [12] have extended the idea of topic modeling to multilingual text summarization. In this row, hierarchical Latent Dirichlet Allocation (hLDA) used semantic information in combination with other features for sentence scoring to generate an effective and robust summary. [13] generated summarization based on LSA and Maximum Marginal Relevance, where the weighted variation strongly determines the accuracy of resulting summary. The purpose of the approach [14] was to summarize the documents in Bahasa (Indonesian Language). It aimed to satisfy the user's need of relevant and consistent summaries. The algorithm is based on sentence features scoring by using LDA and Genetic Algorithm for determining sentence feature weights. In [15], authors used the topic model to identify topics in the input text represented as word distributions. A word distribution represents a topic by appointing high probabilities to words that portray a topic. Also, there are some approaches defined for automatic extractive summarization that can visualize the summary into a graph. The fundamental idea behind the extractive strategy of the text summarization is to discover the importance of the sentences so that the best sentences for the summary can be identified. In [19], authors generated summaries by selecting a subset of the sentences from the original document that emphasized various extractive approaches for single and multi-document summarization. They described some of the most extensively used methods such as topic representation approaches, frequency-driven methods, graph-based and machine learning techniques. [26] proposed the Receivables Management (RM) and LSA based techniques to generate the extractive summary of the input document. The similarity between each sentence and the overall document is calculated with the help of statistical methods, and sentences are ranked on the basis of the similarity. Lim et al. In [20] proposed a specific nonparametric Bayesian topic model for modelling text from social media. The authors focused on posts on Twitter). Cuong et al in [21] investigated the benefits of dropout for preventing topic models from overfitting. They integrated dropout into several stochastic methods for learning, latent Dirichlet allocation. Amplayo and Song in [22] proposed a new approach divided into two parts: sentiment classification and aspect extraction. The first part consists of the building of a three-level sentiment classifier using natural language processing techniques, specific lexical and syntactical techniques.The second level classifiers are two support vector machines classifiers, handling different n-gram feature vectors from different dictionaries. Zhang et al in [23] analysed the sentiment trends over a long period and their relation to announced the news, and the comparison of the human behavior in two different geographical locations affected by this pandemic. Barros et al [24] created a cross-document timeline where a time point contains all the event mentions that refer to the same event. Fu et al.[25] proposed a Variational Hierarchical Model (VHTM) that address summarization with topic inference via encoder-decoder. VHTM is the first attempt to jointly accomplish summarization with topic inference via variational encoder-decoder and merge topics into multi-grained levels through topic embedding and attention.

3.2 Discussion and Limitations

In order to compare the existing approaches and to overcome the challenges described previously, we define here 7 criteria with respect to the defined challenges :

Challenge 1: How to extract topics, identify relationships between them in order to reorganize the summary according to the extracted topics?

- **Criterion 1 (C1): Input type** this criterion refers to the input data which could be : Single Document, Multi-document.
- **Criterion 2 (C2): Granularity** : this criterion describes if the approach based on simple topic or composed topic.
- **Criterion 3 (C3): Topic relationship** this criterion indicates if the approach takes in consideration the relation between the topic (e.g. Yes or No).
- **Criterion 4 (C4): Summarization method** this criterion refers to the techniques deployed to summarize textual document which could be: extractive, abstractive or hybrid.
- **Criterion 5 (C5): User oriented summarization** this criterion indicates of the approach is oriented user (e.g., Yes or No).

Challenge 2: How to represent extracted data/terminologies toward a graphical model based on user profile?

- **Criterion 6 (C6): Output type** this criterion indicates the type of displayed summarized textual data which is a combination of: text summary, sentence summaries, topics or concepts.
- **Criterion 7 (C7): Application domain** this criterion indicates if the system is dedicated to a general or specific domain.

