

Document Summarization and Classification using Concept and Context Similarity Analysis

J.Arun¹, C. Gunavathi M.E²PG scholar, K.S.Rangasamy College of technology, Tiruchengode, Tamil Nadu, India¹Assistant Professor, K.S.Rangasamy College of technology, Tiruchengode, Tamil Nadu, India²

ABSTRACT: “Document summarization and classification using concept and context similarity analysis” deals with an information retrieval task, which aims at extracting a condensed version of the original document. A document summary is useful since it can give an overview of the original document in a shorter period of time. The main goal of a summary is to present the main ideas in a document/set of documents in a short and readable paragraph. Classification is a data mining function that assigns items in a collection to target categories of the documents. Context sensitive document indexing model based on the Bernoulli model of randomness is used for document summarization process. The lexical association between terms is used to produce a context sensitive weight to the document terms. The context sensitive indexing weights are used to compute the sentence similarity matrix 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 positive impact on the quality of the summary

Keywords: Document indexing, Lexical association, Bernoulli model of randomness

I. INTRODUCTION

Modern text retrieval systems principally rely on orthographic, semantic, and statistical analysis. The usual approach is to use white space to identify word boundaries, followed by stemming to conflate words with similar surface forms into a common term. A weight is then computed for each term in every document using the frequency of the term in the document, the selectivity of the term, and the length of the document. In vector space text retrieval, queries are represented in a manner similar to the documents, and the similarity of each document in the collection to the query is then computed as the normalized inner product of the document and query term weight vectors. In probabilistic text retrieval, a term weight is treated as the probability of relevance of a document to a query,

conditioned on the presence of that term in the query. Probabilistic and vector space techniques are often combined with Boolean text retrieval, in which the presence or absence of a term or combination of terms can be explicitly required in the query specification. The principal advantage of vector space and probabilistic text retrieval over a purely Boolean approach is that lists of documents that are ranked in order of decreasing probability of relevance allow users to interactively decide how many documents are worth examining. Unranked Boolean techniques, on the other hand, might be preferred when no user interaction is possible before the next processing stage. In either case, when the document collection is relatively stable it is common to preprocess the collection to produce an index structure on the feature set that can be searched in sub-linear time.

The utility of a text retrieval system depends strongly on how well the query is constructed, and that depends in turn on how well the user understands the collection and the way in which the indexed features can be used to select documents. It is usually fairly straightforward to find some topically relevant documents, but interactive inspection by the user is generally needed if the relevant documents must be more carefully separated from the irrelevant ones. An iterative query reformulation process such as Simulated Nucleation can be used to speed this process, leveraging inspection of a few documents to produce a query that better separates relevant and irrelevant documents.

Information extraction algorithms can be designed to operate without user interaction, making sharp decisions rather than producing ranked lists of candidates. Results from the Message Understanding Conferences (MUC) suggest that simple tasks such as date extraction and named entity recognition can be done quite accurately, but that more complex tasks in which relationships among data items must also be extracted are still somewhat error-prone. Lawrence, et al. have used information extraction techniques to extract bibliographic citations from gray literature on the World Wide Web.

II. PROBLEM STATEMENT

Text documents are unstructured data elements. Documents are grouped using clustering techniques. Similarity measures are used to analyze the document relationship. Document features are used in the classification process. Sentence similarity measures are used for text summarization process. 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. Therefore, the sentence similarity values remain independent of the context. 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. An 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 sentence similarity measure has been used with the baseline graph-based ranking models for sentence extraction. The following drawbacks are identified from the existing system.

- Term weight based indexing model reduces the accuracy level.
- Term relationships are not considered in the similarity analysis.
- Classification process is not supported by the system.
- Context sensitive indexing model is not tuned for semantic analysis scheme.

In a proposed system contain to take classification process, Bernoulli Model of Randomness, Context-Based Word Indexing, system model are used reduced the drawback of existing system

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.

The text summarization experiments have been performed on the single document summarization task over the DUC01 and DUC02 data sets. It has been shown that the proposed model consistently improves the performance of the baseline sentence extraction algorithms under various settings and, thus, can be used as an enhancement over the baseline models. The theoretical foundations along with the empirical results confirm that the proposed model advances the state of the art in document summarization. The main contributions of the system are summarized as follows:

1. To propose the novel idea of using the context-sensitive document indexing to improve the sentence extraction-based document summarization task.
2. To implement the idea by using the lexical association between document terms in a PageRank-based framework. A novel term association metric using the Bernoulli model of randomness has been derived for this purpose.

Empirical evidence has been provided to show that using the derived lexical association metric, average lexical association between the terms in a target summary is higher compared to the association between the terms in a document.

3. Experiments have been conducted over the benchmark document understanding conference (DUC) data sets to empirically validate the effectiveness of the proposed model.

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 H_2 .

Context-Based Word Indexing

Given the lexical association measure between two terms in a document from hypothesis H_2 , the next task is to calculate the context sensitive indexing weight of each term in a document using hypothesis H_3 . A graph-based iterative algorithm is used to find the context sensitive indexing weight of each term. Given a document D_i , a document graph G is built. Let $G = (V, E)$ be an undirected graph to reflect the relationships between the terms in the document D_i . $V = \{v_j | 1 \leq j \leq |V|\}$ denotes the set of vertices, where each vertex is a term appearing in the document. E is a matrix of dimensions $|V| \times |V|$. Each edge $e_{jk} \in E$ corresponds to the lexical association value between the terms corresponding to the vertices v_j and v_k . The lexical association between the same term is set to 0.

Sentence Similarity Using the Context-Based Indexing

The model described above gives a context-sensitive indexing weight to each document term. The next step is to use these indexing weights to calculate the similarity between any two sentences. Given a sentence s_j in the document D_i , the sentence vector is built using the index $W_{(i)}$. The sentence vector s_j is calculated such that if a term v_k appears in s_j , it is given a weight $W_{(v_k)}$, otherwise, it is given a weight 0. The similarity between two sentences s_j and s_l is computed using the dot product, i.e., $\text{sim}(s_j, s_l) = \hat{s}_j \cdot \hat{s}_l$. Besides using the new sentence similarity measure, the paradigm as presented in Wan and Xiao [1], described is used for calculating the score of the sentences. The proposed method will be denoted by “bern” corresponding to the “Bernoulli” measure.

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	73%	84%
100	75%	86%
150	78%	89%
200	80%	92%
250	82%	94%

IV. 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 to the 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 positive impact 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://doi.acm.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 Query-Biased Summaries: A Comparison of Human and System Generated Snippets," Proc. Third Symp. Information Interaction in Context, pp. 195-204, <http://doi.acm.org/10.1145/1840784.1840813>, 2010.
- [4] X. Wan, "Towards a Unified Approach to Simultaneous Single-Document 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.acm.org/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 Multi-Document Summarization," ACM Trans. Information Systems, vol. 28, pp. 13:1-13:47, <http://doi.acm.org/10.1145/1777432.1777436>, July 2010.