International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

An Algorithm of Word Indexing Model for Document Summarization based on Perspective of Document

Meha Shah[1], Chetna Chand[2]

[1]Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

[2]Assistant Professor, Gujarat Technological University, Kalol Institute of Engineering, Kalol, Gujarat, India

Abstract: ***Natural language processing (NLP) is an area of computer science, artificial intelligence, and computational linguistics connected with the communications between computers and natural languages. There are many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Document summarization is a part of it. Many different classes of such process based on machine learning are developed. In researches earlier document summarization mostly use the similarity between sentences in the document to extract the most significant 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: Data mining, Document Summarization, Text mining, Stemming, Sentence Similarity, Context Similarity.

1. Introduction

Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. Generally, in text mining techniques, the phrase frequency of a term is calculated to discover the consequence of the term in the document. Nevertheless, two terms could have similar frequency in the given documents, but one term gives more to the denotation of its sentences, compared to the other term. Clustering, one of the conventional data mining strategies is an unsubstantiated knowledge pattern. Here clustering methods endeavor to recognize intrinsic alignments of the text documents, so that a set of clusters is formed in which clusters display high intra-cluster likeness and low inter-cluster likeness. Normally, text document clustering endeavors to separate out the documents into groups where every group characterizes some subject that is different from the topics characterized by the other groups

2. Overview

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.

Bernoulli Process of Randomness

a Bernoulli process is a finite or infinite sequence of binary random variables, so it is a discrete-time stochastic process that takes only two values, canonically 0 and 1. The two possible values of each Xi are often called "success" and "failure". Thus, when expressed as a number 0 or 1, the outcome may be called the number of successes on the ith "trial".

We approach the system on the base of similarities computed using multiple sets of features: (a) naive lexical features, (b) similarity between character representations of sentences, and (c) similarity between constituent words computed using WordNet, using the eigenword vector representations of words , and using selectors, which generalize words to a set of words that appear in the same context.

The generation process of a summary basically can be either abstractive or extractive [12].In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document is generated, whereas in extractive summarization, the summary is usually formed by selecting salient sentences from the original document. The former requires more sophisticated natural language processing (NLP) techniques, including semantic representation and inference, as well as natural language generation, while this would make abstractive approaches difficult to replicate or extend from constrained domain more general domains. Apart from being abstractive or extractive, a summary may also be generated by considering several other aspects like being generic or

query-oriented summarization, single-document or multi-document summarization, among others. In this paper, we focus exclusively on generic, extractive summarization of spoken documents, since it usually constitutes the essential building block for many other speech summarization tasks.

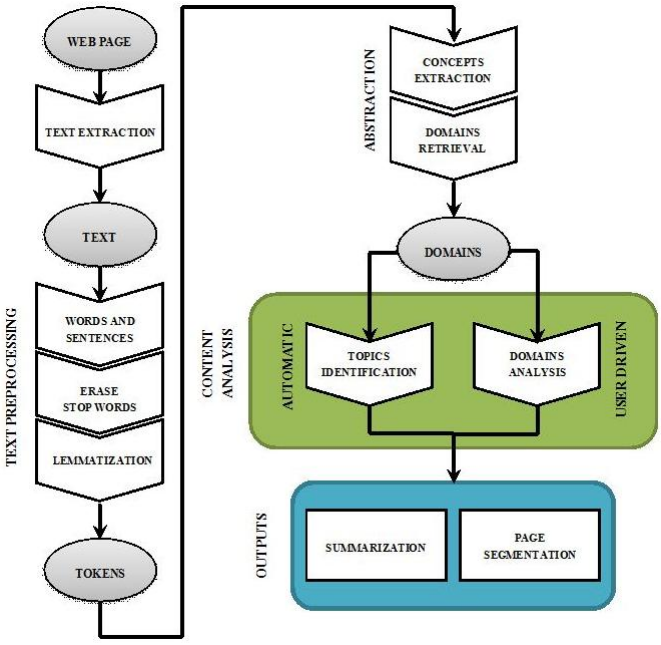

Figure 1: Flow of data summarization process

3. Background Theory

3.1 Document Summarization

Eigenword Vector Summarization

An *Eigenword* is a real-valued vector "embedding" associated with a word that captures its meaning in the sense that distributional similar words have similar eigenword. This page contains links to several sets of eigenword They are computed as the singular vectors of the matrix of co-occurrence of words and their contexts, and used in a variety of spectral NLP methods and applications

Each sentence feature has its unique Contribution and combing them would be advantageous. Therefore we investigate combined sentence features for extractive summarization. [2] Currently, most successful multi-document summarization systems [5] follow the extractive summarization framework. These systems first rank all the sentences in the original document set and then select the most salient sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries, some intensive post-processing approaches are also appended on the extracted sentences. Two Summary Construction Methods are applied first one is Abstractive method where summaries produce generated text from the important parts of the documents and second is Extractive Method where summaries identify important sections of the text and use them in the summary as they are.

