Document Summarization Retrieval System Based on Web User Needs

K.Poojitha[1], M.Revathi[2], Mr.J.Santhosh Kumar[3]

[1, 2]InformationTechnology, AIHT, Chennai, Tamil Nadu, India

[3]Assistant Professor, Information Technology, AIHT, Chennai, Tamil Nadu, India

Abstract

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. 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.

1. 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.

2.1 Architecture

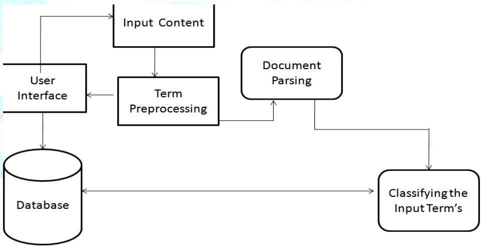

Fig.1: Existing System Architecture

3. Proposed System

The lexical association between terms to give a context sensitive weight to the document terms has been proposed. Proposed system using The traditional indexing schemes cannot distinguish between these terms that are reflected in the sentence similarity values. A cluster are more salient to the document topic. Sentence similarity measures based on cosine similarity was exploited for computing the adjacency matrix. We propose the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task. A novel term association metric using the Bernoulli model of randomness has been derived for this purpose.

3.1 Architecture

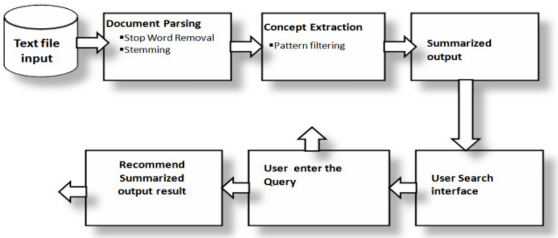

Fig.2: Proposed System Architecture

4. Overview

We propose the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task.

A novel term association metric using the Bernoulli model of randomness has been derived for this purpose. Quickly familiarize themselves with information contained in a large cluster of documents.

The advantages proposed system brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, comprehensive .Interesting observation was that the Bernoulli model for word indexing, when applied with the simplest Intra Link model, performs better than all the baseline models .

5. Modules

5.1 Preprocessing and Stop Word Removal

The preprocessing phase of the study converts the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified. It is the process of incorporating a new document into an information retrieval system. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall). This phase is the most critical and complex process that leads to the representation of each document by a select set of index terms. The main objective of preprocessing is to obtain the key features or key terms from online news text documents and to enhance the relevancy between word and document and the The goal behind preprocessing is to represent each document as a feature vector, that is, to separate the text into individual words. In the proposed classifiers, the text documents are modeled as transactions. Choosing the keyword that is the feature selection process, is the main preprocessing step necessary for the indexing of documents. This step is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning, and discard the words that do not contribute to distinguishing between the documents. Relevancy between word and category.

5.2 Context Based Word Indexing

The summaries shown above clearly reflect the advantage offered by the proposed Bernoulli-based word indexing model. The first two sentences in are very much similar to the first and third sentences in the manual summary. However, these sentences do not appear in the summary provided by Interlink. In the summary provided by Uniform Link, only the first sentence appears at the third position. These two sentences contain a lot of contextual words such as “communist,” “rebels,” “violence,” “police,” and so on. Since the proposed indexing model gives an indexing weight using the lexical association. With all other words, the weights of the contextual words are increased, reflecting on their sentence similarity values. Therefore, these sentences become more central in the “Intra Link + Bernoulli” method than the baseline models, Used without a context-based word indexing.

The lexical association between the terms in a target summary is higher compared to the association between the terms in a document. Thus, the proposed measure satisfies Hypothesis H1. It has been used along with the Page Rank algorithm to give a context-sensitive indexing weight to the document terms to validate Hypothesis H3. The indexing weights, thus, obtained have been used to calculate the sentence similarity values. The underlying assumption in H3 was that the sentence similarity thus obtained would be context sensitive and, therefore, should provide improvements in the sentence extraction task for document summarization. The concept of topical and non topical terms was used to modify the indexing weights of the document terms. Analysis of some of the documents and the corresponding summary figured out the specific advantage offered by the proposed Bernoulli model-based context sensitive indexing.

5.3 Indexing Weight to Each Document Term

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. 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 and/or document 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. A document contains both the content-carrying (topical) terms as well as background (non topical) 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 non topical terms and, thus, influences the sentence similarity values in a positive manner. In this paper, we address 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.

6 Experiments

6.1 Lexical Association Cluster Filtering

There is a known fact that, generally the more a NP appears, the more important it is in the document(s), especially when this NP appears in the 'Subject' position of sentences, and vice versa. From the sentence clusters, we first choose the largest word count cluster. Then, we set up a top threshold, 0.50, and a bottom one, 0.33, for filtering clusters, which means that any cluster whose size issmaller than 1/3 of the largest cluster is discarded; any cluster whose size is larger than 1/2 of the largest cluster is retained as a primary cluster; the other clusters are kept as secondary clusters.

