IJDCST @Oct-Nov-2015, Issue- V-3, I-7, SW-14
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Summarization and Classification for document using Concept and Context Similarity Analysis

Sandhya. Ekkurthi[1], Chilakalapudi Meher Babu[2]

[1] **Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Vatticherukur, Prathipadu Road, Guntur, Andhra Pradesh.**

[2]**Assistant. prof, Malineni Lakshmaiah Women's Engineering College, Pulladigunta, Vatticherukur, Prathipadu Road, Guntur, Andhra Pradesh.**

Abstract: The document summarization mostly use the similarity between sentences in the document to extract the most salient sentences. The documents as well as the sentences are indexed using traditional term indexing measures, which do not take the context into consideration. Therefore, the sentence similarity values remain independent of the context. In this paper, we propose a context sensitive document indexing model based on the Bernoulli model of randomness. The Bernoulli model of randomness has been used to find the probability of the co occurrences of two terms in a large corpus. A new approach using the lexical association between terms to give a context sensitive weight to the document terms has been proposed. The resulting indexing weights are used to compute the sentence similarity matrix. The proposed sentence similarity measure has been used with the baseline graph-based ranking models for sentence extraction. Experiments have been conducted over the benchmark DUC data sets and it has been shown that the proposed Bernoulli-based sentence similarity model provides consistent improvements over the baseline Intra Link and Uniform Link methods.

Keywords: Similarity sentence, bemoulli model, DUC.

Introduction:

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data Mining is the computer-assisted process which analyzes enormous set of data and extracts the meaning of data. Data Mining refers to extracting or “mining” knowledge from large amounts. Such as knowledge mining from data, knowledge extraction, Many people treat data mining for another popularly used term, Knowledge Discovery from Data, or KDD. Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques , data mining helps to recognize significant facts, relationships, trends, patterns and anomalies that might be unnoticed. That technique that is used to perform these feats is called modeling. Modeling is simply the act of building a model based on data from situations. Modelling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect. These

store huge amounts of data, and the computational power to automate modelling techniques to work directly on the data, have been available. Some of the tools used for data mining are:

Artificial neural networks - Predictive models that learn through training and resemble biological neural networks in structure.

Decision trees - Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Rule induction - The extraction of useful if-then rules from data based on statistical significance.

Genetic algorithms - Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.

Nearest neighbour - A classification technique that classifies each record based on the records most similar to it in an historical database. In the short-term, the results of data mining will be in profitable, on business related areas. Advertising will target potential customers with new precision.

In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers. Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe.

DOCUMENT summarization is an information retrieval task, which aims at extracting a condensed version of the original document [2]. A document summary is useful since it can give an overview of the original document in a shorter period of time. Readers may decide whether or not to read the complete document after going through the summary. For example, readers first look at the abstract of a scientific article before reading the complete paper. Search engines also use text summaries to help users make relevance decisions [3].The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents [4]. The task of producing summary from many documents is called multi document Summarization[5],[6],[7],[8],[9],[10].Summarization can also be specific to the information needs of the user, thus called “query-biased” summarization [11], [12], [13].For instance, the QCS system (query, cluster, and summarize, [12]) retrieves relevant documents in response to a query, clusters these documents by topic and produces a summary for each cluster. Opinion summarization [14], [15], [16], [17] is another application of text summarization. Topic summarization deals with the evolution of topics in addition to providing the informative sentences [18].

This paper focuses on sentence extraction-based single document summarization. Most of the previous studies on the sentence extraction-based text summarization task use a graph-based algorithm to calculate the saliency of each sentence in a document and the most salient sentences are extracted to build the document summary. The sentence extraction techniques give an indexing weight to the document terms and use these weights to compute the sentence similarity [1] and/or document centroid [19] and so

on. The sentence similarity calculation remains central to the existing approaches. The indexing weights of the document terms are utilized to compute the sentence similarity values. However, very elementary document features are used to allocate an indexing weight to the document terms, which include the term frequency, document length, occurrence of a term in a background corpus and so on. Therefore, the indexing weight remains independent of the other terms appearing in the document and the context in which the term occurs is overlooked in assigning its indexing weight. This results in "context independent document indexing." To the authors' knowledge, no other work in the existing literature addresses the problem of "context independent document indexing" for the document summarization task.

2. Existing System

Existing models for document summarization mostly use the similarity between sentences in the document to extract the most salient sentences. The documents as well as the sentences are indexed using traditional term indexing measures, which do not take the context into consideration. Existing literature addresses the problem of context independent document indexing for the document summarization task. Problem of retrieves relevant documents in response to a query, clusters these documents by topic and produces a summary for each cluster. Does not provide tool for reading news geographically. News articles available on the applications will be provided from same resources. All the contents are static. User article of cannot get the exact what they want to read.

iii. Proposed System:

The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Summaries can be produced either from a single document or many documents [4]. The task of producing summary from many documents is called multi document summarization [6], [10]. Summarization can also be specific to the information needs of the user, thus called "query-biased" summarization. For instance, the QCS system (query, cluster, and summarize) retrieves relevant documents in response to a query, clusters these documents by topic and produces a summary for each cluster. Opinion summarization [7] is another application of text summarization. Topic summarization deals with the evolution of topics in addition to providing the informative sentences [8].

