TOPIC MODEL IMPLEMENTATION TO FIND RELATED DOCUMENTS IN CORPORATE ARCHIVES IN REAL LIFE: "A CASE SCENARIO ON KNOWLEDGE RETRIEVAL"

İhsan Tolga Medeni
Çankaya University, METU
Specialist, PhD Student
tolgamedeni@gmail.com

Tunç Durmuş Medeni
Yıldırım Beyazıt University, Turksat, METU, Çankaya University
Instructor, Senior Specialist
tuncmedeni@gmial.com

─Abstract ─

Today's organizations were mostly built over their documents. These documents are very crucial sources of knowledge. Even they know the existence of these documents, most of the time, it is nearly impossible to extract captive knowledge inside. In these conditions, organizations choose re-prepare same document again rather than finding proper documents in the archives. On the other hand, finding these documents would save precious time and decrease redundancy of the work. Topic model idea basically focuses on extraction of knowledge from these types of documents. In this study, our aim is to give a summary of Topic Model research and try to explain latest model concept over an imaginary case scenario.

Key Words: *Topic Model, Knowledge Extraction, Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA) ,Latent Dirichlet Allocation (LDA)*
JEL Classification: C60 , D83

1. INTRODUCTION

When an unexpected condition occurs in the organizations, it is needed to find solutions in their organizations flow with creating high volume of documents. In reality, probably this organization could face similar problem in its life, and there is already a set of information and documents exists for solving the condition. As pointed out in Davenport and Prousark's Working Knowledge, if the organization could have proper knowledge structure independent from the individual workers' experience, which could easily lost with the end of this workers carrier in the organization, they would not need to redundant jobs again and again. (Davenport, 2001) One of the solutions to this problem is the topic model concept.

Topic modeling concept defined for the need of extracting information without inclusion of any user queries. (Deerwester,1999) Topic models allow presentation of documents as collections of topics rather than collections of words (Zheng,2009). Working on this concept starts with 90s with the vector space model and continues with latent semantic analysis/indexing (LSA, LSI), probabilistic LSA (pLSA) and Latent Dirichlet Allocation (LDA). Under different research areas from medical science to software engineering, topic models have been used with different names such as Information Retrieval (IR), dimensional reduction, word matching etc. from text mining to the image processing.

In this paper, with focusing on these methods, a general summary will be given. After then, a possible application scenario from organizational application perspective will be given. This paper will be ended with the conclusion part.

2. TOPIC MODEL EVALUTION

2.1. Vector Space model to LSI/LSA

In 1990, Deerwester and his colleagues proposed an approach for automatic indexing and retrieval. (Deerwester,1999) According to them, the existed techniques based on user queries just applied to match words of the user queries. However this approach just does not include any evidential information about the

meaning and concept of the document and topic. Taking statistical approach on one hand, latent semantic index (LSI) analysis (LSA), a semantic space was built to show association of terms and documents. This concept was created over vector space model. Under this concept, text documents represented as vector of terms and relationships between documents and terms represented in a matrix. Cosine of angle in between vectors represents similarity between two documents. (Poshyvanyk,2006) The work of Deerwester differs from simple vector space models with taking the concepts of synonym and polysemy into account. Polysemy is defined as carrying more than one meaning in a simple word.(Deerwester,1999) For example orange, it could be considered as a fruit in one document and on another, it could be taken as a color. On the other hand, synonym referred as ability to refer a concept with more than one single word. Auto and car words could be given as example to this concept. They all could be used to point one meaning, the automobile. Even it was given as supportive to the synonym and polysemy, in reality result of LSI does not show better performance of polysemy when compared to synonymy. (Lukins,2010) Reaching a satisfactory topic set also another problem related with the LSA.
To address these problems, (Hoffmann, 1999) proposed probabilistic Latent Semantic Analysis, pLSA.

