A Survey on Automatic Text Summarization Methods

Janit Chadha[1]

[1]*Student, Dept. of Computer Science Engineering, BNMIT, Karnataka, India*

***

Abstract – *In today's world wealth of data is available o Internet, text mining plays a vital role in many fields. Text mining has been widely used as 80%(approx..) of the data is unstructured. Mining information from voluminous data is a tedious job. Automatic text summarization is a powerful tool used for summarizing the text. The motive of automatic text summarizer text summarization is not only providing summary, but also it is representing the summary in a meaningful manner.*

Key Words: **Text summarization, Text mining, Resource description framework (RDF), Natural Language Processing (NLP).**

1. INTRODUCTION

Textual content analytics may be completed manually, but it's miles an inefficient method. Consequently, textual content analytics software has been created that uses text mining and natural language processing algorithms to locate meaning in huge quantities of text. Text analytics software can provide an early warning of trouble, as it suggests what clients are complaining about. the usage of textual content analytics tools gives you treasured facts from statistics that is not without problems quantified in another way. It turns the unstructured mind of customers into established facts that can be utilized by commercial enterprise. Text Summarization is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives.

Text summarization can broadly be divided into two categories:

Extractive Summarization and Abstractive Summarization.

Extractive Summarization: Extractive methods rely on extracting majority of parts, like phrases and sentences, from a piece of document and combine them together to create a summary. Identifying the right sentences for summarization is of utmost importance in an extractive method.

Abstractive Summarization: Abstractive methods use latest NLP techniques to generate a new summary. Some parts of this summary sometimes not even appear in the original text.

There are also two scales of document summarization:

Single-document summarization: the task of summarizing a standalone document. Note that a" document" could refer to different things depending on the use case (URL, internal PDF file, legal contract, financial report, email, etc.)

Multi-document summarization: the task of assembling a collection of documents (usually through a query against a database or search engine) and generating a summary that incorporates perspectives from across documents.

The main purpose of this research is an overview of different approaches for automatic text summarization.

Text Summarization Applications: -

Media monitoring: The problem of information overload and "content shock" has been widely discussed. Automatic summarization presents an opportunity to condense the continuous torrent of information into smaller pieces of information.

Newsletters: Many weekly newsletters take the form of an introduction followed by a curated selection of relevant articles. Summarization would allow organizations to further enrich newsletters with a stream of summaries (versus a list of links), which can be a particularly convenient format in mobile.

Search marketing and SEO: queries for SEO, it is critical to have a well-rounded understanding of what the competitors are talking about in their content. This has become particularly important since Google updated its algorithm and shifted focus towards topical authority (versus keywords). Multi-document summarization can be a powerful tool to quickly analyze dozens of search results, understand shared themes and skim the most important points.

Internal document workflow: Large companies are constantly producing internal knowledge, which frequently gets stored and under-used in databases as unstructured data. These companies should embrace tools that let them re-use already existing knowledge. Summarization can enable analysts to quickly understand everything the company has already done in a given subject, and quickly assemble reports that incorporate different points of view.

Financial research: Investment banking firms spend large amounts of money acquiring information to drive their decision-making, including automated stock trading. When you are a financial analyst looking at market reports and news everyday, you will inevitably hit a wall and won't be able to read everything. Summarization systems tailored to financial documents like earning reports and financial news can help analysts quickly derive market signals from content.

2. TEXT SUMMARIZATION METHODS

Resourse Desciprtion Famework (RDF):

In this paper [1] RDF is a ontology language for making a ontology system. Ontology system yield skills which can be seen in areas like changes of data concept etc. Firstly, examination of the causes for RDF schema using transitivity rule. Then finally a storage structure and management skills for RDF using RDB is concluded. Different strategies for semantic web are scrutinized.