This paper focuses on sentence extraction-based single document summarization. Most of the previous studies on the sentence extraction-based text summarization task use a graph-based algorithm to calculate the saliency of each sentence in a document and the most salient sentences are extracted to build the document summary. The sentence extraction techniques give an indexing weight to the document terms and use these weights to compute the sentence similarity [1] and/or document centroid and so on. The sentence similarity calculation remains central to the existing approaches. The indexing weights of the document terms are utilized to compute the sentence similarity values. elementary document features are used to allocate an indexing weight to the document

terms, which include the term frequency, document length, occurrence of a term in a background corpus and so on. Therefore, the indexing weight remains independent of the other terms appearing in the document and the context in which the term occurs is overlooked in assigning its indexing weight. This results in "context independent document indexing." To the authors' knowledge, no other work in the existing literature addresses the problem of "context independent document indexing" for the document summarization task.

A document contains both the content-carrying terms as well as background terms. The traditional indexing schemes cannot distinguish between these terms that are reflected in the sentence similarity values. A context sensitive document indexing model gives a higher weight to the topical terms as compared to the nontopical terms and, thus, influences the sentence similarity values in a positive manner.

The system considers the problem of "context independent document indexing" using the lexical association between document terms. In a document, the content carrying words will be highly associated with each other, while the background terms will have very low association with the other terms in the document. The association between terms is captured in this paper by the lexical association, computed through a corpus analysis.

The main motivation behind using the lexical association is the central assumption that the context in which a word appears provides useful information about its meaning. Cooccurrence measures observe the distributional patterns of a term with other terms in the vocabulary and have applications in many tasks pertaining to natural language understanding such as word classification, knowledge acquisition, word sense disambiguation, information retrieval [2], sentence retrieval and word clustering. In this paper, we derive a novel term association metric using the Bernoulli model of randomness. Multivariate Bernoulli models have previously been applied to document indexing and information retrieval. We use the Bernoulli model of randomness to find the probability of the cooccurrences of two terms in a corpus and use the classical semantic information theory to quantify the information contained in the cooccurrences of these two terms.

The lexical association metric, thus, derived is used to propose a context-sensitive document indexing model. The idea is implemented using a PageRank-based algorithm is applied to iteratively compute how informative is each document term. Sentence similarity calculated using the context sensitive indexing should reflect the contextual similarity between two sentences [9]. This will allow two sentences to have different similarity values depending on the context. The hypothesis is that an improved sentence similarity measure would lead to improvements in the document summarization.

iv. Sentence similarity and word indexing:

Bernoulli Model of Randomness:

By using the PMI measure the lexical association between documents terms is higher than between the summary terms. Therefore, the PMI measure may not be a suitable choice for the possible application in document summarization. Using the MI and Bernoulli measure, on the other hand, the average

lexical association between the terms in human summary is higher than that in the original document. As verified by the two different statistical tests, the difference is statistically significant using both these association measures and therefore, the hypothesis holds true for both the MI and Bernoulli measures. However, the significance level as well as the ratio of average lexical association between the target summary and original document is much higher for the Bernoulli measure as compared to the MI measure. Thus, the proposed Bernoulli measure is a better fit for H2.

Context-Based Word Indexing:

Given the lexical association measure between two terms in a document from hypothesis H2, the next task is to calculate the context sensitive indexing weight of each term in a document using hypothesis H3. A graph-based iterative algorithm is used to find the context sensitive indexing weight of each term. Given a document Di , a document graph G is built. Let G = (V,E) be an undirected graph to reflect the relationships between the terms in the document Di . V = {vj |1 ≤ j ≤_ |V|} denotes the set of vertices, where each vertex is a term appearing in the document. E is a matrix of dimensions |V| × |V|. Each edge ejk ε E corresponds to the lexical association value between the terms corresponding to the vertices vj and vk. The lexical association between the same term is set to 0.

v. System Model:

The document indexing and summarization scheme is enhanced with semantic analysis mechanism. Context sensitive index model is improved with semantic weight values. Concept relationship based lexical association measure estimation is performed for index process. Bernoulli lexical association measure is used to perform the document classification process.

The document summarization system is enhanced with document classification process. Concept relationship based semantic weight estimation mechanism is used for document relationship analysis. Ontology based semantic index scheme is used to perform the classification process. The system is divided into five major modules. They are document preprocess, term index process, semantic index process, document summarization, document classification.

The document preprocess module is designed to perform token separation and frequency estimation process. Term indexing process module is designed to estimation term weights and index process. Concept relationship is analyzed under semantic index process. Document summarization module is designed to prepare document summary. Document category assignment is performed under the document classification process.