2.2. LSI/LSA to pLSA

In pLSA, each term modeled over a set of multinominal variables according to related documents. This model build based on probabilistic distribution of terms in each document. Following with this construct, pLSA shows improvements over LSA. (Hoffman, 1999) (Blei, Ng and Jordan,2003)

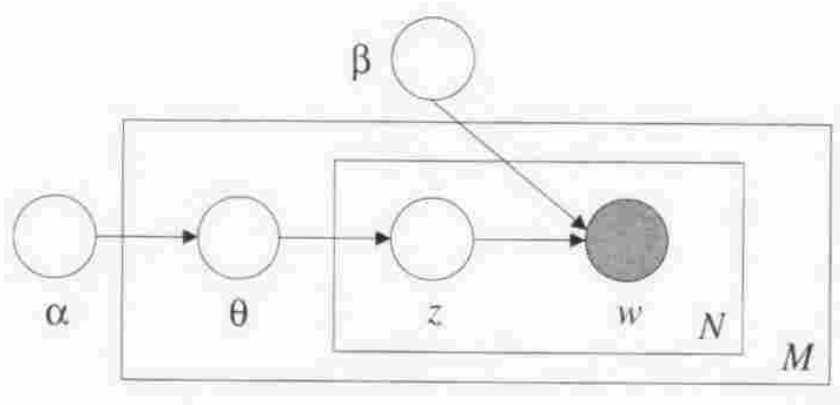
With its improvements, pLSA shows promising results with its document oriented linearly growing model. However this model also introduced an overfitting problem. (Girolami,2003) This problem occurs when new documents are introduced to the previously trained structure. pLSA structure tend to find topics in new documents according to estimated, previously distributed documents. (Blei,2003) (Wei,2006). Similar to pLSA, LDA was introduced to solve this problem with the work of Blei in 2003. (Blei,2003) The studies showed undeniable result to support LDA implementation when comparing [20] results of pLSA, LDA on the same corpora.

2.3. pLSA to LDA

Latent Dirichlet Allocation, provides a means of fitting the Diriclet parameter with a given document set.
In LDA, every document is taken as finite mixture words contains of set of multiple topics. (Blei,2003)(Zheng,2006). In this probabilistic model, word, document and corpus are the main concepts. In (Blei,2003) these concepts are defined as follows;

- A word is the unit basically defines an item from a dictionary indexed from 1 to V.
- A document is the combination of N words and represented by $w=(w_1, w_2, \ldots, w_m)$
- A corpus is a document set that is represented by $D=\{w_1, w_2, \ldots, w_m\}$

Figure-1: Graphical Model Representation of LDA from



Source: Blei and Jordan,2003

In figure 1, the boxes are represented as plates. The outer plate represents documents, while inner plate represents the choice of topics and words within documents. LDA follows the generative process for each document w in a corpus *D*.

- α is the parameter of the Dirichlet prior on the per-document topic distributions.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.
- θ_i is the topic distribution for document i,
- ϕ_k is the word distribution for topic k,
- z_{ij} is the topic for the jth word in document i, and
- w_{ij} is the specific word.

LDA retains advantages over pLSA and LSA, however, when it compares for efficiency it suffers. (Kakkonen,2008) To overcome this problem, it is required to escape from direct computation. This is supported via approximation techniques.(Wie,2006). (Blei,2003) Bayesian based approaches, Marcov Chain Mante Carlo (MCMC) sampling (Gethers,2010), Gibbs sampling (Tian,2009) could be applied for direct estimation of the assigned words to topics given the reflected words in a corpus. (Griffiths,2004) Gibbs sampling applied to generate samples by iteratively sampling and updating each component variables. (Zheng,2006)

To train a LDA model, number of topics should be specified, for both semantic analysis and statistical learning. (Zheng,2006) In a standard LDA model the selected number of topic ranges from 50 to 300 topics. Most preferably, 100 topics used in a document collections up to 20000 documents. (Linstead,2007)With the corpora contains 30000 to 40000 documents three hundred topics become preferable. (Steyvers,2007) (Griffiths,2004) Bayesian model selection approach used to determine optimal number of topics for the purpose of model fitting. (Zheng,2006)

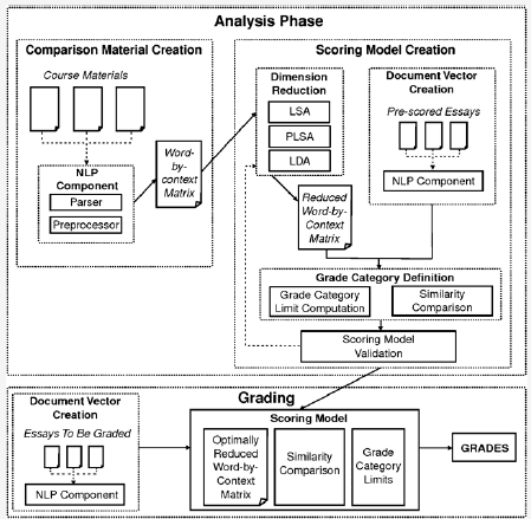
2.4. LDA and Related Studies

Even LDA shows improvements over LSA and pLSA, studies continues on both subject and possible application areas sometimes comparing with LSA, pLSA and sometimes comparing LDA with other concepts originally created from LDA.