Our comparative study shown that the evolution of the text summary is still an open challenge. The table 1 has shown that some existing studies rely only on single document [24] ,[25]. Others studies are able to summarize many documents [22], [23], [21]. For the second criterion, we note that all existing works extract only simple words. The third criterion is shown that none of existing studies is a user oriented summary. So, no approach integrates the user profile to estimate the correlation in topics generation. For the fourth criterion, we note that none of the existing studies is able to generate relationship between topics and none of them identifies concepts from topics. The fifth criterion indicates that most of the existing studies are generic domain [24],[21] and [12] are specific domain. The last criterion gives an idea about the summary output. All existing studies generate text information except [7] ,[15] generating topics and the work of [25] generating a graph summary. The main contribution of this study relies on summarizing a large corpus of textual data into graph model. It's an automatic approach based on hLDA algorithm and integrating user profile in the summarization process. Our approach consists also of linking the topic graph model of the relevant domain knowledge in order to find relevant concepts and provide meaningful and concise summary. After extracting relevant concepts, we provide a personalized visualization model according to the user preferences.

	Challenge 1					Challenge 2	
approach	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
K Garcia et L Berton 2021	Document	Simple	No	Analysis	Tweet	No	Text
Fu et al 2020	Document	Simple	No	Hierarchical topic-Aware	general	No	Text
Barros et al 2019	Document	Not mentioned	No	Abstractive	General	No	Text
Rajendra et al 2019	Multi Document	Not mentioned	No	Extractive	General	No	Text
Cuong et al 2019	Document	Not mentioned	No	Probabilistic	General	No	Text
Zhang et al 2018	Document	Simple	No	Topic Analysis	Chinese tweets	No	Text
Hafeez et al 2018	Multi Document	Simple	No	Extractive	Specific	No	Text
Shiva et al 2018	Document	Simple	No	Extractive	General	No	Text
Litvak et al 2017	Multi Document	Simple	No	Extractive	General	No	Text
Amplayo et al 2017	Document	Simple	No	Extractive	General	No	Text
Wu et al 2017	Multi Document	Simple	No	Extractive	Specific	No	Text
Singh et al. 2017	Multi Document	Simple	No	Extractive	General	No	Text
A. Bashri et al 2017	Document	Not mentioned	No	Extractive	Specific	No	Text
Putri et al 2017	Multi Document	Simple	No	Extractive	Specific	No	Text
Huang et al 2016	Multi Document	Simple	No	Extractive	Specific	No	Text
Lim et al. 2016	Document	Not mentioned	No	Baysien Hierarchical	General	No	Topics
Irawan et al 2016	Document	Not mentioned	No	Extractive	General	No	Text
Silvia et al 2014	Document	Simple	No	Extractive	Specific	No	Text
Kim et al .2012	Document	Simple	No	Extractive	Specific	No	Topics

Table 1: Qualitative Comparison of text Summarization Approaches based on topic modeling

4 METHODOLOGY

Our hLDA-based summarization learning method, called Learn2Sum, uses text information as the data source and the domain knowledge to automatically generate meaningful and concise summary concepts and inter-relationships taking into account the user's profile. Lear2Sum aims mainly at (i) generating topics and topic hierarchies and (ii) matching the resulted hierarchies with the user profile model to estimate the correlation between a given query and the user search. Learn2Sum framework is shown in Figure 1. The proposed architecture consists of four phases : (1) Text preprocessing, (2) Topic extraction, (3) Building graph model, (4) Modeling user profile and (5) Visualizing topic graph. We detail them in what follows.

4.1 Pre-Processing phase

Preprocessing is an important task and critical step in NLP, it consists of transferring text from human language to machine-readable format and it affects substantially the results of the experiments. The preprocessing stage is important to structure the unstructured text and keep the keywords which are useful to represent the category of text topics. Natural language text can contain many words with no specific meaning, such as prepositions, pronouns, etc. So,

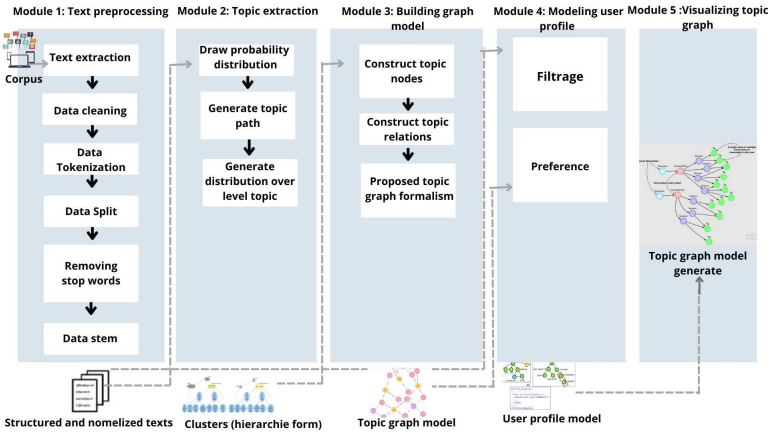

Figure 1: Architecture of our proposed system

after a text is obtained the preprocessing process consists of four steps : **(i) Data cleaning** consisting of shrinking the size of the vocabulary by converting the characters to lowercase, deleting numbers, symbols and removing punctuation. It involves transforming