6.2 Cluster-Reduction

The first round is textual redundant information reduction. In this round, the information of 'Protagonist' and synonyms of verbs and nouns returned from the WorldNet is thoroughly considered and phrase level comparing is carried out through out each cluster. The second round is indexing key units. All the causality-indexing chains in NP and clause levels and the intention-indexing chains provide very useful Connections among sentences and across clusters. All the NPs and clauses involved in these chains are considered as key units. Each chain has its own weight value according to its length, meaning the number of noun phrases or clauses on the chain. The third round is sorting these key units on each chain by the 'Temporality' of their sentences and organizing them based on their original sentence format. Any returned summary from the module of size control is resized by taking the value weight of each chain into account; the lower value weight a chain has, the less important it is.

6.3 Size Control and Output Summary

According to the size of each retained cluster and the total size of all retained clusters, we configure the percentage that each retained cluster should occupy in the final summary. If an extracted summary exceeds the required summary size, e.g. 100 words, the summary will be returned to the module of cluster reduction for further word reduction Once the summary is qualified at the required size, it will become an output of a formal result of summarization.

6.4 Cosine Text Similarity Technique

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Note that these bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.

The technique is also used to measure cohesion within clusters in the field of data mining.

Cosine distance is a term often used for the complement in positive space, that is: $D_C(A,B) = 1 - S_C(A,B)$. It is important to note, however, that this is not a proper metrics it does not have the triangle inequality property and it violates the coincidence axiom; to repair the triangle inequality property whilst maintaining the same ordering, it is necessary to convert to Angular distance.

7 Conclusions

In this paper, the Bernoulli model of randomness has been used to develop a graph-based ranking algorithm for calculating how informative is each of the document terms. We proposed three hypotheses, which were used for the development. Hypothesis H1 is based on the intuition that a document summary contains the most salient information in a text and therefore, the terms in a summary should be more lexically associated with each other than in the original document. This hypothesis was translated into an empirical relation, "The average lexical association between document summaries should be higher than the average lexical association between the original documents." The authors conjectured that if a lexical association measure follows H2, average lexical association will be higher in summaries than in the documents, i.e., it will follow H1. The authors also conjectured that if a lexical association measure follows H2, it will follows H3, that is, it can be used to give a context sensitive indexing weight to the document terms. This hypothesis was correlated with the Rouge scores

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] K.S. Jones, "Automatic Summarising: Factors and Directions,"Advances in Automatic Text Summarization, pp. 1-12, MIT Press,1998.

[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] X. Wan, "An Exploration of Document Impact on Graph-Based Multi-Document Summarization," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 755-762, http://portal.acm.org/citation.cfm?id=1613715.1613811, 2008.

[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.1020,http://portal.acm.org/citation.cfm?id=1747137. 1747140, May 2010.

[7] C. Shen and T. Li, "Multi-Document Summarization via the Minimum Dominating Set," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 984-992, http://portal.acm.org/citation.cfm?id= 1873781.1873892, 2010.

[8] X. Wan and J. Yang, "Multi-Document Summarization Using Cluster-Based Link Analysis," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,pp.299-306, http://doi.acm.org/10.1145/ 1390334.1390386, 2008.

[9] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 307-314, http://doi.acm.org/10.1145/1390334.1390387, 2008.

[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.

[11] H. Daume´ III and D. Marcu, "Bayesian Query-Focused Summarization," Proc. 21st Int'l Conf. Computational Linguistics and the 44th Ann. meeting of the Assoc. for Computational Linguistics, pp. 305-312, http://dx.doi.org/10.3115/1220175.1220214, 2006.

[12] D.M. Dunlavy, D.P. O'Leary, J.M. Conroy, and J.D. Schlesinger ,"QCS: A System for Querying, Clustering and Summarizing Documents," Information Processing and Management,vol.43,pp.1588-1605, http://portal.acm.org/citation.cfm?id=1284916.1285163, Nov. 2007.

[13] R. Varadarajan, V. Hristidis, and T. Li, "Beyond Single-Page Web Search Results," IEEE Trans. Knowledge and Data Eng., vol. 20,no. 3, pp. 411-424, Mar. 2008.

[14] L.-W. Ku, L.-Y. Lee, T.-H. Wu, and H.-H. Chen, "Major Topic Detection and Its Application to Opinion Summarization," Proc.28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 627-628, http://doi.acm.org/10.1145/1076034.1076161, 2005.

[15] E. Lloret, A. Balahur, M. Palomar, and A. Montoyo, "Towards Building a Competitive Opinion Summarization System: Challenges and Keys," Proc. Human Language Technologies: The 2009 Ann. Conference of the North Am. Ch. Assoc. for Computational Linguistics, Companion Vol. : Student Research Workshop and Doctoral Consortium,pp.72,77,http://portal.acm.org/citation.cfm?id= 1620932.1620945, 2009.

[16] J.G. Conrad, J.L. Leidner, F. Schilder, and R. Kondadadi, "Query-Based Opinion Summarization for Legal Blog Entries," Proc. 12th Int'l Conf. Artificial Intelligence and Law, pp. 167-176, http://doi. acm.org/10.1145/1568234.1568253, 2009.

[17] 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.

[18] 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.

[19] D.R. Radev, H. Jing, M. Sty_s, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management, vol. 40, pp. 919-938, http://portal.acm.org/citation.cfm?id=1036118.1036121, Nov. 2004.