A Modern Approach for Urgent Script Cluster Processing and Summarization with Involuntary Length Recognition

K. Nithya, A. Rajiv Kannan

Abstract- Detection the apposite extent of clusters to which credentials should be separation is vital in text cluster. In this dissertation, we suggest a fresh approach, namely DPMTP (Dirichilet Process Model Trait Partition), to realize the embryonic huddle construction based on the DPM model lacking requiring the amount of huddle as key. Elements classify into two class, important expressions and un match terms.Also find the new approach for simultaneouslyclustering and summarization.Probabilistic Hidden Semantic Analysis has been popularly used in document analysis.Topropose Bi-mixture Probabilistic Hidden Semantic Analysis , a new formulation of PHSA that allows the number of latent word classes to be different from the number of latent document classes.Extended method of Bi-PHSA Bi-mixture PHSA with sentence bases (Bi-PHSAS) to simultaneously cluster and summarize the documents utilizing the mutual influence of the document clustering and summarization procedures. Additionally propose a Bayesian nonparametric model for multidocument summarization in order to determine the proper lengths of sum maries.

Keywords- Huddle, DMA, Trait Partition, DPMTP, BNP Summarization.

I. INTRODUCTION

Manuscript-clustering, amalgamation of unlabeled manuscript credentials into significant huddle, is of vital awareness in many application.One hypothesis, taken by deep-rooted manuscript huddle approach as in is that the quantity of huddle Z is known ahead of the method of manuscript huddle. Embryonic, every distinct solitary of credentials rummage around by abusers and fairly accurate Z. This is not just with position to moment overriding other than also out of accomplish other than continually as soon as production with massive manuscript statistics set. Besides, an offensive evaluation of Z valor without doubt hoodwink the huddle progression Document summarization is an important task in the area of natural language processing, which aims to extract the most important information from a single document or a cluster of documents. In various summarization tasks, the summary length is manually de-fined. However, how to find the proper summary length is quite a problem; and keeping all summaries restricted to the same length is not always a good choice. It is obviously improper to generate summaries with the same length for two clusters of documents which contain quite different quantity of information.

Manuscript received March 15, 2014.

K. Nithya, M.E (CSE), K.S.R.College of Engineering, Tiruchengode.
Dr. A. Rajiv Kannan, M.E., Ph.D.,(Head of the Department), K.S.R.College of Engineering, Tiruchengode.

Probabilistic hidden Semantic Analysis (PHSA) has been popularly used in document analysis. However, as it is currently formulated, PHSA strictly requires the number of word latent classes to be equal to the number of document latent classes. In this paper, we propose Bi-mixture PHSA, a new formulation of PLHA that allows the number of latent word classes to be different from the number of latent document classes.

II. BRIEF EXPLAINATION FOR EXISTING SYSTEMS

Text summarization is the process of generating ashort version of a given text to indicate its main top-ics. As the number of documents on the web expo-nentially increases, text summarization has attractedincreasing attention, because it can help people get the most important information within a short time.In most of the existing summarization systems,people need to first define a constant length to restrict all the output summaries.

However, in manycases it is improper to require all summaries are of the same length. Take the multi-document summarization as an example, generating the summaries of the same length for a 5-document cluster and a 50-document cluster is intuitively improper. Morespecifically, consider two different clusters of documents: one cluster contains very similar articles which all focus on the same event at the same time; the other contains different steps of the event but each step has its own topics. The former cluster may need only one or two sentences to explain its information, while the latter needs to include more.

Research on summary length dates back in the late 90s. Goldstein et al. (1999) studied the characteristics of a good summary (single-document summarization for news) and showed an empirical distribution of summary length over document size. However, the length problem has been gradually ignored later, since researchers need to fix the length so as to estimate different summarization models conveniently. A typical instance is the Document Understanding Conferences (DUC)1, which provide authoritative evaluation for summarization systems. The DUC conferences collect news aritcles as the input data and define various summarization tasks, such as generic multi-document summarization, query-focused summarization and update summarization.