raw data into an understandable format. **(ii) Data tokenization**, in this step, we tokenize each word in the source code to remove some numbers and punctuation marks.**(iii) Data Split**, we was split the given text into sentences and each sentence were split into tokens (Words). **(iv) Removing stop words** : we remove common English language stop words ("in, it, for") and key words ("int, return") to reduce the noise and **(v) Data stem**, we stem the corpora to reduce the vocabulary size (e.g., "changing" becomes "change": a word can have different forms in the singular, plural, tenses, and different parts of speech tags. For instance, going, goes, gone, and went all words have the base form 'go').

4.2 Topic Extraction

In this module, we used HLDA [27] because it properly conveys the relevance and structure of the topics. It is able to extract the relations between topics (parent-child and sibling relations) in order to visualize the topic hierarchy output. Moreover, a tree can be viewed as a nested sequence of partitions. Each topic, seen again as a probability distribution across words, is associated with a node in the tree, and therefore, each path is associated with an infinite collection of topics. We used a stochastic process called nested Chinese Restaurant Process (nCRP)[40] that represents an influential non parametric Bayesian used as a prior distribution to help to learn a tree structure and to organize a topic hierarchy into an L-level tree rather than a flat structure. The figure 2 is a

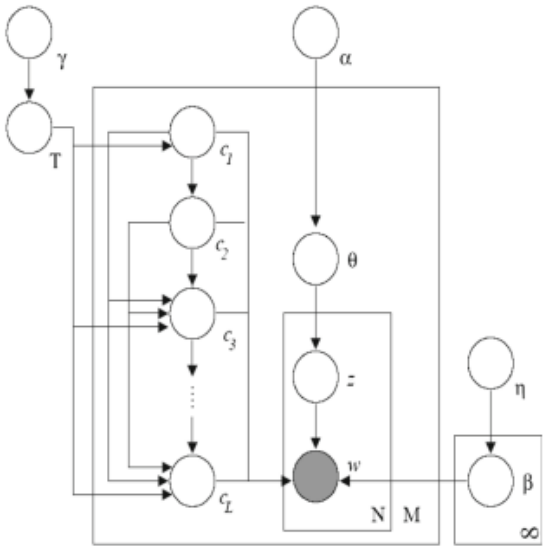

Figure 2: Graphical representation of HLDA model

graphical representation of hLDA model showing different variables of nCRP integrating with finite L-levels and infinite depth tree can be estimated by fixing L to a large number. Suppose a tree with L-levels, each node of the tree only, except leaves with infinite children. The root node has ID1 and each node of the tree has also unique IDs. It is shown the extracted topics in the form of L-level tree. Each topic is represented by tree node t and topic is associated with a distribution over words. All topics and nodes have 1-1 correspondence. The probability distribution of different paths (c1, c2..cn) on a tree defines by nCRP. The basic steps in nCRP to generate documents with L-level tree is:

- Select the path from the root to the leaf of a tree.
- Draw a vector θ for the topic mixing proportion from a Dirichlet with L-dimension.
- Generate the words in a document with the mixing proportion,θ from the distribution of topics along the path.

In hLDA model, all documents are allocated with a path following nCRP technique. All the words win a document d are distributed by a mixture of the topics representing the mixing proportion of a document shown in the algorithm 1. Following is the generative process used to draw a document from a textual corpus: The principle of hLDA consists of:

- Each node N in the tree is assigned to a topic T.
- Each path of length L in the tree consists of L topics.
- A probability distribution P over the topics in the path is defined using the stick-breaking distribution.

We provide below the hlda algorithm used to generate our Topic graph representation.