In (Zheng,2009) the original LDA is taken as fitted LDA and compared with unfitted LDA. In this article, storing the topic model probabilities in the document index taken as original LDA approach, called as fitted LDA, and storing these probabilities as term relations called as unfitted LDA. Their results shows that,

computing fitted LDA do not bring any benefit over the unfitted LDA. One of these studies is compared LSA, pLSA and LDA with each other via essay comparison for grading. (Kakkonen,2008)

Figure-2: The architecture of AES



Source: Kakkonen, Myller, Sutinen and Timonen.(2008).

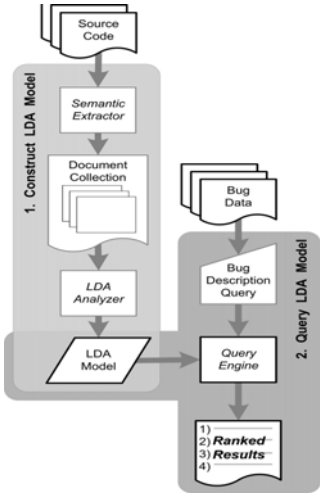
As it can be seen from figure 2, LSA,pLSA and LDA was included as dimension reduction technique. According to this study, still LDA gives proper results when it is compared with other topic models. For empirical comparison of these methods, k-Nearest Neighbors (k-NN) was implemented.

Software related concepts such as bug location (Lukins,2010) and (Gethers,2010) can give possible candidates for application of LDA. In (Lukins,2010), to LDA was implemented to locate possible programming bugs. The concept was tested in five case studies. The following figure gives their structure.

Relations between software source code elements could be calculated through LDA based systems. (Gethers,2010) With an implemented corpus to support these elements, coupling metrics can be built for efficient coding.

To group different software products, (Tian,2009) was used LDA on a Gibbs sampling to create corpus index.

Figure-3: LDA-Based approach applied in

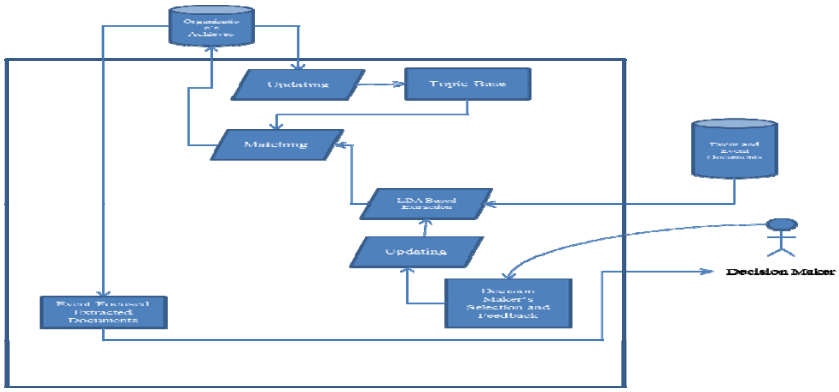


Source: Lukins,Kraft and Etzkorn:2010

In this study, with respect to current evaluation of the topic models, as LDA based system structure will be presented.

3. THE POSSIBLE CASE

Figure-4: LDA-Based System



The enter case scenario, the important parts of the system could be identified as follows;

Topic Base: As it can be seen from the figure 4, Topic Base is related with organization's archive part. At the first run of this system, Topic Base is created with LDA based extraction specific to the organizations digital archives, however, in this figure; the system picture is taken from a time further that point. From time to time, system will need to update the topic base based on new inclusion to the archives so when it is needed to topic base is updated accordingly.

Event: When a new event occurs, the event related documents are submitted to the system. These documents could be any text document such as internal process document, an email… Each document should be sent to the LDA Based extraction to topic extraction from the documents.