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 document length, term frequency, occurrence of a term in a background corpus. Therefore, the indexing weight of the other terms appearing in the document remains independent and the context in which the term occurs is overlooked in assigning its indexing weight for the documents. 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 background terms as well as the content-carrying terms. In the sentence similarity analysis the traditional indexing schemes cannot distinguish between these terms. The higher weight is given by the context sensitive document indexing model to the topical terms where it is compared with the non topical terms and thus influences the sentence similarity values in a positive manner. Using the lexical association between document terms the system considers the problem of "context independent document indexing. The content carrying words will be highly associated with each other in a document, while the background terms will have very low in association with the other terms in the document. The association between terms is stated in this paper by the lexical association and is computed through the corpus analysis.

3.2 Word Indexing

- **Sentence Similarity based**

Sentence similarity assessment is key to most NLP applications. This paper presents a means of calculating the similarity between very short texts and sentences without using an external corpus of literature. This method uses WordNet, common-sense knowledge base and human intuition. Results were verified through experiments. These experiments were performed on two sets of selected sentence pairs. We show that this technique compares favorably to other word-based similarity measures and is flexible enough to allow the user to make comparisons without any additional dictionary or corpus information. We believe that this method can be applied in a variety of text knowledge representation and discovery applications.

- **Context based**

$$\textbf{do}\begin{cases} E \leftarrow 0 \\ \textbf{for} j \leftarrow 1 \textbf{to} |S| \\ \quad \textbf{do}\begin{cases} memoWt[v_j] \leftarrow indexWt[v_j] \\ indexWt[v_j] \leftarrow \mu \cdot \sum_{\forall k \neq j} indexWt[v_k] \cdot \tilde{E_{kj}} \\ + \frac{1-\mu}{|V|} \\ E \leftarrow E + (indexWt[v_j] - memoWt[v_j])^2 \end{cases} \\ E \leftarrow \sqrt{E} \end{cases}$$

$$\textbf{return } indexWt$$

Figure 2: Context Base word Indexing equation

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 terms is set to 0.

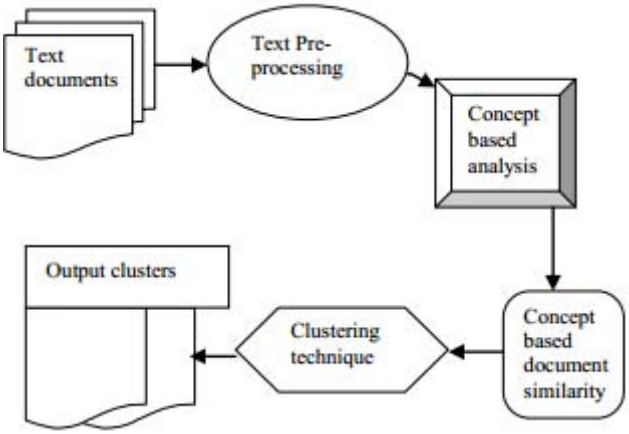

Figure 3: model for concept based analysis of data

The process has involved the above stated steps. Basically they all have one ir the other conceptual technique based on text mining and data mining. We are proposing to use Bernoulli morel and context based similarity indexing for words because the process does not take much time and become efficient than the earlier one.