ALGORITHM 1: Generative process Hlda

Input: large corpus D= d1, d2, d3.. dn
Output: Topic Tree
L the height of topical tree ;
Iter the iteration number of Gibbs sampling ;
α, β, θ: hyperparameters ;
Associate the distribution of vocabulary over topic Z with the node in Topic Tree ;
for each node Z $\in$ $TopicTree\ do$
draw a topic β Z $\sim$ $Dir()$
end for ;
for each document d $\in$ 1, .., $D\ do\ Letcm$, $1betherootnode$
Let the root c1 ;
for each Level L $\in$ 2, .., $Ldo\ Drawthemixingproportionoftopic$
For every position n ;
choose a path for the node by
drawing the level Z dn ;
if $\alpha 1 = \alpha Z\ displaysymmetricDirichletdistribution$
end ;

4.3 Building Topic Graph Model

This phase aims to transform the textual information into a summary graph representation-based on extracted topics. It aims firstly at filtering the most relevant data that must be summarized and then constructing a topic graph summary. We propose a new method transforming the textual information into a model-based graph where relevant topics representing user profile preferences are represented as nodes named Topic Node (TN), and relationship named Topic Relationship (TR) between the nodes. The Topic Graph Model (TGM) i a graphical representation of the extracted topic hierarchy. It is represented by topic nodes and the relationships between them. The proposed TGM defines two main components: Topic Node and Topic Relation.
Formally: **TGM =(TN,TR)** Where:

TN = (IdN, NameN, Val,label) where:

IdN: is the node identifier
NameN:is the node name
Val: is the a value for each node terms
label: is a probability value distribution

TR= (IdR, NameR, nS,nD, label(Length of path= L) Where:

IdDR: is the relation identifier
NameDR: is the relation's name
dnS: is the topic node source of the relation ri
dnD: is the topic node destination of the relation ri
Label: length between nodes represented number of topic (it represent relation labeled)
To model the similarity between the user preference's nodes and generated ones, we assign a weight as an edge label representing a probability calculated using the semantic formula using a Euclidean distance. So, matching user profile with topic graph model is an automatic process based on similarity, calculating between terms representing the user profile and the topics generated by hLDA algorithm. There are a large number of similarity measures in the literature used to find the text similarity. The most used similarity measures are Cosine measure, Jaccard similarity and Euclidean similarity, etc. In our approach we use Euclidean distance:

$$\sqrt{\sum_{N}^{k=1} |topicsik - usertopicsjk|^2}.$$

Figure 3 shows an example of topics hierarchy generated by hlda algorithm from radiology reports.

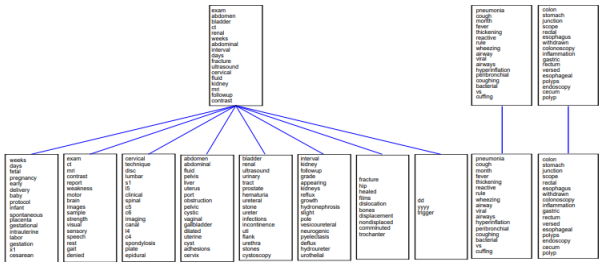

Figure 3: Part of topics hierarchy generated from radiology reports

4.4 User profile modeling

The ultimate goal of the user profile module is to take into account the user' needs. The purpose of this module consists of assigning extracted topics to the appropriate user. The user whose profile matches with the resulting topics. 2. Formally, User profile is defined as bellow:
UP= (Id, PD, TUM, AT, Not Allowed Topic, Interested Topic, Score Interested Topic, Not Interested topic, Score Not Interested Topic, Search History) Where

ALGORITHM 2: Transformation method based on Topic Modeling and user profile requirements