Matching: On the matching step, the extracted topics are tried to match with the current system topics. After that, related documents are obtained from organization archive. These documents are event specific documents because they contain event specific topics inside. The documents presented to the decision maker and if all/or some of them are accepted, the decision is send back to the system with the related feedback by the user for the future extraction processes.

3.1. Case

A proper explanatory case could be given as follows; when the system receives a consumer email related with the problem of a new item, first of all, it will put this

email into LDA Based extraction process to extract related topics from the email. After that, with using the topic base, it matches the findings with the current topic base. From the connections of topic base's topics, the most proper documentation(s) (in this case, a proper solution, a procedure documentation, or a formal documentation to help organization to get rid of legal issues...) will be sent directly to the related officer (in this case that person is the decision maker) with the current email. With these documentations, decision maker could create a response in a short notice. As an optional step, if officer does not satisfied with the current finding, a new inquiry could start with the selected topics or weights with respect to decision of the officer.

4. CONCLUSION

This system is still in the development stage it brings opportunities with it. However, from looking at the current developments, it is easy to implement this system, even without need to implementation of a decision maker role. With a system like this, it is easy to automate all procedure with a parallel scenario like given in the case part.

BIBLIOGRAPHY

Blei, Ng, Jordan,(2003), "Latent Dirichlet Allocation", *Journal of. Machine. Learning*. Vol..3, pp. 993–1022.

Davenport, Prusak, (2000), Working Knowledge:How Organizations Manage What They Know, Boston, Harward Business School Press.

Deerwester, Dumais, Furnas, Landauer, Harshman (1990), "Indexing by Latent Semantic Analysis" *Journal of the American Society for Information Science*, , Vol.41,No.6,pp.391-407.

Gethers, Poshyvanyk,(2010),"Using Relational Topic Models to Capture Coupling among Classes in Object-Oriented Software Systems", *IEEE International Conference on Software Maintenance*, 2010.

Girolami, Kabán, (2003), "On an equivalence between PLSI and LDA", *in: Proc. 22nd Annu. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, Toronto, Ontario, Canada, , pp. 433–434.

Griffiths, Steyvers, (2004) "Finding scientific topics", *Proc. Nat. Acad. Sci.* Vol.101 No.1 , pp. 5228–5235.

Hofmann (1999), "Probabilistic latent semantic indexing", *in: Proc. 22nd Annu. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, Berkeley, CA, USA, , pp. 50–57.

Kakkonen, Myller, Sutinen, Timonen.(2008) "Comparison of Dimension Reduction Methods for Automated Essay Grading", *Educational Technology & Society*;Vol.11, No.3,pp.275-288.

Linstead, Rigor, Bajracharya, Lopes, Baldi,(2007), "Mining concepts from code with probabilistic topic models", *in: Proc. 22nd IEEE/ACM Int. Conf. on Automated*

Lukins, Kraft, Etzkorn.(2010),"Bug localization using latent Dirichlet allocation". *Information and Software Technology* Vol.52, No.9,pp.972-990.

Poshyvanyk, Guéhéneuc, Marcus, G. Antoniol, Rajlich (2006), "Combining probabilistic ranking and latent semantic indexing for feature location", *in:Proc. 14th IEEE Int. Conf. on Program Comprehension,* Athens, Greece, , pp. 137–148.

Steyvers, Griffiths, (2007), "Probabilistic topic models", (in: Landauer, McNamara, Dennis, Kintsch-Ed, *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates..

Tian, Revelle, Poshyvanyk.(2009), "Using Latent Dirichlet Allocation for Automatic Categorization of Software". *6th Ieee International Working Conference on Mining Software Repositories* pp.163-166.

Wei, Croft,(2006) "LDA-based document models for ad-hoc retrieval", *in: Proc. 29th Annu. Int. ACM SIGIR Conf. on Research & Development on Information Retrieval*, WA, USA , pp. 178–185.

Zheng, McLean, Lu, (2006), "Identifying biological concepts from a protein-related corpus with a probabilistic topic model". *Bmc Bioinformatics* Vol.7.

Park, Ramamohanarao,.(2009), "The Sensitivity of Latent Dirichlet Allocation for Information Retrieval". (In: Buntine, Grobelnik, Mladenić, Shawe-Taylor-Ed. ,*Machine Learning and Knowledge Discovery in Databases*): Springer Berlin Heidelberg, pp. 176-188.