Document Preprocess:

The document preprocess is performed to parse the documents into tokens. Stop word elimination process is applied to remove irrelevant terms. Stemming process is applied to carry term suffix analysis. Document vector is constructed with terms and their count values.

Term Index Process:

Statistical weight estimation process is applied with term and its count values. Term weight estimation is performed with Term Frequency (TF) and Inverse Document Frequency (IDF) values. Context sensitive index model uses the term weights for term index process. Latent semantic analysis is applied to estimate relationship values.

Semantic Index Process:

Ontology is a repository that maintains the concept term relationships. Semantic weights are estimated using concept relations. Synonym, hypernym and meronym relationships are used in the concept analysis. Context sensitive index model uses the semantic weight values for index process.

Document Summarization:

Lexical association between terms is used to produce context sensitive weight. Weights are used to compute the sentence similarity matrix. The sentence similarity measure is used with the baseline graph-based ranking models for sentence extraction. Document summary is prepared with sentence similarity values.

Document Classification:

Document classification is carried out to assign document category values. Term weight and semantic weights are used for the classification process. Context sensitive index is used for the document classification process. Sentence similarity is used in classification process.

Experiment result:

The document summarization system is developed to prepare the summary for the text documents. Term weight based context sensitive index and semantic weight based context sensitive index models are used for the document summarization process. Document contents are preprocessed and sentence based similarity analysis is performed to estimate the context sensitive index values. Most important sentences are summarized with reference to the index values. The system is tested with different document count and weight schemes. The Context Sensitive Index with Term weights (CSIT) and Context Sensitive Index with Semantic weights (CSIS) schemes are used for the summarization process. The results are table. The results show that CSIS model improves the summarization accuracy 10% than the CSIT scheme.

Documents	CSIT	CSIS
50	72	84
100	76	87
150	79	89
200	81	91
250	84	93

Conclusion:

In this paper Document summarization methods are used to extract the condensed version of the original document. Document classification methods are used to assign the category of the documents Bernoulli model of randomness is used for document summarization process. The Bernoulli model of randomness is used to find the probability of the

cooccurrences of two terms in a large corpus. The lexical association between terms is used to produce a context sensitive weight to the document terms. The document indexing and summarization scheme is enhanced with semantic analysis mechanism. Context sensitive index model is improved with semantic weight values. Concept relationship based lexical association measure estimation is performed for index process. Bernoulli lexical association measure is used to perform the document classification process. The Java language and Oracle relational database are used for the system development process. The proposed model gives higher weight tothe content-carrying terms and as a result, the sentences are presented in such a way that the most informative sentences appear on the top of the summary, making a positiveimpact on the quality of the summary.

REFERENCES:

[1] X. Wan and J. Xiao, "Exploiting Neighborhood Knowledge for Single Document Summarization and Keyphrase Extraction," ACM Trans. Information Systems, vol. 28, pp. 8:1-8:34, http://doiacm.org/10.1145/1740592.1740596, June 2010.

[2] P. Goyal, L. Behera, and T. McGinnity, "Query Representation Through Lexical Assoc. for Information Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 12, pp. 2260-2273, Dec. 2011.

[3] L.L. Bando, F. Scholer, and A. Turpin, "Constructing QueryBiased Summaries: A Comparison of Human and System Generated Snippets," Proc. Third Symp. Information Interaction in Context, pp. 195-204, http://doi.acm.org/10.1145/18407841840813 ,2010.

[4] X. Wan, "Towards a Unified Approach to Simultaneous SingleDocument and Multi-Document Summarizations," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 1137-1145, http://portal.acm.org/citation.cfm?id=1873781. 1873909 , 2010.

[5] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, "A Study on Position Information in Document Summarization," Proc. 23rd Int'l Conf. Computational Linguistics: Posters, pp. 919-927, http://portal.acmorg/citation.cfm?id=1944566.1944672, 2010.

[6] Q.L. Israel, H. Han, and I.-Y. Song, "Focused Multi-Document Summarization: Human Summarization Activity vs. Automated Systems Techniques," J. Computing Sciences in Colleges, vol. 25, pp. 10-20, http://portal.acm.org/citation.cfm?id=1747137 1747140, May 2010.

[7] H. Nishikawa, T. Hasegawa, Y. Matsuo, and G. Kikui, "Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering," Proc. 23rd Int'l Conf. Computational Linguistics: Posters, pp.910-918., http://portal.acm.org/citation.cfm?id=1944566.1944671, 2010.

[8] C.C. Chen and M.C. Chen, "TSCAN: A Content Anatomy Approach to Temporal Topic Summarization," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 1, pp. 170-183, Jan. 2012.

[9] Pawan Goyal, Laxmidhar Behera, and Thomas Martin McGinnity, "A Context-Based Word Indexing Model for Document Summarization", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, August 2013.

[10] S. Harabagiu and F. Lacatusu, "Using Topic Themes for MultiDocument Summarization," ACM Trans. Information Systems, vol. 28, pp. 13:1-13:47, http://doi.acm.org/10.1145/1777432.1777436, July 2010.