Input: large corpus D= d1,d2,d3..dn
Output: TGM= TN,TR) topic graph summary of the given document based on topic node (TN) and topic relation (TR);
begin;
step1: Preprocess the textual corpus (i.e. stop word removal, punctuation removal, lemmatization,);
SET cleaned textual document ;
FOR EACH di ∈ $corpusD\ SETcleaneddata$
Remove unnecessary characters (document);
Update cleaned date ;
Tokenize and Remove stopwords (cleaned data) ;
Update cleaned data ;
Genarate BIGRAMS and TRIGRAMS (cleaned data);
Update cleaned data Lemmatize Tokens ;
ADD cleaned data TO cleaned dataset;
SET dataset dictionary Generate Dictionary ;
SET dataset corpus ;
Generate Corpus (cleaned textual document);
end for ;
step2 using hLDA on input corpus;
Create a set of topic (using hLDA on input document);
Generate an hierarchie (prior distibution) ;
step3:construct user profile ;
modeling what it's allowed to be visualized in the profile;
modeling interests and attribute scores ;
modeling User search history;
Construct topic nodes;
construct topic relations;
step4 : Matching user profile with topic graph model;
calculate similarity between set of topics generate with hlda and user profile topics;
applied Euclidean Distance similarity measure

$$\sqrt{\sum_{N}^{k=1} |topicik - usertopicjk|^2}$$

step5:visualized topic graph summary;
Return TGM (TN, TR);
end ;

PD: Personnal data, is a set of attributes= p1, p2..pn
TUM: Type user model which can be explicit,or implicit
AT: Allowed Topic is a set of topic allowed to be visualized by the user. AT: val non interest topic = vt1, vt2, v3.. vtn
Not Allowed topic : set of topic that permit to visualized: val-non interest topic= wt1, vt2, vt3..vtn
Interest topic: set of topic that permit to visualized: val-non interest topic= vi1, v2,vi3..vin,score v1.. socre vn
Not interset topic: set of topic that unallowable to visualized: val non interest topic=w1,w2,w3..wm,scorew1..scorewm
Search history: set of topic that user search, (interest topic, score), allowed topic: st1, st2..stn, vi1, vi2, vi3..vin, score vi1..socre vin, vt1,vt2,v3..vtn. The algorithm 2 describes all different steps of our proposed approach.

4.5 Topic graph visualization

This module is responsible for the visual representation of data. It provides visual and interactive visualization and includes interactive techniques to graphically represent the summary.It allows to rapidly find insights in data.

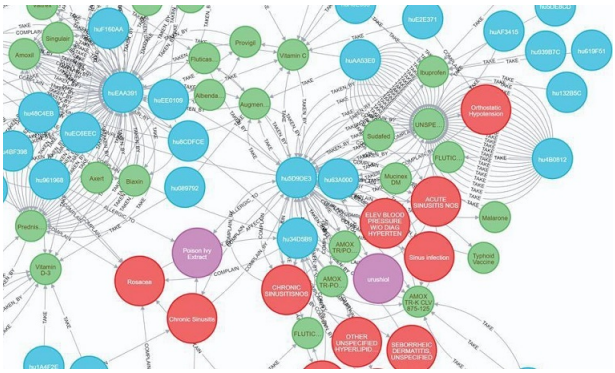

Figure 4: Topic graph based on user profile generated

In figure 4 a topic graph based on user profile is represented mainly by 3 different colors representing the topic node type. The red node represents the interest topics for user, the blue node represents the generated topic from hlda algorithm, and the green node represents the allowed topics to user.

5 EXPERIMENTAL ANALYSIS

Our experiments focused on the medical domain. We used a Pubmed dataset containing 100 textual documents. PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally. The PubMed database contains more than 33 million citations and abstracts of biomedical literature. It does not include full text journal articles; however, links to the full text are often present when available from other sources. In this paper, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[29]. ROUGE is used to measure the summarization performance, which is widely applied by Pubmed for performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. We used ROUGE metric in order to measure the performance of our summary without integrating the user profile and when incorporating it in the summary process. ROUGE-1 and ROUGE-2 are used to compute the F-measure, precison and recall value for unigram and bigram between the system and reference summaries. The F-measure is used to measure the clustering performance so it's obtained from the measurement of recall and precision. Recall is used to calculate the ratio of acquired relevant documents by the total number of documents in documents collection. Meanwhile, precision is used to measure the ratio of the retrieved relevant documents number with a whole number of retrieved document.

Precision = RelevantTopics $\cap RetrievedTopics/RetrievedTopics$
Recall = $RelevantTopics \cap RetrievedTopics/RelevantTopics$

Fscore = 2 * Precision Recall / Precision + Recall

Also, we evaluate our approach among existing ones. Our objective was to demonstrate the importance of relationships between topics and the role of the matching of user profile based on topic with the topic graph model construct to guarantee the pertinence and consistence of information.