Document clustering and multi-document summarization are two fundamentaltools for understandingdocumentdata. Probabilistic Latent Semantic Analysis is a widely used method for document clustering due to the simplicity of the formulation, and efficiency of its EM style computational al-gorithm. The simplicity makes it easy to incorporate PLSA into other machine learning formulations. There are many further developments of PLSA,

such as Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and other topic models see review articles (Steyvers and Griffiths 2007; Blei and Lafferty2009). Theessential formulationofPLSA is the expansion of the co-occurrence probability P(word,doc) into a latent class variable z that separates word distributions from the document distributions given latent class. However, as it is currently formulated, PLSA strictly requires the number of word latent classes to be equal to the num ber of document latent classes (i.e., there is a one-to-one correspondence between word clusters and document clusters). In practical applications, however, this strict requirement may not be satisfied since if we consider documents and words as two different types of objects, they may have their own cluster structures, which are not necessarily same, though related

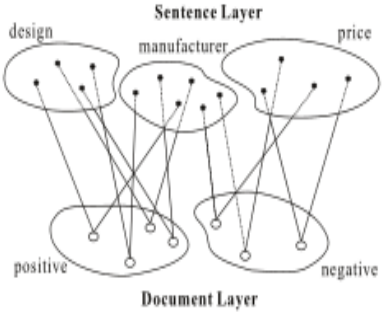


Figure 1: An example showing different cluster structures of documents and sentences.

.

III. PROPOSED SYSTEM

- Probabilistic Hidden Semantic Analysis has been popularly used in document analysis.
- To propose Bi-mixture Probabilistic Hidden Semantic Analysis , a new formulation of PHSA that allows the number of latent word classes to be different from the number of latent document classes.
- Extended method of Bi-PHSA Bi-mixture PHSA with sentence bases (Bi-PHSAS) to simultaneously cluster and summarize the documents utilizing the mutual influence of the document clustering and summarization procedures.
- Additionally propose a Bayesian nonparametric model for multidocument summarization in order to determine the proper lengths of sum maries.

COMPARISION OF EXISTING SYSTEM AND PROPOSED SYSTEM

In Phase 1 only to divide the sentences in documents into discriminative words and non discriminative words for form the clusters.In Phase 2 to perform the clustering the documents and summarize the words which is present in the documents.Additionaly length of the summarize the sentences also find.Comparing phase 1 approach to phase 2 proposed approach provide better result in clustering side and summarize the multidocument side also.

IV. MODULES

There are seven modules in that process of proposed system

- Document Preprocessing
- Identify Discrimination words
- Sentence filtering
- Summarization of sentences
- Clustering process
- Summarized text
- Identification of Summary length

Document Preprocessing

Document preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preparatory data mining practice, document preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Document pre-processing is an often neglected but it is the important step in the data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning projects. Data preparation and filtering steps can take considerable amount of processing time. Document pre-processing includes stop word elimination, stemming, etc.

Identify Discrimination words

This module is designed to fetch word category values based on the topics covered. Each document is represented by a large amount of words including discriminative words and non discriminative words. Only discriminative words are useful for grouping the documents. The discriminative words are the related to the concepts of the documents available in the process. Whereas the non discriminative words are the additional words which are deviating the concepts, i.e. non relevant words of the document. The discriminating and non discriminating words are determined with the help of the Variational Inference Algorithm. For the algorithm of variational inference, it could be applied to infer the document collection structure in a much quicker manner

Sentence filtering

A concept-based similarity measure depends on matching concept at sentence, document, and corpus instead of individual terms. This similarity measure based on three main aspects. First is analyzed label terms that capture semantic structure of each sentence. Second is concept frequency that is used to measure participation of concept in sentence as well as document. Last is the concepts measured from number of documents. A raw document with well defined sentence boundaries is given as input to the proposed system. According to the Semantic analysis, each of the sentences in the document is labeledautomatically. The sentences in the document may have one or more labeled structures.The objective behind the concept-based mechanism is to achieve an accurate analysis of concepts on thE sentence, document, and corpus levels rather than document only.

Clustering process

Clustering is a automatic document organization, topic extraction and fast information retrieval or filtering. It is closely related to data clustering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information.