4. Preprocessing Stage

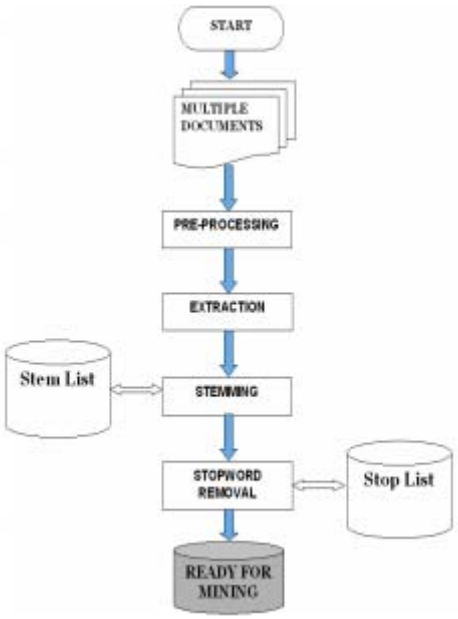

Figure 4: Document Preprocessing

This process consists of the following stages. First the data should pass the process of lexical analysis. It is a process of comparing all the terms with its lexical equivalents and whichever is found match, will be processed for the next stage. Elimination of stop words is the next stage that removes all the text and characters like dot, comma and question marks etc. Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Our system uses the SMART stop word list.[4] Stemming is the next process towards accurate result processing. Stemming means elimination of synonyms and related words. The basic function of both the methods – stemming and lemmatizing is similar. Both of them reduce a word variant to its „stem" in stemming and „lemma" in lemmatizing. There is a very subtle difference between both the concepts. In stemming the „stem" is obtaining after applying a set of rules but without bothering about the part of speech (POS) or the context of the word occurrence. In contrast, lemmatizing deals with obtaining the „lemma" of a word which involves reducing the word forms to its root form after understanding the POS and the context of the word in the given sentence.

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.

5. Proposed Algorithm

5.1 Algorithm

Input : copy and paste the data that has to be processed
Expected Output: Summary of documents
Step 1: Start
Step 2: For each Document
Step 3: Stopping method to remove additional symbols
Step 4: Stemming method to group similar meaning words
Step 5: Using Stemming algorithm removes blank space and extract keywords sentences using wordnet dictionary
End For each
Step 6: For each generated output using stemming algorithm
Sentences index generated by Bernoulli model of randomness.
Context based sentense similarity indexing
Now, Use Context based word indexing on the generated output to create the summarization of text is achieved using equation given in figure : 2.
End for each
Go to next file and repeat above algorithm;
end

5.2 Expected Outcome

Document summarization after processing with random index weight generated by the model and processed with algorithm.

5.3 Performance evaluation

- Comparison of levels of summarization of various documents.
- Demonstrating it in the form of graph
- Preparing a table of analyzed data to show various results

References

[1] Cormen, T. R., C. E. Leiserson, and R. L. Rivest. 1989. Introduction to Algorithms. The MIT Press.
[2] Davison, A. C. and D. V. Hinkley. 1997. Bootstrap Methods and Their Application. Cambridge University Press.
[3] Lin, C.-Y. and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada.
[4] Lin, C.-Y. 2004. Looking for a few good metrics: ROUGE and its evaluation. In Proceedings of NTCIR Workshop 2004, Tokyo, Japan.
[5] Lin, C.-Y. and F. J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of 42nd Annual Meeting of ACL (ACL 2004), Barcelona, Spain.
[6] Mani, I. 2001. Automatic Summarization. John Benjamins Publishing Co.
[7] Melamed, I. D. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In Proceedings of the 3rd Workshop on Very Large Corpora (WVLC3). Boston, U.S.A.
[8] Melamed, I. D., R. Green and J. P. Turian (2003). Precision and recall of machine translation. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada.
[9] Over, P. and J. Yen. 2003. An introduction to DUC 2003 – Intrinsic evaluation of generic news text summarization systems.
[10] A Context-Based Word Indexing Model for Document Summarization Pawan Goyal, Laxmidhar Behera,Senior Member, IEEE, and Thomas Martin McGinnity,Senior Member, IEEE
[11] Context-Based Similarity Analysis for Document Summarization 1 S.Prabha, 2 Dr.K.Duraiswamy, 3 B.Priyanga 1 Associate Professor, Department of Information technology K.S.Rangasamy College of Technology, Tiruchengode – 637215, Tamil Nadu, India 2 Dean Academic
[12] Document Summarization and Classification using Concept and Context Similarity Analysis J.Arun 1, C. Gunavathi M.E 2 PG scholar, K.S.Rangasamy College of technology, Tiruchengode, Tamil Nadu, India
[13] A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model V.M.Navaneethakumar 1, Dr.C.Chandrasekar 2 1 Assistant Professor, Department of Computer Applications, K.S.R College of Engineering, Tiruchengode, Tamilnadu, India
[14] SYSTEM FOR DOCUMENT SUMMARIZATION USING GRAPHS IN TEXT MINING Prashant D. Joshi 1, M. S. Bewoor 2, S. H. Patil 3 1 Deptt. of Computer Engineering, Researcher, BharatiVidyapeeth University, Pune, India.
[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," technical report, Stanford Digital Library Technologies Project, http://citeseer.ist.psu.edu/page98pagerank.html, 1998.
[16] J. Turner and E. Charniak, "Supervised and Unsupervised Learning for Sentence Compression,"Proc. 43rd Ann. Meeting on Assoc. for Computational Linguistics,pp. 290-297, http://dx. doi.org/10.3115/1219840.1219876, 2005.
[17] J. Clarke and M. Lapata, "Discourse Constraints for Document Compression," Computational Linguistics,vol. 36, pp. 411-441, 2010.