RDF size is voluminous and governing it has become a tedious task as existing terminologies are hard and complex. In this paper [2] E-R (Entity relationship) model is used to store RDF data in distinct tables. Hence by applying this, it is way easy to search and update common handling in RDB (Relational Database).
A new RDFS saving strategy is introduced. RDF schema is used for specifying the meaning among RDF data elements and keeping the RDF data into distinct relation tables.

Linked data mining is one of route questions for High performance graph mining. Present Resource description framework (RDF) database engines are not at all flexible and are less authentic in heterogeneous clouds.

A architecture consisting of RDF graph on top of Acacia- RDF is made. Fourth, we have implemented a replication-based fault tolerance mechanism for Acacia RDF. Finally, we have implemented several additional graph algorithms in Acacia's distributed data abstraction.

Specifically, the contributions can be listed as follows,

- Distributed RDF graph database engine -We describe the architecture of a distributed, scalable RDF graph database engine which partitions and persists RDF graphs across multiple workers.
- X10-based SPARQL Executor - We design and implement a SPARQL query processor completely in X10.
- X10-based Fault Tolerance - We leverage X10 language's support for developing fault tolerant graph database server which tolerates the scenarios of failures.
- Evaluation - We provide performance evaluation results of the Acacia-RDF system with experiments conducted using LUBM data sets up to LUBM scale 160.

In this paper [3] Semantic annotation is a dignified representation of content concerned with information resource. Annotations created automatically are in RDF language, expressed using concepts, relations and instances. Most of the research use Recall, Precision and F-measure for evaluation.
The RDF triplets are utilized correctly to evaluate RDF semantic annotations.

In this paper [4] Evaluation protocol for RDF semantic annotations is introduced. Present RDF Stream Processing (RSP) systems are solving the variety lock by explaining a common model for producing, transmitting querying data in RDF model. On the volume and velocity side, the performances of RSP systems need to be improved particularly in terms of joins process between stored and streaming RDF graphs.
Systems need to read the entire local or remote stored RDF data sets while RDF data streams continuously arrived and need to be processed in near real-time. This latency may negatively affect performances in terms of continuous processing and often causes multiple bottlenecks within the network in a distributed environment.

Text summarization:

In this paper [5] Exponentially growing amount of textual information has become a serious research topic in machine generated summarization. Automatic text summarizer is "condensing the source text into a shorter version, while preserving its information content and overall meaning". Heading wise single document summarizer for the improvement of logical meaning of the document. Heading wise summarizer provides feature based extractive single document heading wise summary of source document using statistical and linguistic features.

For utilizing on-line documents usefully, it is important to extract the summary of these documents.

In this paper [6] Summary of multiple document is done through summarizer using semantic nature of web documents and transforming them into triples in the form of subject, verb and object (RDF Triples). A well-known clustering algorithm is used to cluster the triples.

Summarizer based on semantic analysis of web document is introduced.

In this paper [7] Version of KNN (K Nearest Neighbor) where the similarity between feature vectors is computed considering the similarity among attributes or features as well as one among values. Text summarization is viewed into the binary classification task where each paragraph or sentence is classified into the essence or non-essence.

Similarity which considers both attributes and attribute values, modify the KNN into the version based on the similarity, and use the modified version as the approach to the text summarization task.

In this paper [8] Approach for an extractive query focused multi-document summarization which stands on an enhanced knowledge-based short text semantic similarity measure. Summarizer is built primarily on such an improved semantic similarity measure to model relevance, centrality and diversity factors outperforms the best-performing relevant DUC systems and recent closely related studies in at least one or more of the investigated ROUGE metrics.

The primary contributions: -

1) The improvement of WordNet-based similarity by converting all possible loosely encoded and non-hierarchized word categories (e.g. verbs, adverbs and adjectives) to nouns due to the well-structured full-fledged noun taxonomy as compared to other parts of speech encoded in WordNet.
2) A new technique for handling semantic relatedness between named entities based on the co-occurrences of designated names in Wikipedia articles is put forward.
3) The use of these conversion-aided knowledge enriched semantic similarity measures as the chief indicators of salient content for feature-based extractive multi-document summarization.

