Topic modeling (LSA with SVD approach)

Ching-Wen Wang, Mo Xiao

Introduction (Ching-Wen Wang):

Mass amount of data are being generated every second across different platforms. According to techJury (https://techjury.net/blog/how-much-data-is-created-every-day/#gref), a person generates on average around 1.7 MB of data every second. If we must search through every single file to find the desired information from this ocean of data, it would cost huge amount of time and money. Therefore, if we can efficiently associate each document with relevant topics it contains, we can improve the speed and reduce the cost of retrieving a particular information.

This goal can be achieved through topic modeling, which is a method of discovering abstract topics embedded in a collection of documents. Topic modeling is widely adopted in many applications, for example, email spam filter and recommender system. It categorizes and summarizes large collections of unstructured data using methods such as term frequencies or distance between words, so that people can understand large amounts of data quickly, cheaply, and insightfully. Ever since the first model, Latent semantic indexing (LDI), also known as Latent semantic analysis (LSA) was published in 1998, more models are developed one after another. For example, the probabilistic latent semantic analysis (PLSA) in 1999, Latent Dirichlet allocation (LDA) in 2002, which is a very popular method used in industry, and Hierarchical latent tree analysis (HLTA) in 2019.

In this project, we will explore a renowned method of topic modeling, latent semantic indexing (LSI), that uses a singular value decomposition (SVD) approach (Gong & Liu, 2001). Although LSI can enhance information retrieval, in real world applications, there are limitations such as not enough space to store large vocabulary sets or uninterpretable outputs. Therefore, we are keen to find out how LSI can be improved by incorporating syntactic information (Kanejiya, Kumar & Prasad, 2003), part of speech (Kakkonen, Myller & Sutinen, 2006), or implementing regularization methods of lasso and ridge regression (Xu, Hang, L & Craswell, 2011).

Source 1: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis (Mo Xiao)

Gong, Y., & Liu, X. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *SIGIR'01.*
https://www.cs.bham.ac.uk/~pxt/IDA/text_summary.pdf

In this paper, the authors attempt to develop methods for genetic text summarization. They want to present key points of documents to help users quickly extract the ones they need. In the paper, they propose two text summarization methods. The first one uses relevance measure to find sentences that best represent the key ideas of the document while being the least redundant: the document is first broken down into sentences, then a matrix consisted of the weighted frequency of terms for each sentence, and a vector consisted of the weighted frequency of terms of the whole document is created. For each sentence, the relevance score is defined as the inner

product of the matrix and the vector, which represent how relevant each sentence is to the whole document, and the sentence with the highest relevance score is added to the summary sentence set. After a sentence is added to the summary, itself and all the terms in it are eliminated from original document, then the whole process is repeated until our summary reaches a desired length. The second generic text summarization method uses latent semantic analysis and involves singular value decomposition. Similar to the first approach, a matrix with the weighted frequency of terms for all the sentences is created and decomposed by SVD into a matrix with left singular matrix, a diagonal matrix, and a matrix with right singular vectors. Sentences with the largest index value with the first k right singular vectors are used as summary of the document, where k represent our desired sentence number. The authors use 243 documents from the CNN Worldview news program, and compare the selection of sentences with the two methods with the selections of three human summarizers. Both methods have selections around 50%-60% the same as manual selections.

Both approaches seem feasible, but the compotation cost of the two methods is not addressed in the paper. The matrix and vector multiplication required in the relevance measure approach may be especially costly when facing long documents. Both methods reach only 50-60 percent of agreement with manual summaries, which could partly be explained by the disagreement between the three human summarizers. Also, since there is no key words of queries provided in advance, it is hard to determine the effectiveness of the summarization. The evidence for effectiveness is not so much compelling, as the test set of 243 documents may not be sufficient to determine the effectiveness of the summarization approaches, so the claims in the paper could be further refined if we test on more dataset.

Source 2: Automatic evaluation of students' answers using syntactically enhanced LSA (Mo Xiao)

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, *2*, 53–60. https://doi.org/10.3115/1118894.1118902

This paper introduces a syntactically enhanced latent semantic analysis (SELSA) and compare it with the ordinary latent semantic analysis (LSA), in terms of evaluating students' answers based on a given full-mark answer. While the ordinary LSA method measures frequency of words in a document and focus only on semantic, the SELSA method takes the context of words and concludes the syntactic-semantic meaning of words based on it. The mechanism starts by giving all elements in a corpus with a part-of-speech (POS) tag, and the combination of a word and its POS tag is called the word-prevtag pair. A matrix representing the pair (rows) and documents (column) in the corpus is then created, and the frequency of the word-prevtag pairs is found by summing the frequency of the pair across all documents. A matrix consisted with the weighted frequency of each word-prevtag is obtained and decomposed by singular value decomposition (SVD) and keep only the largest n singular values to get the summarization. The effectiveness of LSA and SELSA methods are studied with the experiment of evaluating students' answers. Evaluations with both methods for 5596 documents is compared to manual evaluations from 4 people. The result shows that the LSA method could generate a 51%

similarity with the manual evaluation, while the SELSA method could only generate a 47% similarity, and the average similarity between different human evaluations is 59%.

The SELSA method takes the context of words into consideration, which seems to be reasonable, but the non-satisfying performance indicate that the method need to be refined. The performance of the SELSA method is greatly affected by the corpus and accuracy of POS tag which adds challenges as we first need to obtain a effective POS system. The evaluation section of the paper is very compelling, analyzing the effectiveness with different approaches, and the dataset used is also sufficient for reasonable conclusions. However, the evaluation of the two methods may still requires more testing from documents other than students' answers before we could decide which method is more appropriate in certain circumstances.