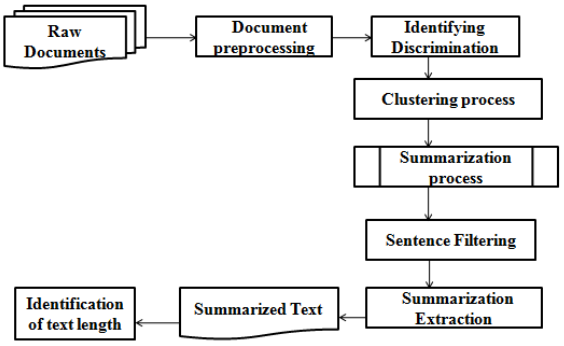


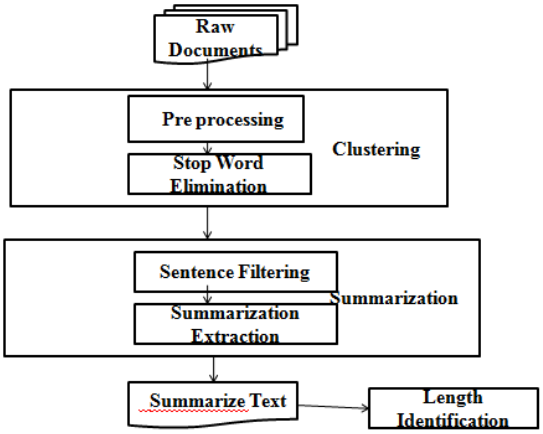
Fig 2:Architecture diagram

Text document clustering gives an important role in providing worthy document retrieval, document browsing, and text document mining. Traditionally, clustering techniques do not consider the semantic relationships between words, such as synonymy, meronym and hypernymy. To utilize semantic relationships, ontology's have been used to improve clustering results.

Identification of Summary length

Propose a Bayesian nonparametric model for multidocument summarization in order to automatically determine the proper lengths of summaries.

MODULE DIAGRAM



V. ALGORITHM USED IN PROPOSED SYSTEM

Bi-mixture PLSA with Sentence Bases

We extend Bi-PLSA to incorporate sentence information. The advantage of sentences over words is thatsentences are more readable, e.g. in extractive summarization methods they are directly used as a summary,while nontrivial extra work is needed to interpret the word clusters, particularly in the form of unigram distributions

Clustering and Summarization via Bi-PLSAS

Once we obtain the parameters p(s|zs), p(d|zd) and p(zd,zs) in the Bi-PLSAS model , we can easily cluster the documents and sentences, and generate the summary**.**

$$\begin{aligned} z(d)^* &= \arg\max_{z_d} p(z_d|d) \\ &= \arg\max_{z_d} p(z_d, d) \\ &= \arg\max_{z_d} p(d|z_d) \sum_{z_s} p(z_d, z_s). \end{aligned}$$

Similarly, the cluster membership of a sentence s can be derived using

$$z(s)^* = \arg\max_{z_s} p(s|z_s) \sum_{z_d} p(z_d, z_s)$$

Summarization: To generate a summary for the document collection, first, the marginal probability of every sentence cluster zsis calculated as zdp(zs,zd), and those clusters with small marginall probability values are removed.Then, the sentences are extracted from the remaining sentence clusters based on p(s|zs).

BNP SUMMARIZATION

Most existing approaches for generic extractive summarization are based on sentence ranking. However, these methods suffer from a severe problem that they cannot make a good trade-off between the coverage and minimum redundancy. Some global optimization algorithms are developed, instead of greedy search, to select the best.One approach to global optimization of summarization is to regard the summarization as a reconstruction process

$$\begin{aligned} x_i &= S(\phi_i \circ z_i) + \epsilon_i \\ S &= [s_1, s_2, ..., s_N] \\ z_{ij} &\sim Bernoulli(\pi_j) \\ \pi_j &\sim Beta(\frac{\alpha\gamma}{N}, \alpha(1 - \frac{\gamma}{N})) \\ \phi_i &\sim \mathcal{N}(0, \sigma_\phi^2 I) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2 I) \end{aligned}$$

where N is the number of sentences in the whole document cluster. The symbol ◦ represents the ele-mentwise multiplication of two vectors.

VI.RESULTS

Comparison for Different Numbers of Sentence/Word Clusters

We first compare the proposed Bi-PLSA and Bi-PLSAS with the baselines:PLSA and FGB. To show the ef ect of two hidden class variables, we evaluate the perfor-mance of Bi-PLSA and Bi-PLSAS with different numbersof word/sentence clusters1.From the, we can see

1) Bi-PLSA outperforms PLSA form o s numbers ofword clusters, since theBi-PLS Ahas freedom to set different value from the number of documentclusters;
2) The better performanceof FGB than PLSA demonstrates the effectiveness of sentence bases in docu-ment clustering;
3) Bi-PLSAS combines the advantage of FGB and Bi-PLSA, and performs the best among all the methods.