Table 3: Comparison text summarization methods for F-Measure into ROUGE 1 / ROUGE2

approach	ROUGE1	ROUGE2
Fu et˙al. 2020	0.420	0.193
Ramesh˙et al 2021	0.422	0.198
Proposed method	0.426	0.201

Table 4: Comparison text summarization methods for Precision into ROUGE 1 / ROUGE2

approach	ROUGE1	ROUGE2
Fu et˙al. 2020	0.176	0.059
Ramesh˙et al 2021	0.179	0.063
Proposed method	0.181	0.067

Table 5: Comparison text summarization methods for Recall into ROUGE 1 / ROUGE2

approach	ROUGE1	ROUGE2
Fu et˙al. 2020	0.549	0.186
Ramesh˙et al 2021	0.556	0.189
Proposed method	0.601	0.192

To evaluate the effectiveness of the proposed method, we Have compared the results with different text summarization Methods mentioned in our related work show in 3,4 and 5 with respect to F-measure, Precession, and Recall for ROUGE-1, ROUGE-2 metrics. We consider that taking a relationship for topic modeling in the summarization process influenced in the pertinence and the performance to generate an effective and robust summary also we validate the primordial role of summarization oriented user profile to generate a performed results that satisfy the user's need of relevant.

6 CONCLUSION

In this paper, we investigated ways to automatically discover a hierarchical topic modeling with user profile incorporating. So our work proposes a new text transformation method based on the unsupervised hLDA algorithm and incorporating user profile. However the existing techniques usually favor brevity instead of incorporating all the essential information present within the source large corpus. We used topic modeling to identify salient topics present in a document to be summarized and have generated topic graph text around those node topics and relation topics and we ameliorate results into user profile integration .In the future work

we will interested for generate concepts from topics by matching topic graph model with domain ontology such as (umls).

REFERENCES

[1] D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, p. 147, 2006.

[2] P. Anupriya and S. Karpagavalli, "LDA based topic modeling of journal abstracts," in 2015 International

[3] Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM 38(11), 39–41 (1995)

[4] Roul, Rajendra Kumar. "Topic modeling combined with classification technique for extractive multi-document text summarization." Soft Computing 25.2 (2021): 1113-1127.

[5] Hafeez, Rubab, et al. "Topic based Summarization of Multiple Documents using Semantic Analysis and Clustering." 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT (HONET-ICT). IEEE, 2018.

[6] Twinandilla, Shiva, et al. "Multi-document summarization using k-means and latent dirichlet allocation (lda)–significance sentences." Procedia Computer Science 135 (2018): 663-670.

[7] Litvak, Marina, Natalia Vanetik, and Lei Li. "Summarizing Weibo with Topics Compression." International Conference on Computational Linguistics and Intelligent Text Processing. Springer, Cham, 2017.

[8] Wu, Zongda, et al. "A topic modeling based approach to novel document automatic summarization." Expert Systems with Applications 84 (2017): 12-23.

[9] Singh, Ksh, H. Mamata Devi, and Anjana Kakoti Mahanta. "Document representation techniques and their effect on the document Clustering and Classification: A Review." International Journal of Advanced Research in Computer Science 8.5 (2017).

[10] Bashir, Muazzam, Azilawati Rozaimee, and Wan Malini Wan Isa. "Automatic Hausa language text summarization based on feature extraction using Naive Bayes Model." World Applied Science Journal 35.9 (2017): 2074-2080.

[11] Putri, Indiati Restu, and Retno Kusumaningrum. "Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia." Journal of Physics: Conference Series. Vol. 801. No. 1. IOP Publishing, 2017.

[12] Huang, Taiwen, Lei Li, and Yazhao Zhang. "Multilingual multi-document summarization with enhanced hLDA features." Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer, Cham, 2016. 299-312.

[13] Irawan, Santun. "Studi Awal Peringkasan Dokumen Bahasa Indonesia Menggunakan Metode Latent Semantik Analysis dan Maximum Marginal Relevance." Annual Research Seminar (ARS). Vol. 2. No. 1. 2017.

[14] Chiru, Costin-Gabriel, Traian Rebedea, and Silvia Ciotec. "Comparison between LSA-LDA-Lexical Chains." WEBIST (2). 2014.