Source 3: APPLYING PART-OF-SPEECH ENHANCED LSA TO AUTOMATIC ESSAY GRADING (Ching-Wen Wang)

Kakkonen, T., Myller N., & Sutinen E. (2006). Applying Part-of-Speech Enhanced LSA to Automatic Essay Grading. Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education. https://arxiv.org/abs/cs/0610118

This paper is written by authors who attempt to increase the performance of automatic essay grading systems by applying part of speech(POS) enhanced LSA instead of traditional LSA. The process of automatic grading involves giving scores by comparing the similarity between word-context matrix (WCM) generated from student's essays and the WCM retrieved from passing in course materials into the LSA model. Each entry of the word-context matrix (WCM) represents the number of occurrences of each word, excluding the stop words, in the corresponding context, such as document, paragraph, or sentence. However, the limitation with using LSA is that it is not sensitive in capturing the word to word relationship of sentences and the word to context relationship. Therefore, since structure of sentences can provide information on semantics of words, the authors decide to add different combinations of POS tags to each distinct word in WCM and evaluate the models by its score correlation with human-graded essays. The model that receives the highest correlation 0.829 is the one with POS tags of the previous ,current, and next word. It's accuracy improves 10.77 percent compared to traditional LSA.

The group of words feeding into the models are combinations of all words and content words (noun, verb, adjective), with ambiguous words, which are words that have the same basic form but mean and spell differently. The approach that this experiment took is to display the difference in accuracy with the baseline LSA model every time one or a group of new POS tags are added. The experiment result is considered to be reliable, because the test sets are chosen among 3 different domains: education, communication, and Software engineering with 3 different graders, and the evaluation method is accurate given the purpose of automatic grading is to mimic human graders. Although the effectiveness of this model is convincing, the greatest limitation is the size of storage space. The best model generates an extremely large WCM matrix with content words and ambiguous words that has 3 components in each word. Therefore, in general, these models can only be implemented with small collections of documents. Also, the model is experimented in Finnish, so it is unknown whether it can achieve a significant amount of difference compared to the baseline LSA. In conclusion, this model offers insights on how

models with POS tags can perform, as well as pointing out future directions for research, such as integrating sentence structure with another topic modeling method Probabilistic Latent Semantic Analysis (PLSA) or implementing this POSELSA model in a different language.

Source 4: Regularized Latent Semantic Indexing for Topic Modeling (Ching-Wen Wang)

Wang, Q., Xu, J., Li, H., & Craswell, N. (2013). Regularized latent semantic indexing: A new approach to large-scale topic modeling. ACM Transactions on Information Systems (TOIS), 31(1), 1-44. http://www.hangli-hl.com/uploads/3/1/6/8/3168008/rlsi-tois-revision.pdf

Previous solutions to topic modeling on problems of handling large inputs of document collections and the running complexity are to decrease the input vocabulary collections or approximate the output topics. However, the author's showed that these methods can decrease the accuracy of generated topics and their readabilities. Therefore, they proposed a new method of topic modeling, Regularized Latent Semantic Indexing (RLSI), that involves optimizing the 2 decomposed matrices: term-topic matrix and topic-document matrix from the term-document matrix with regularization of L1 and L2 norm over several processors. By making it parallelizable, storage space is saved by shrinking the number of terms and optimizing the matching of topics to documents. RLSI differs with LSI in that parallelization is not achievable on LSI, since LSI has to satisfy orthogonality constraints by applying SVD. The results of online and batch RLSI models are evaluated empirically using 3 different criterias, topic readability, topic compactness, and retrieval performance, on 3 different TREC datasets and web datasets. Through critical evaluation on experimental results, RLSI performance is proved to perform significantly better on topic readability and distinguishability than LSI and some other existing methods. Also, online RLSI is shown to be capable of processing 1.6 million documents and 10 thousand queries with improved performance and tracking the dynamic changes of topics over time.

However, the major limitation of RLSI is the computation cost. While online RLSI doesn't occupy a lot of storage space, it is costly in its computation process. Online RLSI utilizes stochastic learning algorithms to update its term-topic and topic-document matrices. This method only speeds up the computational time of the topic-document matrix, and leaves most of the costs coming from computing term-topic matrix. Another limitation that RLSI has is that the overall performance of RLSI did not show to exceed Latent Dirichlet Allocation(LDA), which is a probabilistic topic modeling, so more study may be required to enhance the model. Overall, the paper showed detailed theoretical proofs of model convergence, experimental procedure and results, empirical evaluation method of relevance ranking, and calculated time and space complexity to show the effectiveness of its approach in topic modeling. Also, future directions are discussed in improving the current version of RLSI by tuning parameters, and designing new algorithms that lower space complexity and computational costs at the same time for larger datasets.

Conclusion (Mo Xiao):

The latent semantic approach plays an important role in text summarization and document comparison, which greatly boosts the efficiency of sifting vast information. Refined

approaches of latent semantic analysis, including enhanced latent semantic analysis and regulated latent semantic indexing are introduced for better performance. Current applications of latent semantic techniques include document summarization, evaluation of students' answers and essays, extracting key topics, and so on. However, many methods still face the problem of computational complexity, and some of methods do not perform so well. More training and testing are needed to improve the effectiveness of LSI and find the best fit under different circumstances.