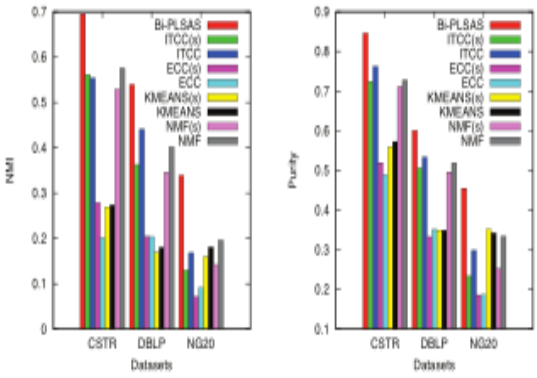


Fig 4: Comparision with co-clustering methods

Experiments with summarization length results

The Random model selects sentences randomly for each document cluster.

- The MMR strives to reduce redundancy while maintaining relevance. For generic summarization, we replace the query relevance with the relevance to documents.
- The Lexrank model is a graph-based method which choose sentences based on the concept of eigenvector centrality.
- The Linear Representation model has the same assumption as ours and it can be seen as an approximation of the constant-length version of our model.

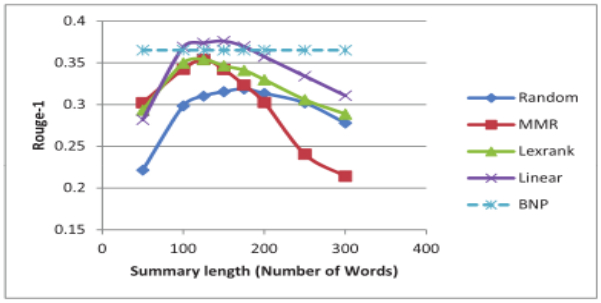


Figure 1: Rouge-1 values on DUC2004 dataset.

Turn to the length determination. We take advantage of the Linear Representation model to approximate the constant-length version of our model.Comparing the summaries generated at different predefined lengths.

In future we may consider more human factors, and prove the summary length determined by our system agrees with human preference. In addition, in the experiments,we only use the imbalanced datasets as the example that intuitively needs varying the summary length.

VI. CONCLUSION

Find a new problem of exploring a proper summary length for multi-document summarization based on the document content. A Bayesian nonparametric model is proposed to solve this problem. We use the beta processas the prior to construct a Bayesian framework for summary sentence selection.Additionaly find the problem we propose a new formulation of PLSA to incorporate the sentence information, allowing the number of latent sentence classes to be different from the number of latent document classes.

In future, we may extend thework by studying more cases that need varying sum mary length.

REFERENCES

1. Akaike, H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19(6):716–723.
2. Blei, D., and Lafferty, J. 2009. Topic models. Text mining: classification, clustering, and applications 71.
3. Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. The Journal of Machine Learning Research.
4. Cho, H.; Dhillon, I.; Guan, Y.; and Sra, S. 2004. Minimum sum-squared residue co-clustering of gene expression data. In SDM, 114–125.
5. Dhillon, I.; Mallela, S.; and Modha, D. 2003. Information-theorticco-clustering. In SIGKDD.
6. Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In SIGKDD.
7. Ding, C.; Li, T.; and Peng, W. 2006. Nonnegative matrix factorization and probabilistic latent.
8. K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchel, “Text Classification from Labeled and Unlabeled Documents
9. Using Em,” J. Machine Learning, vol. 39, no. 2, pp. 103-134, 2000.
10. G. Yu, R. Huang, and Z. Wang, “Document Clustering via Dirichlet Process Mixture Model with Feature Selection,” Proc.
11. ACM Int’l Conf. Knowledge Discovery and Data Mining, pp. 763-772, 2010.
12. D. Blei and M. Jordan, “Variational Inference for Dirichlet Process Mixtures,” Bayesian Analysis, vol. 1, no. 1, pp. 121-144, 2006.
13. Zhanying He, Chun Chen, Jiajun Bu, CanWang, Lijun Zhang, Deng Cai and Xiaofei He. 2012. Document Summarization Based on Data Reconstruction. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence.
14. Michael Kaisser, Marti A. Hearst, John B. Lowe. 2008.Improving Search Results Quality by Customizing Summary Lengths. Proceedings of ACL-08: HLT,pages 701-709.
15. Chin-YewLin, GuihongCao, JianfengGao, andJian-YunNie. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. Proceedings ofNAACL2006, pages 463-470.