[15] Kim, Hyun Duk, et al. "Enriching text representation with frequent pattern mining for probabilistic topic modeling." Proceedings of the American Society for Information Science and Technology 49.1 (2012): 1-10.

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, 2003, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Volume 3, pp. 993-1022.

[17] Zhou Tong, and Haiyi Zhang, 2016, "A Text Mining Research Based on LDA Topic Modeling", The Sixth International Conference on Computer Science, Engineering and Information Technology, Volume 6, pp. 201-210.

[18] Melissa Ailem, Bowen Zhang, and Fei Sha, 2019, "Topic Augmented Generator for Abstractive Summarization", ArXiv

[19] Allahyari M, Pouriyeh S, Assef M, Safaei S, Trippe ED, Gutierrez JB,Kochut K (2017) Text summarization techniques: a brief survey. arXiv preprint arXiv:170702268

[20] Lim KW, Buntine W, Chen C, Du L (2016) Nonparametric bayesian topic modelling with the hierarchical pitman-yor processes. Int J Approx Reason 78:172–191

[21] Cuong HN, Tran VD, Van LN, Than K (2019) Eliminating overftting of probabilistic topic models on short and noisy text: the role of dropout. Int J Approx Reason.

[22] Amplayo RK, Song M (2017) An adaptable fne-grained sentiment analysis for summarization of multiple short online reviews. Data Knowl Eng 110:54–67

[23] Zhang L, Wu Z, Bu Z, Jiang Y, Cao J (2018a) A pattern-based topic detection and analysis system on Chinese tweets. J Comput Sci 28:369–381

[24] Barros C, Lloret E, Saquete E, Navarro-Colorado B (2019) Natsum:Narrative abstractive summarization through cross-document timeline generation. Inform Process Manag 56(5):1775–1793

[25] Fu X, Wang J, Zhang J, Wei J, Yang Z (2020) Document summarization with vhtm: Variational hierarchical topic-aware mechanism. In: AAAI, pp 7740–7747

[26] Widyassari, Adhika Pramita, et al. "Review of automatic text summarization techniques methods." Journal of King Saud University-Computer and Information Sciences (2020).

[27] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[28] Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T,Mcnamara D, Dennis S, Kintsch W (Eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007

[29] Lin CY (2004) Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.

[30] Zukerman, I., Albrecht, D.: Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction 11(2), 5–18 (2001)

[31] Blei D M, McAuliffe J. Supervised topic models. In: Advances in Neural Information Processing Systems (NIPS) 21. Cambridge, MA, MIT Press, 2007, 121–128nd

[32] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. Amer. Soc. Inf.Sci., vol. 41, no. 6, pp. 391–407, 1990

[33] T. Hofmann, "Probabilistic latent semantic analysis," in Proc. 15th Conf. Uncertain. Artif. Intell., San Francisco, CA, USA, 1999, pp. 289–296.

[34] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. 23rd Int.Conf. Mach. Learn., Pittsburgh, PA, USA, Jun. 2006, pp. 113–120

[35] Mimno, David, Wei Li, and Andrew McCallum. "Mixtures of hierarchical topics with pachinko allocation." Proceedings of the 24th international conference on Machine learning. 2007.

[36] Yang, G., Wen, D., Chen, N. S., Sutinen, E., 2015. A novel contextual topic model for multi-document summarization. Expert Systems with Applications 42 (3), 1340–1352

[37] Bairi, R., Iyer, R., Ramakrishnan, G., Bilmes, J., 2015. Summarization of multi-document topic hierarchies using submodular mixtures. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 553–563.

[38] Riddell, A., 2013. Demography of literary form: Probabilistic models for literary history. Ph.D. thesis, Duke University

[39] Yuan, J., Sivrikaya, F., Hopfgartner, F., Lommatzsch, A., Mu, M., 2015. Context-aware lda: Balancing relevance and diversity in tv content recommenders. In: Proceedings of the 2nd Workshop on Recommendation Systems for Television and Online Video

[40] Ali, W., Rehman, Z., Rehman, A. U., Slaman, M. (2018, November). Detection of plagiarism in Urdu text documents. In 2018 14th International Conference on Emerging Technologies (ICET) (pp. 1-6). IEEE