<u>Word disambiguation:</u>

In this paper [9] A new model model of WordNet that is used to disambiguate the correct sense of polysemy word based on the clue words. The related words for each sense of a polysemy word as well as single sense word are referred to as the clue words. These clue words for each sense of a polysemy word as well as for single sense word are used to disambiguate the correct meaning of the polysemy word in the given context using knowledge-based Word Sense Disambiguation (WSD) algorithms.

In this paper [10] Context word of the ambiguous word is an important basis for word sense disambiguation (WSD). Adopted three methods to compute WSD weight of context words, which are the method to assign uniform weight for each word, the method to assign weight with exponential function and the method to assign weight with power function.

An extension of a semantic similarity measure for large linked data sources such as DBPedia. To evaluate and compare similarity measures, an experiment is described to collect human evaluations of the similarity between pairs of movies. Introduced an extension of a semantic similarity measure to be used with real linked data sources and then present an experiment
to collect human evaluations of the similarity between pairs of movies to evaluate the measure.

In this paper [11] Measuring the semantic similarity between words is an important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. Despite the usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words (or entities) remains a challenging task. The information in web is rapidly growing so finding the semantic similarity between words does not possible in Lexical Based search and Semantic Based Search only finds the similarity.

3. Conclusion

We have provided a very brief introductions to the text mining. Then several general applications are described to know text mining in the overall perspective. Then we classified text mining work as resource description framework, text summarization and word disambiguation. Text mining is a new direction of artificial intelligence, and with the continuous improvement of the text mining technology, its application areas will be growing.

ACKNOWLEDGEMENT (Optional)

I would like to thank my parents and Dr. Shashikala for providing me proper guidance.

REFERENCES

1) ByungGon Kim, YounHee Kim and Haechull Lim "Storage structure and management skills for RDF Schema versioning in RDB environment" Feb..20-22, 2006 ICA0T2006

2) LiLi Xu, SangWon Lee, Seokhyun Kim "E-R Model based RDF Data Storage in RDB" 978-1-4244-5540-9/10 20 1 0 IEEE.

3) Miyuru Dayarathna, Isuru Herath, Yasima Dewmini, Gayan Mettananda, Sameera Nandasiri, Sanath Jayasena, Toyotaro Suzumura "Acacia-RDF: An X10-based Scalable Distributed RDF Graph Database Engine" 2016 IEEE 9th International Conference on Cloud Computing

4) Rim Teyeb, Mouna Torjmen, Afef Latrach "Towards an evaluation protocol for RDF semantic annotations (RDF SemAnnotEval Protocol)" ICEMIS2017, Monastir, Tunisia

5) P.Krishnaveni, Dr.S. R. Balasundaram "Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence" Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC)

6) Angelin Florence, Vijaya Padmadas " A Summarizer System Based on a Semantic Analysis of Web Documents" 2015 International Conference on

Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India

7) Taeho Jo “K Nearest Neighbor for Text Summarization using Feature Similarity” 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), Khartoum, Sudan

8) Muhidin A. Mohamed, Mourad Oussalah “Similarity-based Query-focused Multi-document Summarization using Crowdsourced and Manually-built Lexical-Semantic Resources” 2015 IEEE Trustcom/BigDataSE/ISPA

9) Udaya Raj Dhungana, Subarna Shakya, Kabita BaraP and Bharat Sharma “Word Sense Disambiguation using WSD Specific WordNet of Polysemy Words” Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)

10) Valentin Grou`es, Yannick Naudet, Odej Kao “Adaptation and Evaluation of a Semantic Similarity Measure for DBPedia: A First Experiment” 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization

11) Samhith.K, Arun Tilak.S, Prof G.Panda “Word Sense Disambiguation using WordNet Lexical Categories” International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016